Universität Stuttgart

Fachbereich Mathematik

A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning

Hanyuan Hang, Ingo Steinwart

Preprint 2015/016

Fachbereich Mathematik
Fakultät Mathematik und Physik
Universität Stuttgart
Pfaffenwaldring 57
D-70 569 Stuttgart

E-Mail: preprints@mathematik.uni-stuttgart.de
WWW: http://www.mathematik.uni-stuttgart.de/preprints

ISSN **1613-8309**

A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning

H. Hang and I. Steinwart
Institute for Stochastics and Applications
University of Stuttgart
D-70569 Stuttgart
{hanghn,ingo.steinwart}@mathematik.uni-stuttgart.de

January 13, 2015

Abstract

We establish a Bernstein-type inequality for a class of stochastic processes that include the classical geometrically ϕ-mixing processes, Rio's generalization of these processes, as well as many time-discrete dynamical systems. Modulo a logarithmic factor and some constants, our Bernstein-type inequality coincides with the classical Bernstein inequality for i.i.d. data. We further use this new Bernstein-type inequality to derive an oracle inequality for generic regularized empirical risk minimization algorithms and data generated by such processes. Applying this oracle inequality to support vector machines using the Gaussian kernels for both least squares and quantile regression, it turns out that the resulting learning rates match, up to some arbitrarily small extra term in the exponent, the optimal rates for i.i.d. processes.

1 Introduction

Concentration inequalities such as Hoeffding's inequality, Bernstein's inequality, McDiarmid's inequality, and Talagrand's inequality play an important role in many areas of probability. For example, the analysis of various methods from non-parametric statistics and machine learning crucially depend on these inequalities, see e.g. [19, 20, 22, 42]. Here, stronger results can typically be achieved by Bernstein's inequality and/or Talagrand's inequality, since these inequalities allow for localization due to their specific dependence on the variance. In particular, most derivations of minimax optimal learning rates are based on one of these inequalities.

The concentration inequalities mentioned above all assume the data to be generated by an i.i.d. process. Unfortunately, however, this assumption is often violated in several important areas of applications including financial prediction, signal processing, system observation and diagnosis, text and speech recognition, and time series forecasting. For this and other reasons there has been some effort to establish concentration inequalities for non-i.i.d. processes, too. For example, generalizations of Bernstein's inequality to α-mixing and ϕ-mixing processes have been found [10, 33, 32] and [38], respectively. Among many other applications, the Bernstein-type inequality established in [10] was used in [50] to obtain convergence rates for sieve estimates from α-mixing strictly stationary processes in the special case of neural networks. Furthermore, [23] applied the Bernstein-type inequality in [33] to derive an oracle inequality for generic regularized empirical risk minimization algorithms learning from stationary α-mixing processes. Moreover, by employing the Bernstein-type inequality in [32], [7] derived almost sure uniform rates of convergence for the estimated Lévy density both in mixed-frequency and low-frequency setups and proved that these rates are optimal in the minimax sense. Finally, in the particular case of the least square loss, [2] obtained the optimal learning rate for ϕ-mixing processes by applying the Bernstein-type inequality established in [38].

However, there exist many dynamical systems such as the uniformly expanding maps given in [17, p. 41] that are not α-mixing. To deal with such non-mixing processes Rio [34] introduced so-called $\tilde{\phi}$-mixing coefficients, which extend the classical ϕ-mixing coefficients. For dynamical systems with exponentially decreasing, *modified* $\tilde{\phi}$-coefficients, [47] derived a Bernstein-type inequality, which turns out to be the same as the one for i.i.d. processes modulo some logarithmic factor. However, this modification seems to be significant stronger than Rio's original $\tilde{\phi}$-mixing, so it remains unclear when the Bernstein-type inequality in [47] is applicable. In addition, the $\tilde{\phi}$-mixing concept is still not large enough to cover many commonly considered dynamical systems. To include such dynamical systems, [31] proposed the $\mathcal{C}$-mixing coefficients, which further generalize $\tilde{\phi}$-mixing coefficients.

In this work, we establish a Bernstein-type inequality for geometrically $\mathcal{C}$-mixing processes, which, modulo a logarithmic factor and some constants, coincides with the classical one for i.i.d. processes. Using the techniques developed in [23], we then derive an oracle inequality for generic regularized empirical risk minimization and $\mathcal{C}$-mixing processes. We further apply this oracle inequality to a state-of-the-art learning method, namely support vector machines (SVMs) with Gaussian kernels. Here it turns out that for both, least squares and quantile regression, we can recover the (essentially) optimal rates recently found for the i.i.d. case, see [21], when the data is generated by a geometrically $\mathcal{C}$-mixing process. Finally, we establish an oracle inequality for the problem of forecasting an unknown dynamical system. This oracle will make it possible to extend the purely asymptotic analysis in [41] to learning rates.

The rest of this work is organized as follows: In Section 2, we recall the notion of (time-reversed) $\mathcal{C}$-mixing processes. We further illustrate this class of processes by some examples and discuss the relation between $\mathcal{C}$-mixing and other notions of mixing. As the main result of this work, a Bernstein-type inequality for geometrically (time-reversed) $\mathcal{C}$-mixing processes will be formulated in Section 3. There, we also compare our new Bernstein-type inequality to previously established concentration inequalities. As an application of our Bernstein-type inequality, we will derive the oracle inequality for regularized risk minimization schemes in Section 4. We additionally derive learning rates for SVMs and an oracle inequality for forecasting certain dynamical systems. All proofs can be found in the last section.

2 $\mathcal{C}$-mixing processes

In this section we recall two classes of stationary stochastic processes called (time-reversed) $\mathcal{C}$-mixing processes that have a certain decay of correlations for suitable pairs of functions. We also present some examples of such processes including certain dynamical systems.

Let us begin by introducing some notations. In the following, $(\Omega, \mathcal{A}, \mu)$ always denotes a probability space. As usual, we write $L_p(\mu)$ for the space of (equivalence classes of) measurable functions $f : \Omega \to \mathbb{R}$ with finite L_p-norm $\|f\|_p$. It is well-known that $L_p(\mu)$ together with $\|f\|_p$ forms a Banach space. Moreover, if $\mathcal{A}' \subset \mathcal{A}$ is a sub-σ-algebra, then $L_1(\mathcal{A}', \mu)$ denotes the space of all $\mathcal{A}'$-measurable functions $f \in L_1(\mu)$. In the following, for a Banach space E, we write B_E for its closed unit ball.

Given a semi-norm $\|\cdot\|$ on a vector space E of bounded measurable functions $f : Z \to \mathbb{R}$, we define the $\mathcal{C}$-Norm by

$$\|f\|_{\mathcal{C}} := \|f\|_\infty + \|f\| \tag{1}$$

and denote the space of all bounded $\mathcal{C}$-functions by

$$\mathcal{C}(Z) := \big\{ f : Z \to \mathbb{R} \,\big|\, \|f\|_{\mathcal{C}} < \infty \big\}. \tag{2}$$

Throughout this work, we only consider the semi-norms $\|\cdot\|$ in (1) that satisfy the inequality

$$\left\|e^f\right\| \leq \left\|e^f\right\|_\infty \|f\| \tag{3}$$

for all $f \in \mathcal{C}(Z)$. We are mostly interested in the following examples of semi-norms satisfying (3).

Example 2.1. Let Z be an arbitrary set and suppose that we have $\|f\| = 0$ for all $f : Z \to \mathbb{R}$. Then, it is obviously to see that $\|e^f\| = \|f\| = 0$. Hence, (3) is satisfied.

Example 2.2. Let $Z \subset \mathbb{R}$ be an interval. A function $f : Z \to \mathbb{R}$ is said to have bounded variation on Z if its total variation $\|f\|_{BV(Z)}$ is bounded. Denote by $BV(Z)$ the set of all functions of bounded variation. It is well-known that $BV(Z)$ together with $\|f\|_\infty + \|f\|_{BV(Z)}$ forms a Banach space. Moreover, we have (3), i.e. we have for all $f \in \mathcal{C}(Z)$:

$$\left\|e^f\right\|_{BV(Z)} \leq \left\|e^f\right\|_\infty \|f\|_{BV(Z)}.$$

Example 2.3. Let Z be a subset of $\mathbb{R}^d$ and $C_b(Z)$ be the set of bounded continuous functions on Z. For $f \in C_b(Z)$ and $0 < \alpha \leq 1$ let

$$\|f\| := |f|_\alpha := \sup_{z \neq z'} \frac{|f(z) - f(z')|}{|z - z'|^\alpha}.$$

Clearly, f is α-Hölder continuous if and only if $|f|_\alpha < \infty$. The collection of bounded, α-Hölder continuous functions on Z will be denoted by

$$C_{b,\alpha}(Z) := \{f \in C_b(Z) : |f|_\alpha < \infty\}.$$

Note that, if Z is compact, then $C_{b,\alpha}(Z)$ together with the norm $\|f\|_{C_{b,\alpha}} := \|f\|_\infty + |f|_\alpha$ forms a Banach space. Moreover, the inequality (3) is also valid for $f \in C_{b,\alpha}(Z)$. As usual, we speak of Lipschitz continuous functions if $\alpha = 1$ and write $\mathrm{Lip}(Z) := C_{b,1}(Z)$.

Example 2.4. Let $Z \subset \mathbb{R}^d$ be an open subset. For a continuously differentiable function $f : Z \to \mathbb{R}$ we write

$$\|f\| := \sup_{z \in Z} |f'(z)|$$

and $C^1(Z) := \left\{f : Z \to \mathbb{R} \,|\, f \text{ continuously differentiable and } \|f\|_\infty + \|f\| < \infty\right\}$. It is well-known, that $C^1(Z)$ is a Banach space with respect to the norm $\|\cdot\|_\infty + \|\cdot\|$ and the chain rule gives

$$\left\|e^f\right\| = \left\|\left(e^f\right)'\right\|_\infty = \left\|e^f \cdot f'\right\|_\infty \leq \left\|e^f\right\|_\infty \|f'\|_\infty = \left\|e^f\right\|_\infty \|f\|,$$

for all $f \in C^1(Z)$, i.e. (3) is satisfied.

Let us now assume that we also have a measurable space $(Z, \mathcal{B})$ and a measurable map $\chi : \Omega \to Z$. Then $\sigma(\chi)$ denotes the smallest σ-algebra on Ω for which χ is measurable. Moreover, μ_χ denotes the χ-image measure of μ, which is defined by $\mu_\chi(B) := \mu(\chi^{-1}(B))$, $B \in \mathcal{B}$.

Let $\mathcal{Z} := (Z_n)_{n \geq 0}$ be a Z-valued stochastic process on $(\Omega, \mathcal{A}, \mu)$, and $\mathcal{A}_0^i$ and $\mathcal{A}_{i+n}^\infty$ be the σ-algebras generated by $(Z_0, \ldots, Z_i)$ and $(Z_{i+n}, Z_{i+n+1}, \ldots)$, respectively. The process $\mathcal{Z}$ is called *stationary* if $\mu_{(Z_{i_1+i}, \ldots, Z_{i_n+i})} = \mu_{(Z_{i_1}, \ldots, Z_{i_n})}$ for all $n, i, i_1, \ldots, i_n \geq 1$. In this case, we always write $P := \mu_{Z_0}$. Moreover, to define certain dependency coefficients for $\mathcal{Z}$, we denote, for $\psi, \varphi \in L_1(\mu)$ satisfying $\psi\varphi \in L_1(\mu)$ the correlation of ψ and φ by

$$\mathrm{cor}(\psi, \varphi) := \int_\Omega \psi \cdot \varphi \, d\mu - \int_\Omega \psi \, d\mu \cdot \int_\Omega \varphi \, d\mu \,.$$

Several dependency coefficients for $\mathcal{Z}$ can be expressed by imposing restrictions on ψ and φ. The following definition, which is taken from [31], introduces the restrictions on ψ and φ we consider throughout this work.

Definition 2.5. *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $(Z, \mathcal{B})$ be a measurable space, $\mathcal{Z} := (Z_i)_{i\geq 0}$ be a Z-valued, stationary process on Ω, and $\|\cdot\|_{\mathcal{C}}$ be defined by* (1) *for some semi-norm $\|\cdot\|$. Then, for $n \geq 0$, we define:*

(i) the $\mathcal{C}$-mixing coefficients by

$$\phi_{\mathcal{C}}(\mathcal{Z}, n) := \sup\left\{\operatorname{cor}(\psi, h \circ Z_{k+n}) : k \geq 0, \psi \in B_{L_1(\mathcal{A}_0^k, \mu)}, h \in B_{\mathcal{C}(Z)}\right\} \tag{4}$$

(ii) the time-reversed $\mathcal{C}$-mixing coefficients by

$$\phi_{\mathcal{C},rev}(\mathcal{Z}, n) := \sup\left\{\operatorname{cor}(h \circ Z_k, \varphi) : k \geq 0, h \in B_{\mathcal{C}(Z)}, \varphi \in B_{L_1(\mathcal{A}_{k+n}^\infty, \mu)}\right\}. \tag{5}$$

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

$$d_n := c\exp\left(-bn^\gamma\right), \qquad n \geq 1, \tag{6}$$

for some constants $b > 0$, $c \geq 0$, and $\gamma > 0$, then $\mathcal{Z}$ is called *geometrically* (time-reversed) $\mathcal{C}$-mixing.

Obviously, $\mathcal{Z}$ is $\mathcal{C}$-mixing with rate $(d_n)_{n\geq 0}$, if and only if for all $k, n \geq 0$, all $\psi \in L_1(\mathcal{A}_0^k, \mu)$, and all $h \in \mathcal{C}(Z)$, we have

$$\operatorname{cor}(\psi, h \circ Z_{k+n}) \leq \|\psi\|_{L_1(\mu)} \|h\|_{\mathcal{C}}\, d_n, \tag{7}$$

or similarly, time-reversed $\mathcal{C}$-mixing with rate $(d_n)_{n\geq 0}$, if and only if for all $k, n \geq 0$, all $h \in \mathcal{C}(Z)$, and all $\varphi \in L_1(\mathcal{A}_{k+n}^\infty, \mu)$, we have

$$\operatorname{cor}(h \circ Z_k, \varphi) \leq \|h\|_{\mathcal{C}} \|\varphi\|_{L_1(\mu)}\, d_n. \tag{8}$$

In the rest of this section we consider examples of (time-reversed) $\mathcal{C}$-mixing processes. To begin with, let us assume that $\mathcal{Z}$ is a stationary ϕ-mixing process [25] with rate $(d_n)_{n\geq 0}$. By [16, Inequality (1.1)] we then have

$$\operatorname{cor}(\psi, \varphi) \leq \|\psi\|_{L_1(\mu)} \|\varphi\|_{L_\infty(\mu)} d_n, \quad n \geq 1, \tag{9}$$

for all $\mathcal{A}_0^k$-measurable $\psi \in L_1(\mu)$ and all $\mathcal{A}_{k+n}^\infty$-measurable $\varphi \in L_\infty(\mu)$. By taking $\|\cdot\|_{\mathcal{C}} := \|\cdot\|_\infty$ and $\varphi := h \circ Z_{k+n}$, we then see that (7) is satisfied, i.e. $\mathcal{Z}$ is $\mathcal{C}$-mixing with rate $(d_n)_{n\geq 0}$. Finally, by similar arguments we can deduce that time-reversed ϕ-mixing processes [12, Section 3.13] are also time-reversed $\mathcal{C}$-mixing with the same rate. In other words we have found

$$\phi_{L_\infty(\mu)}(\mathcal{Z}, n) = \phi(\mathcal{Z}, n) \qquad \text{and} \qquad \phi_{L_\infty(\mu),\mathrm{rev}}(\mathcal{Z}, n) = \phi_{\mathrm{rev}}(\mathcal{Z}, n).$$

To deal with processes that are not α-mixing [35], Rio [34] introduced the following relaxation of ϕ-mixing coefficients

$$\begin{aligned}\tilde{\phi}(\mathcal{Z}, n) &:= \sup_{\substack{k\geq 0,\\ f\in BV_1}} \left\|\mathbb{E}\left(f(Z_{k+n})\middle|\mathcal{A}_0^k\right) - \mathbb{E}f(Z_{k+n})\right\|_\infty \\ &= \sup\left\{\operatorname{cor}(\psi, h \circ Z_{k+n}) : k \geq 0, \psi \in B_{L_1(\mathcal{A}_0^k, \mu)}, h \in B_{BV(Z)}\right\}\end{aligned} \tag{10}$$

and an analogous time-reversed coefficient

$$\tilde{\phi}_{\mathrm{rev}}(\mathcal{Z}, n) := \sup_{\substack{k\geq 0,\\ f\in BV_1}} \left\|\mathbb{E}\left(f(Z_k)\middle|A_{k+n}^\infty\right) - \mathbb{E}f(Z_k)\right\|_\infty$$

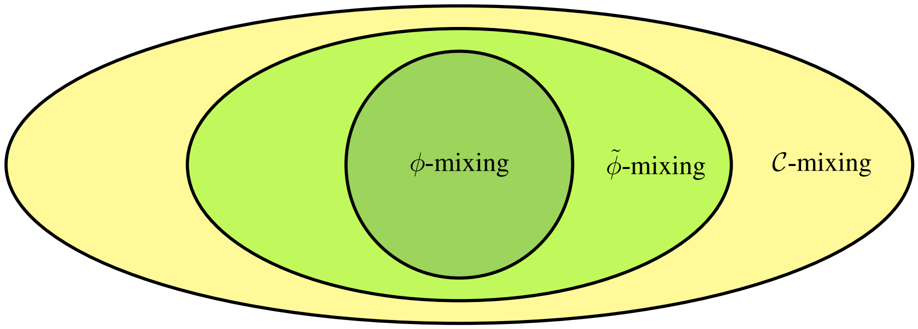

Figure 1: Relationship between ϕ-, $\tilde{\phi}$-, and $\mathcal{C}$-mixing processes

$$= \sup \left\{ \operatorname{cor}(h \circ Z_k, \varphi) : k \geq 0, \varphi \in B_{L_1(\mathcal{A}_{k+n}^{\infty}, \mu)}, h \in B_{BV(Z)} \right\},$$

where the two identities follow from [18, Lemma 4]. In other words we have

$$\phi_{BV(Z)}(\mathcal{Z}, n) = \tilde{\phi}(\mathcal{Z}, n) \qquad \text{and} \qquad \phi_{BV(Z),\text{rev}}(\mathcal{Z}, n) = \tilde{\phi}_{\text{rev}}(\mathcal{Z}, n)$$

Moreover, [17, p. 41] shows that some uniformly expanding maps are $\tilde{\phi}$-mixing but not α-mixing. Figure 1 summarizes the relations between ϕ, $\tilde{\phi}$, and $\mathcal{C}$-mixing.

Our next goal is to relate $\mathcal{C}$-mixing to some well-known results on the decay of correlations for dynamical systems. To this end, recall that $(\Omega, \mathcal{A}, \mu, T)$ is a dynamical system, if $T : \Omega \to \Omega$ is a measurable map satisfying $\mu(T^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{A}$. Let us consider the stationary stochastic process $\mathcal{Z} := (Z_n)_{n\geq 0}$ defined by $Z_n := T^n$ for $n \geq 0$. Since $\mathcal{A}_{n+1}^{n+1} \subset \mathcal{A}_n^n$ for all $n \geq 0$, we conclude that $\mathcal{A}_{k+n}^{\infty} = \mathcal{A}_{k+n}^{k+n}$. Consequently, φ is $\mathcal{A}_{k+n}^{\infty}$-measurable, if and only if it is $\mathcal{A}_{k+n}^{k+n}$-measurable. Moreover $\mathcal{A}_{k+n}^{k+n}$ is the σ-algebra generated by T^{k+n}, and hence φ is $\mathcal{A}_{k+n}^{k+n}$-measurable, if and only if it is of the form $\varphi = g \circ T^{k+n}$ for some suitable, measurable $g : \Omega \to \mathbb{R}$. Let us now suppose that $\|\cdot\|_{\mathcal{C}(\Omega)}$ is defined by (1) for some semi-norm $\|\cdot\|$. For $h \in \mathcal{C}(\Omega)$ we then find

$$\begin{aligned} \operatorname{cor}(h \circ Z_k, \varphi) = \operatorname{cor}(h \circ Z_k, g \circ Z_{k+n}) &= \operatorname{cor}(h, g \circ Z_n) \\ &= \int_\Omega h \cdot (g \circ T^n)\, d\mu - \int_\Omega h\, d\mu \cdot \int_\Omega g\, d\mu \\ &=: \operatorname{cor}_{T,n}(h, g)\,. \end{aligned}$$

The next result shows that $\mathcal{Z}$ is time-reversed $\mathcal{C}$-mixing even if we only have generic constants $C(h, g)$ in (8).

Theorem 2.6. *Let $(\Omega, \mathcal{A}, \mu, T)$ be a dynamical system and the stochastic process $\mathcal{Z} := (Z_n)_{n\geq 0}$ be defined by $Z_n := T^n$ for $n \geq 0$. Moreover, Let $\|\cdot\|_{\mathcal{C}}$ be defined by* (1) *for some semi-norm $\|\cdot\|$. Then, $\mathcal{Z}$ is time-reversed $\mathcal{C}$-mixing with rate $(d_n)_{n\geq 0}$ iff for all $h \in \mathcal{C}(\Omega)$ and all $g \in L_1(\mu)$ there exists a constant $C(h, g)$ such that*

$$\operatorname{cor}_{T,n}(h, g) \leq C(h, g) d_n, \qquad n \geq 0.$$

Thus, we see that $\mathcal{Z}$ is time-reversed $\mathcal{C}$-mixing, if $\operatorname{cor}_{T,n}(h, g)$ converges to zero for all $h \in \mathcal{C}(\Omega)$ and $g \in L_1(\mu)$ with a rate that is independent of h and g.

For concrete examples, let us first mention that [31] presents some discrete dynamical systems that are time-reversed geometrically $\mathcal{C}$-mixing such as Lasota-Yorke maps, uni-modal maps, piecewise expanding maps in higher dimension. Here, the involved spaces are either $BV(Z)$ or $\mathrm{Lip}(Z)$.

In dynamical systems where chaos is weak, correlations often decay polynomially, i.e. the correlations satisfy

$$|\mathrm{cor}_{T,n}(h,g)| \leq C(h,g)\cdot n^{-b}\,, \qquad\qquad n\geq 0, \tag{11}$$

for some constants $b > 0$ and $C(h, g) \geq 0$ depending on the functions h and g. Young [49] developed a powerful method for studying correlations in systems with weak chaos where correlations decay at a polynomial rate for bounded g and Hölder continuous h. Her method was applied to billiards with slow mixing rates, such as Bunimovich billiards, see [6, Theorem 3.5]. For example, modulo some logarithmic factors [30, 14] obtained (11) with $b = 1$ and $b = 2$ for certain forms of Bunimovich billiards and Hölder continuous h and g. Besides these results, Baladi [5] also compiles a list of "parabolic" or "intermittent" systems having a polynomial decay.

It is well-known that, if the functions h and g are sufficient smooth, there exist dynamical systems where chaos is strong enough such that the correlations decay exponentially fast, that is,

$$|\mathrm{cor}_{T,n}(h,g)| \leq C(h,g)\cdot \exp\bigl(-bn^{\gamma}\bigr)\,, \qquad\qquad n\geq 0, \tag{12}$$

for some constants $b > 0$, $\gamma > 0$, and $C(h, g) \geq 0$ depending on h and g. Again, Baladi [5] has listed some simple examples of dynamical systems enjoying (12) for analytic h and g such as the angle doubling map and the Arnold's cat map. Moreover, for continuously differentiable h and g, [36, 39] proved (12) for two closely related classes of systems, more precisely, $C^{1+\varepsilon}$ Anosov or the Axiom-A diffeomorphisms with Gibbs invariant measures and topological Markov chains, which are also known as subshifts of finite type, see also [11]. These results were then extended by [24, 37] to expanding interval maps with smooth invariant measures for functions h and g of bounded variation. In the 1990s, similar results for Hölder continuous h and g were proved for systems with somewhat weaker chaotic behavior which is characterized by nonuniform hyperbolicity, such as quadratic interval maps, see [48], [27] and the Hénon map [8], and then extended to chaotic systems with singularities by [28] and specifically to Sinai billiards in a torus by [48, 13]. For some of these extensions, such as smooth expanding dynamics, smooth nonuniformly hyperbolic systems, and hyperbolic systems with singularities, we refer to [4] as well. Recently, for h of bounded variation and bounded g, [29] obtained (12) for a class of piecewise smooth one-dimensional maps with critical points and singularities. Moreover, [3] has deduced (12) for $h, g \in \mathrm{Lip}(Z)$ and a suitable iterate of Poincaré's first return map T of a large class of singular hyperbolic flows.

3 A Bernstein-type inequality

In this section, we present the key result of this work, a Bernstein-type inequality for stationary geometrically (time-reversed) $\mathcal{C}$-mixing process.

Theorem 3.1. *Let $\mathcal{Z} := (Z_n)_{n\geq 0}$ be a Z-valued stationary geometrically (time-reversed) $\mathcal{C}$-mixing process on $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\geq 0}$ as in (6), $\|\cdot\|_{\mathcal{C}}$ be defined by* (1) *for some semi-norm $\|\cdot\|$ satisfying (3), and $P := \mu_{Z_0}$. Moreover, let $h \in \mathcal{C}(Z)$ with $\mathbb{E}_P h = 0$ and assume that there exist some $A > 0$, $B > 0$, and $\sigma \geq 0$ such that $\|h\| \leq A$, $\|h\|_\infty \leq B$, and $\mathbb{E}_P h^2 \leq \sigma^2$. Then, for all $\varepsilon > 0$ and all*

$$n \geq n_0 := \max\left\{\min\left\{m\geq 3 : m^2 \geq \frac{808c(3A+B)}{B} \textit{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \geq 4\right\}, e^{\frac{3}{b}}\right\}, \tag{13}$$

we have

$$\mu\left(\left\{\omega\in\Omega : \frac{1}{n}\sum_{i=1}^{n} h\circ Z_i \geq \varepsilon\right\}\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{8(\log n)^{\frac{2}{\gamma}}(\sigma^2+\varepsilon B/3)}\right), \tag{14}$$

or alternatively, for all $n \geq n_0$ and $\tau > 0$, we have

$$\mu\left(\left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^{n} h(Z_i(\omega)) \geq \sqrt{\frac{8(\log n)^{\frac{2}{\gamma}}\sigma^2\tau}{n}} + \frac{8(\log n)^{\frac{2}{\gamma}}B\tau}{3n}\right\}\right) \leq 2e^{-\tau}. \tag{15}$$

Note that besides the additional logarithmic factor $4(\log n)^{\frac{2}{\gamma}}$ and the constant 2 in front of the exponential, (14) coincides with Bernstein's classical inequality for i.i.d. processes.

In the remainder of this section, we compare Theorem 3.1 with some other concentration inequalities for non-i.i.d. processes $\mathcal{Z}$. Here, $\mathcal{Z}$ is real-valued and h is the identity map if not specified otherwise.

Example 3.2. Theorem 2.3 in [4] shows that smooth expanding systems on $[0,1]$ have exponential decay of correlations (7). Moreover, if, for such expanding systems, the transformation T is Lipschitz continuous and satisfies the conditions at the end of Section 4 in [18] and the ergodic measure μ satisfies [18, condition (4.8)], then [18, Theorem 2] shows that for all $\varepsilon \geq 0$ and $n \geq 1$, the left-hand side of (14) is bounded by

$$\exp\left(-\frac{\varepsilon^2 n}{C}\right)$$

where C is some constant independent of n. The same result has been proved in [15, Theorem III.1] as well. Obviously, this is a Hoeffding-type bound instead of a Bernstein-type one. Hence, it is always larger than ours if the denominator of the exponent in (14) is smaller than C.

Example 3.3. For dynamical systems with exponentially decreasing $\tilde{\phi}$-coefficients, see [47, condition (3.1)], [47, Theorem 3.1] provides a Bernstein-type inequality for 1-Lipschitz functions $h : Z \to [-1/2, 1/2]$ w.r.t. some metric d on Z, in which the left-hand side of (14) is bounded by

$$\exp\left(-\frac{C\varepsilon^2 n}{\sigma^2 + \varepsilon \log f(n)}\right) \tag{16}$$

for some constant C independent of n and $f(n)$ being some function monotonically increasing in n. Note that modulo the logarithmic factor $\log f(n)$ the bound (16) is the same as the one for i.i.d. processes. Moreover, if $f(n)$ grows polynomially, cf. [47, Section 3.3], then (16) has the same asymptotic behaviour as our bound. However, geometrically $\mathcal{C}$-mixing is weaker than Condition (3.1) in [47]: Indeed, the required exponential form of Condition (3.1) in [47], i.e.

$$\sup_{k\geq 0} \tilde{\phi}(\mathcal{A}_0^k, \mathbf{Z}_{k+n}^{k+2n-1}) := \sup_{k\geq 0} \sup_{f\in\mathcal{F}^n} \left\|\mathbb{E}\big(f(\mathbf{Z}_{k+n}^{k+2n-1})\big|\mathcal{A}_0^k\big) - \mathbb{E}f(\mathbf{Z}_{k+n}^{k+2n-1})\right\|_\infty \leq c \cdot e^{-bn}$$

for some $c, b > 0$ and all $n \geq 1$, where $\mathbf{Z}_{k+n}^{k+2n-1} := (Z_{k+n}, \ldots, Z_{k+2n-1})$ and $\mathcal{F}^n$ is the set of 1-Lipschitz functions $f : Z^n \to [-\frac{1}{2}, \frac{1}{2}]$ w.r.t. the metric $d^n(x,y) := \frac{1}{n}\sum_{i=1}^n d(x_i, y_i)$, implies

$$\sup_{k\geq 0} \sup_{f\in\mathcal{F}} \left\|\mathbb{E}\big(f(Z_{k+n})\big|\mathcal{A}_0^k\big) - \mathbb{E}f(Z_{k+n})\right\|_\infty \leq c \cdot ne^{-bn} \leq c \cdot e^{-\tilde{b}n}$$

for some $c, \tilde{b} > 0$ and all $n \geq 1$, where $\mathcal{F}$ is the set of 1-Lipschitz functions $f : Z \to [-\frac{1}{2}, \frac{1}{2}]$ w.r.t. the metric d. In other words, processes satisfying Condition (3.1) in [47] are $\tilde{\phi}$-mixing, see (10), which is stronger than geometrically $\mathcal{C}$-mixing, see again Figure 1. Moreover, our result holds for all $\gamma > 0$, while [47] only considers the case $\gamma = 1$.

Example 3.4. For an α-mixing sequence of centered and bounded random variables satisfying $\alpha(n) \leq c\exp(-bn^\gamma)$ for some constants $b > 0$, $c \geq 0$, and $\gamma > 0$, [33, Theorem 4.3] bounds the left-hand side of (14) by

$$(1 + 4e^{-2}c)\exp\left(-\frac{3\varepsilon^2 n^{(\gamma)}}{6\sigma^2 + 2\varepsilon B}\right) \quad \text{with } n^{(\gamma)} \asymp n^{\frac{\gamma}{\gamma+1}} \tag{17}$$

for all $n \geq 1$ and all $\varepsilon > 0$. In general, this bound and our result are not comparable, since not every α-mixing process satisfies (7) and conversely, not every process satisfying (7) is necessarily α-mixing, see Figure 2. Nevertheless, for ϕ-mixing processes, it is easily seen that this bound is always worse than ours for a fixed $\gamma > 0$, if n is large enough.

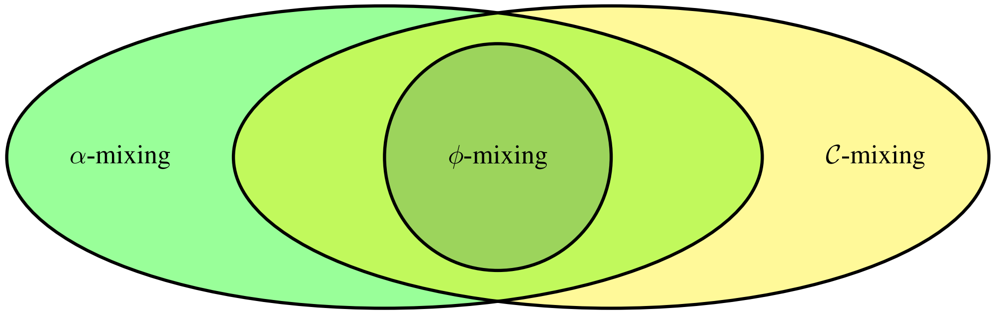

Figure 2: Relationship between α-, ϕ-, and $\mathcal{C}$-mixing processes

Example 3.5. For an α-mixing stationary sequence of centered and bounded random variables satisfying $\alpha(n) \leq \exp(-2cn)$ for some $c > 0$, [32, Theorem 2] bounds the left-hand side of (14) by

$$\exp\left(-\frac{C\varepsilon^2 n}{v^2 + \varepsilon B(\log n)^2 + n^{-1}B^2}\right), \tag{18}$$

where $C > 0$ is some constant and

$$v^2 := \sigma^2 + 2\sum_{2\leq i\leq n} |\text{cov}(X_1, X_i)|. \tag{19}$$

By applying the covariance inequality for α-mixing processes, see [16, the corollary to Lemma 2.1], we obtain $v^2 \leq C_\delta \|X_1\|_{2+\delta}^2$ for an arbitrary $\delta > 0$ and a constant C_δ only depending on δ. If the additional $\delta > 0$ is ignored, (18) has therefore the same asymptotic behavior as our bound. In general, however, the additional δ does influence the asymptotic behavior. For example, the oracle inequality we obtain in the next section would be slower by a factor of n^ξ, where $\xi > 0$ is arbitrary, if we used (18) instead. Finally, note that in general the bound (18) and ours are not comparable, see again Figure 2.

In particular, Inequality (18) can be applied to geometrically ϕ-mixing processes with $\gamma = 1$. By using the covariance inequality (1.1) for ϕ-mixing processes in [16], we can bound v^2 defined as in (19) by $C\sigma^2$ with some constant C independent of n. Modulo the term $n^{-1}B$ in the denominator, the bound (18) coincides with ours for geometrically ϕ-mixing processes with $\gamma = 1$. However, our bound also holds for such processes with $\gamma \in (0, 1)$.

Example 3.6. For stationary, geometrically α-mixing Markov chains with centered and bounded random variables, [1] bounds the left-hand side of (14) by

$$\exp\left(-\frac{n\varepsilon^2}{\tilde{\sigma}^2 + \varepsilon B \log n}\right), \tag{20}$$

where $\tilde{\sigma}^2 = \lim_{n\to\infty} \frac{1}{n}\text{Var}\sum_{i=1}^n X_i$. By a similar argument as in Example 3.5 we obtain

$$\text{Var}\sum_{i=1}^n X_i = n\sigma^2 + 2\sum_{1\leq i<j\leq n} |\text{cov}(X_i, X_j)| \leq n\sigma^2 + \tilde{C}_\delta n \|X_1\|_{2+\delta}^2$$

for an arbitrary $\delta > 0$ and a constant $\tilde{C}_\delta$ depending only on δ. Consequently we conclude that modulo some arbitrary small number $\delta > 0$ and the logarithmic factor $\log n$ instead of $(\log n)^2$, the bound (20) coincides with ours. Again, this bound and our result are not comparable, see Figure 2.

Example 3.7. For stationary, weakly dependent processes of centered and bounded random variables with $|\text{cov}(X_1, X_n)| \leq c \cdot \exp(-bn)$ for some $c, b > 0$ and all $n \geq 1$, [26, Theorem 2.1] bounds the left-hand side of (14) by

$$\exp\left(-\frac{\varepsilon^2 n}{C_1 + C_2\varepsilon^{5/3}n^{2/3}}\right) \tag{21}$$

where C_1 is some constant depending on c and b, and C_2 is some constant depending on c, b, and B. Note that the denominator in (21) is at least C_1, and therefore the bound (21) is more of Hoeffding type.

4 Applications to Statistical Learning

In this section, we apply the Bernstein inequality from the last section to deduce oracle inequalities for some widely used learning methods and observations generated by a geometrically $\mathcal{C}$-mixing processes. More precisely, in Subsection 4.1, we recall some basic concepts of statistical learning and formulate an oracle inequality for learning methods that are based on (regularized) empirical risk minimization. Then, in the Subsection 4.2, we illustrate this oracle inequality by deriving the learning rates for SVMs. Finally, in Subsection 4.3, we present an oracle inequality for forecasting of dynamical systems.

4.1 Oracle inequality for CR-ERMs

In this section, let X always be a measurable space if not mentioned otherwise and $Y \subset \mathbb{R}$ always be a closed subset. Recall that in the (supervised) statistical learning, our aim 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. To evaluate the quality of such functions f, we need a loss function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ that is measurable. Following [42, Definition 2.22], we say that a loss L can be clipped at $M > 0$, if, for all $(x, y, t) \in X \times Y \times \mathbb{R}$, we have

$$L(x, y, \widehat{t}\,) \leq L(x, y, t), \tag{22}$$

where $\widehat{t}$ denotes the clipped value of t at $\pm M$, that is $\widehat{t} := t$ if $t \in [-M, M]$, $\widehat{t} := -M$ if $t < -M$, $\widehat{t} := M$ if $t > M$. Various often used loss functions can be clipped. For example, if $Y := \{-1, 1\}$ and L is a convex, margin-based loss represented by $\varphi : \mathbb{R} \to [0, \infty)$, that is $L(y, t) = \varphi(yt)$ for all $y \in Y$ and $t \in \mathbb{R}$, then L can be clipped, if and only if φ has a global minimum, see [42, Lemma 2.23]. 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. Moreover, if $Y := [-M, M]$ and L is a convex, distance-based loss represented by some $\psi : \mathbb{R} \to [0, \infty)$, that is $L(y, t) = \psi(y - t)$ for all $y \in Y$ and $t \in \mathbb{R}$, then L can be clipped whenever $\psi(0) = 0$, see again [42, Lemma 2.23]. In particular, the least squares loss

$$L(y, t) = (y - t)^2 \tag{23}$$

and the τ-pinball loss

$$L_\tau(y, t) := \psi(y - t) = \begin{cases} -(1 - \tau)(y - t), & \text{if } y - t < 0 \\ \tau(y - t), & \text{if } y - t \geq 0 \end{cases} \tag{24}$$

used for quantile regression can be clipped, if the space of labels Y is bounded.

Now we summarize assumptions on the loss function L that will be used throughout this work.

Assumption 4.1. *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 $L(x, y, t) \leq 1$ and locally Lipschitz continuous, that is,*

$$|L(x, y, t) - L(x, y, t')| \leq |t - t'| \,. \tag{25}$$

Here both inequalites are supposed to hold for all $(x, y) \in X \times Y$ and $t, t' \in [-M, M]$. Note that the former assumption can typically be enforced by scaling.

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, see also [42, Definitions 2.2 & 2.3].

Definition 4.2. *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

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

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.

Informally, the goal of learning from a training set $D \in (X \times Y)^n$ 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 4.3. *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 $f_D : X \to \mathbb{R}$.*

Let us now describe the learning algorithms we are interested in. 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 pre-compact with respect to the supremum norm $\| \cdot \|_\infty$. Since $\mathcal{F}$ can be infinite, we need to recall the following, classical concept, which will enable us to approximate infinite $\mathcal{F}$ by finite subsets.

Definition 4.4. *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 pre-compact, and hence for all $\varepsilon > 0$, the covering number $\mathcal{N}(\mathcal{F}, \| \cdot \|_\infty, \varepsilon)$ is finite.

In order to introduce our generic learning algorithms, 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 the $X \times Y$-valued process $\mathcal{Z} = (Z_i)_{i \geq 1}$. Furthermore, we write $D_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$, 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 to the data set D. Finally, 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 can now introduce the class of learning methods we are interested in, see also [42, Definition 7.18].

Definition 4.5. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss that can be clipped at some $M > 0$, $\mathcal{F}$ be a hypothesis set, that is, a set of measurable functions $f : X \to \mathbb{R}$, with $0 \in \mathcal{F}$, and Υ be a regularizer on $\mathcal{F}$, that is, a function $\Upsilon : \mathcal{F} \to [0, \infty)$ with $\Upsilon(0) = 0$. Then, for $\delta \geq 0$, a learning method whose decision functions $f_{D_n,\Upsilon} \in \mathcal{F}$ satisfy*

$$\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,D_n}(\widehat{f}_{D_n,\Upsilon}) \leq \inf_{f\in\mathcal{F}} \left(\Upsilon(f) + \mathcal{R}_{L,D_n}(f)\right) + \delta \tag{26}$$

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.

Moreover, in the case $\delta = 0$, we simply speak of clipped regularized empirical risk minimization (CR-ERM).

Note that on the right-hand side of (26) the unclipped loss is considered, and hence CR-ERMs do not necessarily minimize the regularized clipped empirical risk $\Upsilon(\cdot) + \mathcal{R}_{L,D_n}(\widehat{\cdot})$. 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 (26) the clipped function is considered. However, if we have a minimizer of the unclipped regularized risk, then it automatically satisfies (26). As an example of CR-ERMs, SVMs will be discussed in Section 4.2.

Before we present the oracle inequality for δ-CR-ERMs, we need to introduce a few more notations. Let $\mathcal{F}$ be a hypothesis set in the sense of Definition 4.5. For

$$r^* := \inf_{f\in\mathcal{F}} \Upsilon(f) + \mathcal{R}_{L,P}(\widehat{f}\,) - \mathcal{R}^*_{L,P} \tag{27}$$

and $r > r^*$, we write

$$\mathcal{F}_r := \left\{ f \in \mathcal{F} : \Upsilon(f) + \mathcal{R}_{L,P}(\widehat{f}\,) - \mathcal{R}^*_{L,P} \leq r \right\}. \tag{28}$$

Then we have $r^* \leq 1$, since $L(x, y, 0) \leq 1$, $0 \in \mathcal{F}$, and $\Upsilon(0) = 0$. Furthermore, we assume that we have a monotonic decreasing sequence $(A_r)_{r\in(0,1]}$ such that

$$\|L \circ \widehat{f}\| \leq A_r \quad \text{for all } f \in \mathcal{F}_r \text{ and } r \in (0, 1]\,, \tag{29}$$

where $\|\cdot\|$ is a semi-norm satisfying (3). Because of the definition (28), it is easily to conclude that $\|L \circ \widehat{f}\| \leq A_1$ for all $f \in \mathcal{F}_r$ and $r \in (0, 1]$. Finally, we assume that there exists a function $\varphi : (0, \infty) \to (0, \infty)$ and a $p \in (0, 1]$ such that, for all $r > 0$ and $\varepsilon > 0$, we have

$$\ln \mathcal{N}(\mathcal{F}_r, \|\cdot\|_\infty, \varepsilon) \leq \varphi(\varepsilon) r^p. \tag{30}$$

Note that there are actually many hypothesis sets satisfying Assumption (30), see [23, Section 4] for some examples.

Now the oracle inequality for CR-ERMs reads as follows:

Theorem 4.6. *Let $\mathcal{Z} := (Z_n)_{n\geq0}$ be a Z-valued stationary geometrically (time-reversed) $\mathcal{C}$-mixing process on $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\geq0}$ as in (6), $\|\cdot\|_\mathcal{C}$ be defined by (1) for some semi-norm $\|\cdot\|$ satisfying (3), and $P := \mu_{Z_0}$. Moreover, let L be a loss satisfying Assumption 4.1. In addition, assume that there exist a Bayes decision function $f^*_{L,P}$ and constants $\vartheta \in [0, 1]$ and $V \geq 1$ such that*

$$\mathbb{E}_P(L \circ \widehat{f} - L \circ f^*_{L,P})^2 \leq V \cdot \left(\mathbb{E}_P(L \circ \widehat{f} - L \circ f^*_{L,P})\right)^\vartheta, \qquad f \in \mathcal{F}, \tag{31}$$

where $\mathcal{F}$ is a hypothesis set with $0 \in \mathcal{F}$. We define r^*, $\mathcal{F}_r$, and A_r by (27), (28), and (29), respectively and assume that (30) 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 $A_0, A^* \geq 0$*,* $B_0 \geq 1$ *be constants such that* $\|L \circ f_0\| \leq A_0$*,* $\|L \circ \widehat{f}_0\| \leq A_0$*,* $\|L \circ f^*_{L,P}\| \leq A^*$ *and* $\|L \circ f_0\|_\infty \leq B_0$*. Then, for all fixed* $\varepsilon > 0$*,* $\delta \geq 0$*,* $\tau \geq 1$*, and*

$$n \geq n_0^* := \max\left\{\min\left\{m \geq 3 : m^2 \geq K \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \geq 4\right\}, e^{\frac{3}{b}}\right\} \tag{32}$$

with $K = 1212c(4A_0 + A^* + A_1 + 1)$*, and* $r \in (0,1]$ *satisfying*

$$r \geq \max\left\{\left(\frac{c_V (\log n)^{\frac{2}{\gamma}}(\tau + \varphi(\varepsilon/2)2^p r^p)}{n}\right)^{\frac{1}{2-\vartheta}}, \frac{20(\log n)^{\frac{2}{\gamma}} B_0 \tau}{n}, r^*\right\} \tag{33}$$

with $c_V := 512(12V + 1)/3$*, every learning method defined by (26) satisfies with probability* μ *not less than* $1 - 16e^{-\tau}$*:*

$$\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P} < 2\Upsilon(f_0) + 4\mathcal{R}_{L,P}(f_0) - 4\mathcal{R}^*_{L,P} + 4r + 5\varepsilon + 2\delta. \tag{34}$$

Let us briefly discuss the variance bound (31). For example, if $Y = [-M, M]$ and L is the least squares loss, then it is well-known that (31) is satisfied for $V := 16M^2$ and $\vartheta = 1$, see e.g. [42, Example 7.3]. Moreover, under some assumptions on the distribution P, [43] established a variance bound of the form (31) for the pinball loss used for quantile regression. In addition, for the hinge loss, (31) is satisfied for $\vartheta := q/(q + 1)$, if Tsybakov's noise assumption [46] holds for q, see [42, Theorem 8.24]. Finally, based on [9], [40] established a variance bound with $\vartheta = 1$ for the earlier mentioned clippable modifications of strictly convex, twice continuously differentiable margin-based loss functions.

One might wonder, why the constants A_0 and B_0 are necessary in Theorem 4.6, since it appears to add further complexity. However, a closer look reveals that the constants A_1 and B are the bounds for functions of the form $L \circ \widehat{f}$, while A_0 and B_0 are valid for the function $L \circ f_0$ for an *unclipped* $f_0 \in \mathcal{F}$. Since we do not assume that all $f \in \mathcal{F}$ satisfy $\widehat{f} = f$, we conclude that in general A_0 and B_0 are necessary.

The following lemma shows that the required bounds on $\|L \circ f\|$ do hold for specific loss functions, if $\mathcal{C} = \mathrm{Lip}$ and the involved functions $f \in \mathcal{F}$ are Lipschitz, too.

Lemma 4.7. *Let* (X, d) *be a metric space,* $Y \subset [-M, M]$ *with* $M > 0$*. Moreover, let* $f : X \to \mathbb{R}$ *be a bounded, Lipschitz continuous function. Then the following statements hold true:*

(i) For the least square loss L*, see (23), we have*

$$|L \circ f|_1 \leq 2\sqrt{2}\,(M + \|f\|_\infty)\,(1 + |f|_1).$$

(ii) For the τ*-pinball loss* L*, see (24), we have*

$$|L \circ f|_1 \leq \sqrt{2}(1 + |f|_1).$$

4.2 Learning rates for SVMs

Let us begin by briefly recalling SVMs, see [42] for details. 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

$$f_{D_n,\lambda} = \arg\min_{f \in H} \left(\lambda \|f\|_H^2 + \mathcal{R}_{L,D_n}(f)\right). \tag{35}$$

In particular, SVMs using the least-squares loss (23) are called least-squares SVMs (LS-SVMs), while SVMs using the τ-pinball loss (24) are called SVMs for quantile regression.

Note that SVM decision functions (35) satisfy (26) for the regularizer $\Upsilon := \lambda\|\cdot\|_H^2$ and $\delta := 0$. In other words, SVMs are CR-ERMs. Consequently we can use the oracle inequality in Theorem 4.6 to derive the learning rates for SVMs.

Assumption 4.1 implies that

$$\lambda\|f_{D_n,\lambda}\|_H^2 \leq \lambda\|f_{D_n,\lambda}\|_H^2 + \mathcal{R}_{L,D_n}(f) = \min_{f\in H}\left(\lambda\|f\|_H^2 + \mathcal{R}_{L,D_n}(f)\right) \leq \mathcal{R}_{L,D_n}(0) \leq 1.$$

In other words, for a fix $\lambda > 0$, we have

$$f_{D_n,\lambda} \in \lambda^{-1/2} B_H, \tag{36}$$

where B_H denotes the closed unit ball of the RKHS H.

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

$$k_\sigma(x, x') := \exp\left(-\frac{\|x - x'\|_2^2}{\sigma^2}\right), \qquad x, x' \in X,$$

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 [44]. The entropy numbers for Gaussian kernels [42, Theorem 6.27] and the equivalence of covering and entropy numbers [42, Lemma 6.21] yield that

$$\ln \mathcal{N}(B_{H_\sigma}, \|\cdot\|_\infty, \varepsilon) \leq a\sigma^{-d}\varepsilon^{-2p}, \qquad \varepsilon > 0, \tag{37}$$

for some constants $a > 0$ and $p \in (0, 1)$.

Because of (36), we can choose the hypothesis set as $\mathcal{F} = \lambda^{-1/2} B_{H_\sigma}$. Then the definition (28) implies that $\mathcal{F}_r \subset r^{1/2}\lambda^{-1/2} B_{H_\sigma}$ and consequently we have

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

and thus, for the function φ in Theorem 4.6, we can choose

$$\varphi(\varepsilon) := a\sigma^{-d}\lambda^{-p}\varepsilon^{-2p}. \tag{38}$$

Now, with some additional assumptions below, we can use the oracle inequality in Theorem 4.6 to derive the learning rates for the SVMs using Gaussian kernels. In the following, $B_{2s,\infty}^t$ denotes the usual Besov space with the smoothness parameter t, more details see [21, Section 2].

Theorem 4.8 (Least Square Regression with Gaussian Kernels)**.** *Let $Y := [-M, M]$ for $M > 0$, and P be a distribution on $\mathbb{R}^d \times Y$ such that $X := \operatorname{supp}P_X \subset B_{\ell_2^d}$ is a bounded domain with $\mu(\partial X) = 0$, where $B_{\ell_2^d}$ denotes the closed unit ball of d-dimensional Euclidean space ℓ_2^d. Furthermore, let P_X be absolutely continuous w.r.t. the Lebesgue measure μ on X with associated density $g : \mathbb{R}^d \to \mathbb{R}$ such that $g \in L_q(X)$ for some $q \geq 1$. Moreover, let $f_{L,P}^* : \mathbb{R}^d \to \mathbb{R}$ be a Bayes decision function such that $f_{L,P}^* \in L_2(\mathbb{R}^d) \cap \operatorname{Lip}(\mathbb{R}^d)$ as well as $f_{L,P}^* \in B_{2s,\infty}^t$ for some $t \geq 1$ and $s \geq 1$ with $\frac{1}{q} + \frac{1}{s} = 1$. Then, for all $\xi > 0$, the LS-SVM using Gaussian RKHS H_σ and*

$$\lambda_n = n^{-1} \quad \textit{and} \quad \sigma_n = n^{-\frac{1}{2t+d}}, \tag{39}$$

learns with rate

$$n^{-\frac{2t}{2t+d}+\xi}. \tag{40}$$

It turns out that, modulo the arbitrarily small $\xi > 0$, these learning rates are optimal, see e.g. [45, Theorem 13] or [22, Theorem 3.2].

To achieve these rates, however, we need to set λ_n and σ_n as in (39), which in turn requires us to know t. Since in practice we usually do not know these values nor their existence, we can use the training/validation approach TV-SVM, see e.g. [42, Chapters 6.5, 7.4, 8.2], to achieve the same rates adaptively, i.e. without knowing t. To this end, let $\Lambda := (\Lambda_n)$ and $\Sigma := (\Sigma_n)$ be sequences of finite subsets $\Lambda_n, \Sigma_n \subset (0,1]$ such that Λ_n is an ϵ_n-net of $(0,1]$ and Σ_n is an δ_n-net of $(0,1]$ with $\epsilon_n \leq n^{-1}$ and $\delta_n \leq n^{-\frac{1}{2+d}}$. Furthermore, assume that the cardinalities $|\Lambda_n|$ and $|\Sigma_n|$ grow polynomially in n. For a data set $D := ((x_1, y_1), \ldots, (x_n, y_n))$, we define

$$D_1 := ((x_1, y_1), \ldots, (x_m, y_m))$$
$$D_2 := ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n))$$

where $m := \left\lfloor \frac{n}{2} \right\rfloor + 1$ and $n \geq 4$. We will use D_1 as a training set by computing the SVM decision functions

$$f_{\mathrm{D}_1,\lambda,\sigma} := \arg\min_{f \in H_\sigma} \lambda \|f\|^2_{H_\sigma} + \mathcal{R}_{L,\mathrm{D}_1}(f), \qquad (\lambda,\sigma) \in \Lambda_n \times \Sigma_n$$

and use D_2 to determine (λ, σ) by choosing a $(\lambda_{\mathrm{D}_2}, \sigma_{\mathrm{D}_2}) \in \Lambda_n \times \Sigma_n$ such that

$$\mathcal{R}_{L,\mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\sigma_{\mathrm{D}_2}}\right) = \min_{(\lambda,\sigma)\in\Lambda_n\times\Sigma_n} \mathcal{R}_{L,\mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_1,\lambda,\sigma}\right) .$$

Then, analogous to the proof of Theorem 3.3 in [21] we can show that for all $\zeta > 0$ and $\xi > 0$, the TV-SVM producing the decision functions $f_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\sigma_{\mathrm{D}_2}}$ with the above learning rates (40).

The following remark discusses learning rates for SVMs for quantile regression. For more information on such SVMs we refer to [21, Section 4].

Remark 4.9 (Quantile Regression with Gaussian Kernels). Let $Y := [-1, 1]$, and P be a distribution on $\mathbb{R}^d \times Y$ such that $X := \mathrm{supp} P_X \subset B_{\ell_2^d}$ be a domain. Furthermore, we assume that, for P_X-almost all $x \in X$, the conditional measure $P(\cdot|x)$ is absolutely continuous w.r.t. the Lebesgue measure on Y and the conditional density $h(\cdot, x)$ of $P(\cdot|x)$ is bounded from 0 and ∞, see also [21, Example 4.5]. Moreover, let P_X be absolutely continuous w.r.t. 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 $f^*_{\tau,P} \in L_2(\mathbb{R}^d) \cap \mathrm{Lip}(\mathbb{R}^d)$. 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 [43, Theorem 2.8] yields a variance bound of the form

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

for all $f : X \to \mathbb{R}$, where V is a suitable constant and L_τ is the τ-pinball loss. Similar arguments to Theorem 4.8 shows that the essentially optimal learning rate (40) can be achieved as well. Note that the rate (40) is for the excess L_τ-risk, but since [43, Theorem 2.7] shows

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

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

4.3 Forecasting of dynamical systems

In this section, we proceed with the study of the forecasting problem of dynamical systems considered in [41]. First, let us recall some basic notations and assumptions. Let Ω be a compact subset of $\mathbb{R}^d$, $(\Omega, \mathcal{A}, \mu, T)$ be a dynamical system, and $S_0 \in \Omega$ be a random variable describing the true but unknown state at time 0. Moreover, for $E > 0$, assume that all observations of the stochastic process described by the sequence $\mathcal{T} := (T^n)_{n\geq 0}$ are additively corrupted by some i.i.d., $[-E,E]^d$-valued noise process $\mathcal{E} = (\varepsilon_n)_{n\geq 0}$ defined on the probability space $(\Theta, \mathcal{C}, \nu)$ which is (stochastically) independent of $\mathcal{T}$. It follows that all possible observations of the system at time $n \geq 0$ are of the form

$$X_n = T^n(S_0) + \varepsilon_n. \tag{41}$$

In other words, the process that generates the noisy observations (41) is $(T^n(S_0)+\varepsilon_n)_{n\geq 0}$. In particular, a sequence of observations $(X_0, \dots, X_n)$ generated by this process is of the form (41) for a conjoint initial state S_0.

Now, given an observation of the process $\mathcal{T} := (T^n)_{n\geq 0}$ at some arbitrary time, our goal is to forecast the next *observable* state. To do so, we will use the training set

$$\begin{aligned} \boldsymbol{D}_n &= ((X_0, X_1), \dots, (X_{n-1}, X_n)) \\ &= \big((S_0 + \varepsilon_0, T(S_0) + \varepsilon_1), \dots, \big(T^{n-1}(S_0) + \varepsilon_{n-1}, T^n(S_0) + \varepsilon_n\big)\big) \end{aligned}$$

whose input/output pairs are consecutive observable states. In other words, our goal is to use $\boldsymbol{D}_n$ to build a forecaster

$$\boldsymbol{f}_{\boldsymbol{D}_n} : \mathbb{R}^d \to \mathbb{R}^d$$

whose average forecasting performance on future noisy observations is as small as possible. In order to render this goal, we will use the forecaster

$$\boldsymbol{f}_{\boldsymbol{D}_n} := \left(f_{\boldsymbol{D}_n^{(1)}}, \dots, f_{\boldsymbol{D}_n^{(d)}}\right), \tag{42}$$

where $f_{\boldsymbol{D}_n^{(j)}}$ is the forecaster obtained by using the training set

$$\boldsymbol{D}_n^{(j)} := ((X_0, \pi_j(X_1)), \dots, (X_{n-1}, \pi_j(X_n)))$$

which is obtained by projecting the output variable of $\boldsymbol{D}_n$ onto its jth-coordinate via the coordinate projection $\pi_j : \mathbb{R}^d \to \mathbb{R}$.

In other words, we build the forecaster $\boldsymbol{f}_{\boldsymbol{D}_n}$ by training separately d different decision functions on the training sets $\boldsymbol{D}_n^{(1)}, \dots, \boldsymbol{D}_n^{(d)}$. These problems can be considered as the (supervised) statistical learning problems formulated in Subsection 4.1 with the help of the following Notations.

For $E > 0$ and a fixed $j \in \{1, \dots, d\}$, we write $X := K + [-E,E]^d$, $Y := \pi_j(X)$ and $Z := X \times Y$. Moreover, we define the $X \times Y$-valued process $\mathcal{Z} = (Z_n)_{n\geq 0} = (X_n, Y_n)_{n\geq 0}$ on $(K \times \Theta, \mathcal{B}\otimes\mathcal{C}, \mu\otimes\nu)$ by $X_n := T^n + \varepsilon_n$ and $Y_n := \pi_j(T^{n+1} + \varepsilon_{n+1})$. In addition, we write $P := (\mu\otimes\nu)_{(X_0,Y_0)}$. Obviously, if the stochastic process $\mathcal{T}$ is $\mathcal{C}$-mixing and the noise process $\mathcal{E}$ is i.i.d, then the stochastic processes

$$\mathcal{Z} = (X_n, Y_n)_{n\geq 0} = (T^n(S_0) + \varepsilon_n, \pi_j(T^{n+1}(S_0) + \varepsilon_{n+1}))_{n\geq 0}$$

is $\mathcal{C}$-mixing as well.

To formulate the oracle inequality for our original d-dimensional problem, we need to introduce the following concepts. Firstly, for the decision function $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d$, it is necessary to introduce a loss function $\boldsymbol{L} : \mathbb{R}^d \to [0, \infty)$ such that

$$\boldsymbol{L}\left(X_i - \boldsymbol{f}(X_{i-1})\right) = \boldsymbol{L}\left(T^i(S_0) + \varepsilon_i - \boldsymbol{f}(T^{i-1}(S_0) + \varepsilon_{i-1})\right)$$

gives a value for the discrepancy between the forecast $\boldsymbol{f}(T^{i-1}(S_0)+\varepsilon_{i-1})$ and the observation of the next state $T^i(S_0)+\varepsilon_i$. We say that a loss $\boldsymbol{L}:\mathbb{R}^d\to[0,\infty)$ can be *clipped* at $M>0$, if, for all $\boldsymbol{t}=(t_1,\dots,t_d)\in\mathbb{R}^d$, we have $\boldsymbol{L}(\widehat{\boldsymbol{t}})\le\boldsymbol{L}(\boldsymbol{t})$, where $\widehat{\boldsymbol{t}}=(\widehat{t}_1,\dots,\widehat{t}_d)$ denotes the clipped value of $\boldsymbol{t}$ at $\{\pm M\}^d$. Moreover, the loss function $\boldsymbol{L}:\mathbb{R}^d\to[0,\infty)$ is called *separable*, if there exists a distance-based loss $L:X\times Y\times\mathbb{R}\to[0,\infty)$ such that its representing function $\psi:\mathbb{R}\to[0,\infty)$ has a unique global minimum at 0 and satisfies

$$\boldsymbol{L}(\boldsymbol{r})=\psi(r_1)+\cdots+\psi(r_d),\qquad \boldsymbol{r}=(r_1,\dots,r_d)\in\mathbb{R}^d. \tag{43}$$

In our problem-setting, the average forecasting performance is given by the $\boldsymbol{L}$-risk

$$\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f}):=\iint\boldsymbol{L}\left(T(x)+\varepsilon_1-\boldsymbol{f}(x+\varepsilon_0)\right)\,\nu(d\varepsilon)\,\mu(dx), \tag{44}$$

where $\varepsilon=(\varepsilon_i)_{i\ge0}$ and $\boldsymbol{P}:=\nu\otimes\mu$. Naturally, the smaller the risk, the better the forecaster is. Hence, we ideally would like to have a forecaster $\boldsymbol{f}^*_{\boldsymbol{L},\boldsymbol{P}}:\mathbb{R}^d\to\mathbb{R}^d$ that attains the minimal $\boldsymbol{L}$-risk

$$\mathcal{R}^*_{\boldsymbol{L},\boldsymbol{P}}:=\inf\left\{\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f})|\boldsymbol{f}:\mathbb{R}^d\to\mathbb{R}^d\text{ measurable}\right\}. \tag{45}$$

The assumption (43) then implies $\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f})=\sum_{j=1}^d\mathcal{R}_{L,P}(f_{\boldsymbol{D}_n{}^{(j)}})$ and

$$\mathcal{R}_{\boldsymbol{L},\mathbf{D}_n}(\boldsymbol{f}_{\boldsymbol{D}_n})=\sum_{j=1}^d\mathcal{R}_{L,\mathbf{D}_n^{(j)}}(f_{\boldsymbol{D}_n^{(j)}}),$$

where $\mathbf{D}_n$, $\mathbf{D}_n^{(j)}$ are the empirical measures associated to $\boldsymbol{D}_n$, $\boldsymbol{D}_n^{(j)}$ respectively.

Finally, let $\boldsymbol{L}:\mathbb{R}^d\to[0,\infty)$ be a clippable loss and $\mathcal{F}$ be a hypothesis set with $0\in\mathcal{F}$. A regularizer $\boldsymbol{\Upsilon}$ on $\mathcal{F}^d$, that is, a function $\boldsymbol{\Upsilon}:\mathcal{F}^d\to[0,\infty)$, is also said to be *separable*, if there exists a regularizer Υ on $\mathcal{F}$ with $\Upsilon(0)=0$ such that $\boldsymbol{\Upsilon}(\boldsymbol{f})=\sum_{j=1}^d\Upsilon(f_j)$ for $\boldsymbol{f}=(f_1,\dots,f_d)$. Then, for $\delta\ge0$, a learning method whose decision functions $\boldsymbol{f}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}}\in\mathcal{F}^d$ satisfy

$$\boldsymbol{\Upsilon}(\boldsymbol{f}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}})+\mathcal{R}_{\boldsymbol{L},\boldsymbol{D}_n}(\widehat{\boldsymbol{f}}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}})<\inf_{\boldsymbol{f}\in\mathcal{F}^d}\left(\boldsymbol{\Upsilon}(\boldsymbol{f})+\mathcal{R}_{\boldsymbol{L},\boldsymbol{D}_n}(\boldsymbol{f})\right)+d\delta \tag{46}$$

for all $n\ge1$ and $\boldsymbol{D}_n\in(X\times Y)^{dn}$ is called $d\delta$-approximate clipped regularized empirical risk minimization ($d\delta$-CR-ERM) with respect to $\boldsymbol{L}$, $\mathcal{F}^d$, and $\boldsymbol{\Upsilon}$.

With all these preparations above, the oracle inequality for geometrically $\mathcal{C}$-mixing dynamical systems with i.i.d noise processes, can be stated as following:

Theorem 4.10. *Let $\Omega\subset\mathbb{R}^d$ be compact and $(\Omega,\mathcal{A},\mu,T)$ be a dynamical system. Suppose that the stationary stochastic process $\mathcal{T}:=(T^n)_{n\ge0}$ is geometrically time-reversed $\mathcal{C}$-mixing and $\mathcal{E}=(\varepsilon_n)_{n\ge0}$ is some i.i.d. noise process defined on $(\Theta,\mathcal{C},\nu)$ which is independent of $\mathcal{T}$. Furthermore, let $\boldsymbol{L}:\mathbb{R}^d\to[0,\infty)$ be a clippable and separable loss function with the corresponding loss function $L:X\times Y\times\mathbb{R}\to[0,\infty)$ satisfying the properties described as in Theorem 4.6. Finally, let $\boldsymbol{\Upsilon}:\mathcal{F}^d\to[0,\infty)$ be a separable regularizer. Then, for all fixed $\boldsymbol{f}_0=(f_0,\dots,f_0)$, $\varepsilon>0$, $\delta\ge0$, $\tau\ge1$, $n\ge n_0$ as in Theorem 4.6, and $r\in(0,1]$ satisfying (33), every learning method defined by (46) satisfies with probability $\mu\otimes\nu$ not less than $1-16e^{-\tau}$:*

$$\boldsymbol{\Upsilon}(\boldsymbol{f}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}})+\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\widehat{\boldsymbol{f}}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}})-\mathcal{R}^*_{\boldsymbol{L},\boldsymbol{P}}<2\boldsymbol{\Upsilon}(\boldsymbol{f}_0)+4\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f}_0)-4\mathcal{R}^*_{\boldsymbol{L},\boldsymbol{P}}+4dr+5d\varepsilon+2d\delta. \tag{47}$$

Again, this general oracle inequality can be applied to SVMs. We omit the details for the sake of brevity and only mention that such applications would lead to learning rates and not only consistency as in [41].

5 Proofs

5.1 Proofs of Section 2

Proof of Example 2.2. Consider the collection Π of ordered $n+1$-ples of points $z_0 < z_1 < \ldots < z_n \in Z$, where n is an arbitrary natural number. The total variation of a function $f : I \to \mathbb{R}$ is given by

$$\|f\|_{BV(Z)} := \sup_{(z_0,z_1,\ldots,z_n)\in\Pi} \sum_{i=1}^{n} |f(z_i) - f(z_{i-1})|.$$

Let us now assume that we have an $1 \leq i \leq n$ with $f(z_{i-1}) \leq f(z_i)$. Moreover, for $t \leq 0$, it is not difficult to verify that $|1 - e^t| \leq |t|$. This implies

$$\left| e^{f(z_i)} - e^{f(z_{i-1})} \right| = e^{f(z_i)} \left| 1 - e^{f(z_{i-1})-f(z_i)} \right| \leq \left\| e^f \right\|_\infty |f(z_i) - f(z_{i-1})| \,.$$

By interchanging the roles of $f(z_i)$ and $f(z_{i-1})$ we find the same estimate in the case of $f(z_{i-1}) \geq f(z_i)$. Consequently we obtain

$$\sum_{i=1}^{n} |e^{f(z_i)} - e^{f(z_{i-1})}| \leq \left\| e^f \right\|_\infty \sum_{i=1}^{n} |f(z_i) - f(z_{i-1})|$$

for all collection Π. Taking the supremum we get $\|e^f\|_{BV} \leq \|e^f\|_\infty \|f\|_{BV}$, i.e. (3) is satisfied. □

Proof of Example 2.3. Given a function $f \in C_{b,\alpha}(Z)$, we assume that $f(z) \geq f(z')$. Again, by using $|1 - e^t| \leq |t|$, $t \leq 0$, we obtain

$$\left| e^{f(z)} - e^{f(z')} \right| = e^{f(z)} \left| 1 - e^{f(z')-f(z)} \right| \leq \left\| e^f \right\|_\infty |f(z') - f(z)| \leq \left\| e^f \right\|_\infty |f|_\alpha |z - z'|^\alpha.$$

By interchanging the roles of $f(z)$ and $f(z')$ we find the same estimate in the case of $f(z') \geq f(z)$. Consequently we obtain $\|e^f\| \leq \|e^f\|_\infty |f|_\alpha$, i.e. (3) is satisfied. □

Proof of Theorem 2.6. ($\Rightarrow$) The proof is straightforward.
($\Leftarrow$) For $p, q \in [1, \infty]$ with $1/p + 1/q = 1$, let E_1 and E_2 be Banach spaces that are continuously embedded into $L_p(\mu)$ and $L_q(\mu)$, respectively, and let F be a Banach space that is continuously embedded into ℓ_∞. Analysis similar to that in the proof of [41, Theorem 5.1] shows that if, for all $n \geq 0$, and all $h \in E_1$, $g \in E_2$, the correlation sequence satisfies

$$\mathrm{cor}_{T,n}(h, g) \in F,$$

then there exists a constant $c \in [0, \infty)$ such that

$$\|\mathrm{cor}_{T,n}(h, g)\|_F \leq c \cdot \|h\|_{E_1} \|g\|_{E_2}, \qquad h \in E_1,\ g \in E_2. \tag{48}$$

In particular, (48) holds for $E_1 = \mathcal{C}(\Omega)$ and $E_2 = L_1(\mu)$ and the assertion is proved. □

5.2 Proofs of Section 3

The following lemma, which may be of independent interest, supplies the key to the proof of Theorem 3.1.

Lemma 5.1. *Let $\mathcal{Z} := (Z_n)_{n\geq 0}$ be a Z-valued stationary (time-reversed) $\mathcal{C}$-mixing process on the probability space $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\geq 0}$, and $P := \mu_{Z_0}$. Moreover, for $f : Z \to [0, \infty)$, suppose that $f \in \mathcal{C}(Z)$ and write $f_n := f \circ Z_n$. Finally, assume that we have natural numbers k and l satisfying*

$$2l \cdot \|f\|_{\mathcal{C}} \cdot d_k \leq \|f\|_{L_1(P)}. \tag{49}$$

Then we have

$$\mathbb{E}_\mu \prod_{j=0}^{l} f_{jk} \leq 2\|f\|_{L_1(P)}^{l+1}.$$

Proof of Lemma 5.1. We divide the proof into two parts.
(i) Suppose that the correlation inequality (7) holds. Obviously the case $f = 0$ P-a.s. is trivial. For $f \neq 0$, we define

$$D_l := \left| \mathbb{E}_\mu \prod_{j=0}^{l} f_{jk} - \prod_{j=0}^{l} \mathbb{E}_\mu f_{jk} \right|. \tag{50}$$

Then we have

$$\begin{aligned} D_l &\leq \left| \mathbb{E}_\mu \left(\prod_{j=0}^{l-1} f_{jk} \right) f_{lk} - \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk}\, \mathbb{E}_\mu f_{lk} \right| + \left| \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk}\, \mathbb{E}_\mu f_{lk} - \prod_{j=0}^{l} \mathbb{E}_\mu f_{jk} \right| \\ &= \left| \mathbb{E}_\mu \left(\prod_{j=0}^{l-1} f_{jk} \right) f_{lk} - \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk}\, \mathbb{E}_\mu f_{lk} \right| + \left| \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk}\, \mathbb{E}_\mu f_{lk} - \prod_{j=0}^{l-1} \mathbb{E}_\mu f_{jk}\, \mathbb{E}_\mu f_{lk} \right|. \end{aligned}$$

Since the stochastic process $\mathcal{Z}$ is stationary, the decay of correlations (7) together with $\psi := \prod_{j=0}^{l-1} f_{jk}$, $h := f$, and the assumption $f \geq 0$ yields

$$\begin{aligned} \left| \mathbb{E}_\mu \left(\prod_{j=0}^{l-1} f_{jk} \right) f_{lk} - \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk}\, \mathbb{E}_\mu f_{lk} \right| &\leq \left\| \prod_{j=0}^{l-1} f_{jk} \right\|_{L_1(\mu)} \|f\|_{\mathcal{C}}\, d_k = \left| \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk} \right| \|f\|_{\mathcal{C}}\, d_k \\ &\leq \left(\left| \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk} - \prod_{j=0}^{l-1} \mathbb{E}_\mu f_{jk} \right| + \prod_{j=0}^{l-1} \mathbb{E}_\mu f_{jk} \right) \|f\|_{\mathcal{C}}\, d_k \\ &= \left(D_{l-1} + \|f\|_{L_1(P)}^l \right) \|f\|_{\mathcal{C}}\, d_k. \end{aligned}$$

Moreover, for the second term, we find

$$\left| \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk}\, \mathbb{E}_\mu f_{lk} - \prod_{j=0}^{l-1} \mathbb{E}_\mu f_{jk}\, \mathbb{E}_\mu f_{lk} \right| = \|f\|_{L_1(P)} \left| \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk} - \prod_{j=0}^{l-1} \mathbb{E}_\mu f_{jk} \right| = \|f\|_{L_1(P)} D_{l-1}.$$

These estimates together imply that

$$\begin{aligned} D_l &\leq \left(D_{l-1} + \|f\|_{L_1(P)}^l \right) \|f\|_{\mathcal{C}}\, d_k + \|f\|_{L_1(P)} D_{l-1} \\ &= \left(\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}}\, d_k \right) D_{l-1} + \|f\|_{\mathcal{C}} \|f\|_{L_1(P)}^l\, d_k. \end{aligned} \tag{51}$$

In the following, we will show by induction that the latter estimate implies

$$D_l \leq \|f\|_{L_1(P)} \left(\left(\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}}\, d_k \right)^l - \|f\|_{L_1(P)}^l \right). \tag{52}$$

When $l = 1$, (52) is true because of (7). Now let $l \geq 1$ be given and suppose (52) is true for l. Then (51) and (52) imply

$$\begin{aligned}
D_{l+1} &\leq \left(\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}}\, d_k\right) D_l + \|f\|_{\mathcal{C}} \|f\|_{L_1(P)}^{l+1}\, d_k \\
&\leq \left(\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}}\, d_k\right) \left(\|f\|_{L_1(P)} \left(\left(\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}}\, d_k\right)^l - \|f\|_{L_1(P)}^l\right)\right) + \|f\|_{\mathcal{C}} \|f\|_{L_1(P)}^{l+1}\, d_k \\
&= \|f\|_{L_1(P)} \left(\left(\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}}\, d_k\right)^{l+1} - \|f\|_{L_1(P)}^{l+1}\right).
\end{aligned}$$

Thus, (52) holds for $l+1$, and the proof of the induction step is complete. By the principle of induction, (52) is thus true for all $l \geq 1$.

Using the binomial formula, we obtain

$$D_l \leq \|f\|_{L_1(P)} \left(\sum_{i=0}^{l} \binom{l}{i} \|f\|_{L_1(P)}^{l-i} \left(\|f\|_{\mathcal{C}}\, d_k\right)^i - \|f\|_{L_1(P)}^l\right).$$

For $i = 0, \dots, l$ we now set

$$a_i := \binom{l}{i} \|f\|_{L_1(P)}^{l-i} \left(\|f\|_{\mathcal{C}}\, d_k\right)^i.$$

The assumption (49) implies for $i = 0, \dots, l-1$

$$\begin{aligned}
\frac{a_{i+1}}{a_i} &= \frac{\binom{l}{i+1} \|f\|_{L_1(P)}^{l-i-1} \left(\|f\|_{\mathcal{C}}\, d_k\right)^{i+1}}{\binom{l}{i} \|f\|_{L_1(P)}^{l-i} \left(\|f\|_{\mathcal{C}}\, d_k\right)^i} = \frac{\frac{l!}{(i+1)!(l-i-1)!}}{\frac{l!}{i!(l-i)!}} \frac{\|f\|_{\mathcal{C}}\, d_k}{\|f\|_{L_1(P)}} \\
&= \frac{l-i}{i+1} \frac{\|f\|_{\mathcal{C}}\, d_k}{\|f\|_{L_1(P)}} \leq l \cdot \frac{\|f\|_{\mathcal{C}}}{\|f\|_{L_1(P)}} \cdot d_k \leq \frac{1}{2}.
\end{aligned}$$

This gives $a_i \leq 2^{-i} a_0$ for all $i = 0, \dots, l$ and consequently we have

$$\sum_{i=0}^{l} a_i = a_0 + \sum_{i=1}^{l} a_i \leq a_0 + \sum_{i=1}^{l} 2^{-i} a_0 = a_0 \cdot \left(\sum_{i=1}^{l} 2^{-i}\right) \leq 2a_0.$$

This implies

$$\begin{aligned}
D_l &\leq \|f\|_{L_1(P)} \left(\sum_{i=0}^{l} a_i - \|f\|_{L_1(P)}^l\right) \leq \|f\|_{L_1(P)} \left(2a_0 - \|f\|_{L_1(P)}^l\right) \\
&= \|f\|_{L_1(P)} \left(2\|f\|_{L_1(P)}^l - \|f\|_{L_1(P)}^l\right) = \|f\|_{L_1(P)}^{l+1}.
\end{aligned}$$

Using the definition of D_l we thus obtain

$$\mathbb{E}_\mu \prod_{j=0}^{l} f_{jk} \leq 2\|f\|_{L_1(P)}^{l+1}.$$

(ii) Suppose that the correlation inequality (8) holds.
Again, the case $f = 0$ P-a.s. is trivial. For $f \neq 0$, we estimate D_l defined as in (50) in a slightly different way from above:

$$D_l \leq \left|\mathbb{E}_\mu f_0 \prod_{j=1}^{l} f_{jk} - \mathbb{E}_\mu f_0 \mathbb{E}_\mu \prod_{j=1}^{l} f_{jk}\right| + \left|\mathbb{E}_\mu f_0 \mathbb{E}_\mu \prod_{j=1}^{l} f_{jk} - \prod_{j=0}^{l} \mathbb{E}_\mu f_{jk}\right|$$

$$= \left| \mathbb{E}_\mu f_0 \prod_{j=1}^{l} f_{jk} - \mathbb{E}_\mu f_0 \mathbb{E}_\mu \prod_{j=1}^{l} f_{jk} \right| + \left| \mathbb{E}_\mu f_0 \mathbb{E}_\mu \prod_{j=1}^{l} f_{jk} - \mathbb{E}_\mu f_0 \prod_{j=1}^{l} \mathbb{E}_\mu f_{jk} \right|.$$

Since the stochastic process $\mathcal{Z}$ is stationary, the decay of correlations (8) together with $h := f$, $\phi := \prod_{j=1}^{l} f_{jk}$, and the assumption $f \geq 0$ yields

$$\begin{aligned}
\left| \mathbb{E}_\mu f_0 \prod_{j=1}^{l} f_{jk} - \mathbb{E}_\mu f_0 \mathbb{E}_\mu \prod_{j=1}^{l} f_{jk} \right| &\leq \|f\|_{\mathcal{C}} \Big\| \prod_{j=1}^{l} f_{jk} \Big\|_{L_1(\mu)} d_k \\
&= \|f\|_{\mathcal{C}} \Big| \mathbb{E}_\mu \prod_{j=1}^{l} f_{jk} \Big| d_k = \|f\|_{\mathcal{C}} \Big| \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk} \Big| d_k \\
&\leq \|f\|_{\mathcal{C}} \left(\left| \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk} - \prod_{j=0}^{l-1} \mathbb{E}_\mu f_{jk} \right| + \prod_{j=0}^{l-1} \mathbb{E}_\mu f_{jk} \right) d_k \\
&= \|f\|_{\mathcal{C}} \left(D_{l-1} + \|f\|_{L_1(P)}^{l} \right) d_k.
\end{aligned}$$

Moreover, for the second term, since the stochastic process $\mathcal{Z}$ is stationary, we find

$$\begin{aligned}
\left| \mathbb{E}_\mu f_0 \mathbb{E}_\mu \prod_{j=1}^{l} f_{jk} - \mathbb{E}_\mu f_0 \prod_{j=1}^{l} \mathbb{E}_\mu f_{jk} \right| &= \|f\|_{L_1(P)} \left| \mathbb{E}_\mu \prod_{j=1}^{l} f_{jk} - \prod_{j=1}^{l} \mathbb{E}_\mu f_{jk} \right| \\
&= \|f\|_{L_1(P)} \left| \mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk} - \prod_{j=0}^{l-1} \mathbb{E}_\mu f_{jk} \right| \\
&= \|f\|_{L_1(P)} D_{l-1}.
\end{aligned}$$

Combining the above estimates, we get

$$\begin{aligned}
D_l &\leq \|f\|_{\mathcal{C}} \left(D_{l-1} + \|f\|_{L_1(P)}^{l} \right) d_k + \|f\|_{L_1(P)} D_{l-1} \\
&= \left(\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}}\, d_k \right) D_{l-1} + \|f\|_{\mathcal{C}} \|f\|_{L_1(P)}^{l}\, d_k.
\end{aligned}$$

This estimate coincides with (51). The rest of the argument is the same as in *(i)*, and the assertion is proved. □

To prove Theorem 3.1, we need to introduce some notations. In the following, for $t \in \mathbb{R}$, $\lfloor t \rfloor$ is the largest integer n satisfying $n \leq t$, and similarly, $\lceil t \rceil$ is the smallest integer n satisfying $n \geq t$. We write $h_i := h \circ Z_i$ and

$$S_n = \sum_{i=1}^{n} h_i = \sum_{i=1}^{n} h \circ Z_i.$$

We now recall the so-called blocking method. To this end, we partition the set $\{1, 2, \ldots, n\}$ into k blocks. Each block will contain approximatively $l := \lfloor n/k \rfloor$ terms. Let $r := n - k \cdot l < k$ denote the remainder when we divide n by k.

We now construct k blocks as follows. Define I_i, the indexes of terms in the i-th block, as

$$I_i = \begin{cases} \{i, i+k, \ldots, i+(l+1)k\}, & \text{if } 1 \leq i \leq r, \\ \{i, i+k, \ldots, i+lk\}, & \text{if } r+1 \leq i \leq k. \end{cases}$$

Note that the number of the terms satisfies

$$|I_i| = \begin{cases} l+1, & \text{for } 1 \le i \le r, \\ l, & \text{for } r+1 \le i \le k. \end{cases}$$

In other words, the first r blocks each contain $l+1$ terms, while the last $(k-r)$ blocks each contain l terms. Moreover, we have

$$\sum_{i=1}^{k} |I_i| = \sum_{i=1}^{r} |I_i| + \sum_{i=r+1}^{k} |I_i| = r(l+1) + (k-r)l = n. \tag{53}$$

Furthermore, for $i = 1, 2, \ldots, k$, we define the i-th block sum as

$$g_i = \sum_{j \in I_i} h_j \tag{54}$$

such that

$$S_n = \sum_{i=1}^{k} g_i. \tag{55}$$

Finally, for $i = 1, 2, \ldots, k$, define

$$p_i := \frac{|I_i|}{n}. \tag{56}$$

It follows from (53) that

$$\sum_{i=1}^{k} p_i = \frac{1}{n} \sum_{i=1}^{k} |I_i| = 1.$$

The following three lemmas will derive the upper bounds for the expected value of the exponentials of S_n.

Lemma 5.2. *Let $\mathcal{Z} := (Z_n)_{n \ge 0}$ be a Z-valued stationary stochastic process on the probability space $(\Omega, \mathcal{A}, \mu)$ and $P := \mu_{Z_0}$. Moreover, let k and l be defined as above, and for a bounded $h : Z \to \mathbb{R}$ we define g_i and S_n by (54) and (55), respectively. Then, for all $t > 0$, we have*

$$\mathbb{E}_\mu \exp\left(t \frac{S_n}{n}\right) \le \sum_{i=1}^{k} p_i \mathbb{E}_\mu \exp\left(t \frac{g_i}{|I_i|}\right).$$

Proof of Lemma 5.2. It is well-known that the exponential function is convex. Jensen's inequality together with $\sum_{i=1}^{k} p_i = 1$, (55), and (56) yields

$$\mathbb{E}_\mu \exp\left(t \frac{S_n}{n}\right) = \mathbb{E}_\mu \exp\left(\sum_{i=1}^{k} t p_i \frac{g_i}{|I_i|}\right) \le \sum_{i=1}^{k} p_i \mathbb{E}_\mu \exp\left(t \frac{g_i}{|I_i|}\right).$$

□

Lemma 5.3. *Let $\mathcal{Z} := (Z_n)_{n\geq 0}$ be a Z-valued stationary (time-reversed) $\mathcal{C}$-mixing process on the probability space $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\geq 0}$, and $P := \mu_{Z_0}$. Moreover, for $h : Z \to [0, \infty)$, we write $h_n := h \circ Z_n$. Finally, let k and l be defined as above. Then, for all $t > 0$ satisfying*

$$e^{\frac{t}{|I_i|}h} \in \mathcal{C}(Z) \ \text{ and } \ 2l \cdot \|e^{\frac{t}{|I_i|}h}\|_{\mathcal{C}} \cdot d_k \leq \|e^{\frac{t}{|I_i|}h}\|_{L_1(P)}, \tag{57}$$

we have

$$\mathbb{E}_\mu \exp\left(t\frac{g_i}{|I_i|}\right) \leq 2\left(\mathbb{E}_P \exp\left(t\frac{h}{|I_i|}\right)\right)^{|I_i|}.$$

Proof of Lemma 5.3. The ith block sum g_i in (54) depends only on h_{i+jk} with j ranging from 0 through $|I_i| - 1$. Since $\mathcal{Z}$ is stationary, Lemma 5.1 with $f := \exp(\frac{t}{|I_i|}h)$ then yields

$$\begin{aligned}
\mathbb{E}_\mu \exp\left(t\frac{g_i}{|I_i|}\right) &= \mathbb{E}_\mu \exp\left(\frac{t}{|I_i|}\sum_{j=0}^{|I_i|-1} h_{i+jk}\right) = \mathbb{E}_\mu \exp\left(\frac{t}{|I_i|}\sum_{j=0}^{|I_i|-1} h_{jk}\right) \\
&= \mathbb{E}_\mu \prod_{j=0}^{|I_i|-1} \exp\left(\frac{t}{|I_i|}h_{jk}\right) \leq 2\left(\mathbb{E}_P \exp\left(t\frac{h}{|I_i|}\right)\right)^{|I_i|}.
\end{aligned}$$

□

Lemma 5.4. *Let $\mathcal{Z} := (Z_n)_{n\geq 0}$ be a Z-valued stationary (time-reversed) $\mathcal{C}$-mixing process on the probability space $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\geq 0}$, and $P := \mu_{Z_0}$. Moreover, for $h : Z \to [0, \infty)$, we write $h_n := h \circ Z_n$ and suppose that $\mathbb{E}_P h = 0$, $\|h\| \leq A$, $\|h\|_\infty \leq B$, and $\mathbb{E}_P h^2 \leq \sigma^2$ for some $A > 0$, $B > 0$ and $\sigma \geq 0$. Finally, let k and l be defined as above. Then, for all $i = 1, \ldots, k$, and all $t > 0$ satisfying $0 < t < 3l/B$ and (57), we have*

$$\mathbb{E}_\mu \exp\left(t\frac{g_i}{|I_i|}\right) \leq 2\exp\left(\frac{t^2\sigma^2}{2(l - tB/3)}\right).$$

Proof of Lemma 5.4. Because of $\|h\|_\infty \leq B$ and $2 \cdot 3^{j-2} \leq j!$, we obtain

$$\begin{aligned}
\exp\left(\frac{t}{|I_i|}h\right) &= 1 + \frac{t}{|I_i|}h + \sum_{j=2}^{\infty}\left(\frac{t}{|I_i|}\right)^j \frac{h^j}{j!} \\
&\leq 1 + \frac{t}{|I_i|}h + \sum_{j=2}^{\infty}\left(\frac{t}{|I_i|}\right)^j \frac{h^2 B^{j-2}}{2 \cdot 3^{j-2}} \\
&= 1 + \frac{t}{|I_i|}h + \frac{1}{2}\left(\frac{t}{|I_i|}\right)^2 h^2 \sum_{j=2}^{\infty}\left(\frac{tB}{3|I_i|}\right)^{j-2} \\
&= 1 + \frac{t}{|I_i|}h + \frac{1}{2}\left(\frac{t}{|I_i|}\right)^2 h^2 \frac{1}{1 - tB/(3|I_i|)}
\end{aligned}$$

if $tB/(3|I_i|) < 1$. This, together with $\mathbb{E}_P h = 0$, $1 + x \leq e^x$, and $l \leq |I_i| \leq l + 1$, implies

$$\begin{aligned}
\left(\mathbb{E}_P \exp\left(t\frac{h}{|I_i|}\right)\right)^{|I_i|} &\leq \left(1 + \frac{1}{2}\left(\frac{t}{|I_i|}\right)^2 \sigma^2 \frac{1}{1 - tB/(3|I_i|)}\right)^{|I_i|} \\
&\leq \left(\exp\left(\frac{1}{2}\left(\frac{t}{|I_i|}\right)^2 \sigma^2 \frac{1}{1 - tB/(3|I_i|)}\right)\right)^{|I_i|}
\end{aligned}$$

$$= \exp\left(\frac{t^2\sigma^2}{2(|I_i| - tB/3)}\right)$$
$$\leq \exp\left(\frac{t^2\sigma^2}{2(l - tB/3)}\right), \tag{58}$$

since the assumed $tB/(3l) < 1$ implies $tB/(3|I_i|) < 1$. Lemma 5.3 then yields

$$\mathbb{E}_\mu \exp\left(t\frac{g_i}{|I_i|}\right) \leq 2\exp\left(\frac{t^2\sigma^2}{2(l - tB/3)}\right).$$

□

Proof of Theorem 3.1. For k and l as above we define

$$t := \frac{l\varepsilon}{\sigma^2 + \varepsilon B/3}. \tag{59}$$

Then we have

$$\frac{t}{|I_i|} \leq \frac{t}{l} = \frac{\varepsilon}{\sigma^2 + \varepsilon B/3} \leq \frac{\varepsilon}{\varepsilon B/3} = \frac{3}{B}. \tag{60}$$

In particular, this t satisfies $0 < t < 3l/B$. Moreover, we find

$$\left\|\exp\left(\frac{t}{|I_i|}h\right)\right\|_\infty \leq \exp\left(\frac{3}{B}\cdot B\right) = e^3. \tag{61}$$

Then, the assumption (3) together with the bounds (61) and (60) implies

$$\left\|\exp\left(\frac{t}{|I_i|}h\right)\right\| \leq \left\|\exp\left(\frac{t}{|I_i|}h\right)\right\|_\infty \left\|\frac{t}{|I_i|}h\right\| \leq e^3 \cdot \frac{t}{|I_i|}\|h\| \leq \frac{3e^3A}{B}. \tag{62}$$

Since $-B \leq h \leq B$, we further find

$$\left\|\exp\left(\frac{t}{|I_i|}h\right)\right\|_{L_1(P)} = \mathbb{E}_P \exp\left(\frac{t}{|I_i|}h\right) \geq \exp\left(\frac{3}{B}\cdot(-B)\right) = e^{-3}. \tag{63}$$

Now we choose $k := \lfloor (\log n)^{\frac{2}{\gamma}} \rfloor + 1$, which implies $k \geq (\log n)^{\frac{2}{\gamma}}$. On the other hand, since $(\log n)^{\frac{2}{\gamma}} \geq 1$ for $n \geq n_0 \geq 3$, we have $k \leq 2(\log n)^{\frac{2}{\gamma}}$. This implies

$$l = \frac{n-r}{k} \geq \frac{n}{k} - 1 \geq \frac{1}{2}\frac{n}{(\log n)^{\frac{2}{\gamma}}} - 1 \geq \frac{1}{4}\frac{n}{(\log n)^{\frac{2}{\gamma}}}, \tag{64}$$

since we have $n \geq 4(\log n)^{\frac{2}{\gamma}}$ for $n \geq n_0$. Now, by (61), (62), (63), (6), and (13) we obtain

$$\begin{aligned} l \cdot \frac{\|e^{\frac{t}{|I_i|}h}\|_{\mathcal{C}}}{\|e^{\frac{t}{|I_i|}h}\|_{L_1(P)}} \cdot d_k &\leq l \cdot \frac{\|e^{\frac{t}{|I_i|}h}\|_\infty + \|e^{\frac{t}{|I_i|}h}\|}{\|e^{\frac{t}{|I_i|}h}\|_{L_1(P)}} \cdot c \cdot \exp\left(-bk^\gamma\right) \\ &\leq n \cdot \frac{e^3 + \frac{3e^3A}{B}}{e^{-3}} \cdot c \cdot \exp\left(-b(\log n)^2\right) \\ &\leq n \cdot \frac{404c(3A+B)}{B} \cdot \exp\left(-b\log n \cdot \frac{3}{b}\right) \\ &\leq n \cdot \frac{n^2}{2} \cdot n^{-3} = \frac{1}{2}, \end{aligned}$$

i.e., the assumption (57) is valid.

Summarizing, the value of t defined as in (59) satisfies $0 < t < 3l/B$ and the assumption (57). In other words, all the requirements on t in Lemma 5.4 are satisfied.

Now, for this t, by using Markov's inequality, Lemma 5.2, and Lemma 5.4, we obtain for any $\varepsilon > 0$,

$$\begin{aligned} P\left(\frac{S_n}{n} > \varepsilon\right) &= P\left(\exp\left(t\frac{S_n}{n}\right) > \exp\left(t\varepsilon\right)\right) \\ &\leq \exp\left(-t\varepsilon\right)\mathbb{E}_\mu \exp\left(t\frac{S_n}{n}\right) \\ &\leq \exp\left(-t\varepsilon\right)\sum_{i=1}^k p_i \mathbb{E}_\mu \exp\left(t\frac{g_i}{|I_i|}\right) \\ &\leq \exp\left(-t\varepsilon\right)\cdot 2\exp\left(\frac{t^2\sigma^2}{2(l-tB/3)}\right)\sum_{i=1}^k p_i \\ &= 2\exp\left(-t\varepsilon + \frac{t^2\sigma^2}{2(l-tB/3)}\right). \end{aligned} \tag{65}$$

Substituting the definition of t into the exponent of inequality (65), we get

$$\begin{aligned} -t\varepsilon + \frac{t^2\sigma^2}{2(l-tB/3)} &= -\frac{l\varepsilon^2}{\sigma^2+\varepsilon B/3} + \frac{l^2\varepsilon^2}{(\sigma^2+\varepsilon B/3)^2}\cdot\frac{\sigma^2}{2\left(l-\frac{l\varepsilon B/3}{\sigma^2+\varepsilon B/3}\right)} \\ &= -\frac{l\varepsilon^2}{\sigma^2+\varepsilon B/3} + \frac{l\varepsilon^2}{\sigma^2+\varepsilon B/3}\cdot\frac{\sigma^2}{2\left(\sigma^2+\varepsilon B/3-\varepsilon B/3\right)} \\ &= \frac{-l\varepsilon^2}{2\left(\sigma^2+\varepsilon B/3\right)}, \end{aligned}$$

hence

$$\mathbb{P}\left(\frac{1}{n}S_n > \varepsilon\right) \leq 2\exp\left(-\frac{-l\varepsilon^2}{2\left(\sigma^2+\varepsilon B/3\right)}\right).$$

Using the estimate (64), we thus obtain

$$\mathbb{P}\left(\frac{1}{n}S_n > \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{8(\log n)^{\frac{2}{\gamma}}\left(\sigma^2+\varepsilon B/3\right)}\right),$$

for all $n \geq n_0$ and $\varepsilon > 0$. Setting $\tau := \frac{n\varepsilon^2}{8(\log n)^{\frac{2}{\gamma}}(\sigma^2+\varepsilon B/3)}$, we then have

$$\mu\left(\left\{\omega\in\Omega : \frac{1}{n}\sum_{i=1}^n h(Z_i(\omega)) \geq \varepsilon\right\}\right) \leq 2e^{-\tau}, \qquad n \geq n_0.$$

Simple transformations and estimations then yield

$$\mu\left(\left\{\omega\in\Omega : \frac{1}{n}\sum_{i=1}^n h(Z_i(\omega)) \geq \sqrt{\frac{8(\log n)^{\frac{2}{\gamma}}\tau\sigma^2}{n}} + \frac{8(\log n)^{\frac{2}{\gamma}}B\tau}{3n}\right\}\right) \leq 2e^{-\tau}$$

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

5.3 Proofs of Section 4

Proof of Lemma 4.7. *(i)* For the least square loss (23), by using $a + b \leq (2(a^2 + b^2))^{1/2}$, we obtain

$$\begin{aligned}|L(x,y,f(x)) - L(x',y',f(x'))| &= |(y - f(x))^2 - (y' - f(x'))^2| \\ &= |y - f(x) + y' - f(x')| \cdot |y - f(x) - y' + f(x')| \\ &\leq \big(|y + y'| + |f(x) + f(x')|\big)\big(|y - y'| + |f(x) - f(x')|\big) \\ &\leq 2\left(M + \|f\|_\infty\right)\big(|y - y'| + |f|_1|x - x'|\big) \\ &\leq 2\left(M + \|f\|_\infty\right)(1 + |f|_1)\big(|y - y'| + |x - x'|\big) \\ &\leq 2\sqrt{2}\left(M + \|f\|_\infty\right)(1 + |f|_1)\|(x,y) - (x',y')\|_2\end{aligned}$$

for all $(x,y),(x',y') \in X \times Y$, that is, we have proved the assertion.

(ii) Let L be the the τ-pinball loss (24) and define

$$D := L(x,y,f(x)) - L(x',y',f(x')).$$

We divide the proof into the following four cases. If $y \geq f(x)$ and $y' \geq f(x')$, we have

$$|D| = |\tau(y - f(x)) - \tau(y' - f(x'))| = \tau|(y - y') - (f(x) - f(x'))|.$$

If $y < f(x)$ and $y' < f(x')$, in an exactly similar way we obtain

$$|D| = (1 - \tau)|(y - y') - (f(x) - f(x'))|.$$

Moreover, in case of $y \geq f(x)$ and $y' < f(x')$, we get

$$|D| = |\tau(y - f(x)) + (1 - \tau)(y' - f(x'))| \leq |(y - f(x)) + (f(x') - y')|.$$

Similar arguments to the case $y < f(x)$ and $y' \geq f(x')$ show that

$$|D| = |-(1 - \tau)(y - f(x)) - \tau(y' - f(x'))| \leq |(y - f(x)) + (f(x') - y')|.$$

Summarizing, for all $(x,y),(x',y') \in X \times Y$, we have

$$|L(x,y,f(x)) - L(x',y',f(x'))| \leq |(y - y') - (f(x) - f(x'))| \leq |y - y'| + |f(x) - f(x')|.$$

The rest of the argument is similar to that of part *(i)*, and the assertion is proved. □

For our proof of Theorem 4.6 we need the following simple and well-known lemma (see e.g. [42, Lemma 7.1]):

Lemma 5.5. *For $q \in (1,\infty)$, define $q' \in (1,\infty)$ by $1/q + 1/q' = 1$. Then, for all $a, b \geq 0$, we have $(qa)^{2/q}(q'b)^{2/q'} \leq (a + b)^2$ and $ab \leq a^q/q + b^{q'}/q'$.*

Apart from the semi-norm bounds involving A_0, A_1, and A^* and some constants, for example, the constant n_0 and the constants on the right side of the oracle inequality, the proof of Theorem 4.6 is almost identical to the proof of [23, Theorem 3.1]. For this reason, a few parts of the proof will be omitted.

Proof of Theorem 4.6. **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

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

and consequently we obtain

$$
\begin{aligned}
&\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P} \\
&= \Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} \\
&\leq \Upsilon(f_0) + \mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_{D_n} h_{\widehat{f}_{D_n,\Upsilon}} + \mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} + \delta \\
&= (\Upsilon(f_0) + \mathbb{E}_P h_{f_0}) + (\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_P h_{f_0}) + (\mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} - \mathbb{E}_{D_n} h_{\widehat{f}_{D_n,\Upsilon}}) + \delta. \qquad (66)
\end{aligned}
$$

Estimating the First Stochastic Term. Let us 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

$$
\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_P h_{f_0} = \Big(\mathbb{E}_{D_n}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})\Big) + (\mathbb{E}_{D_n} h_{\widehat{f}_0} - \mathbb{E}_P h_{\widehat{f}_0}). \qquad (67)
$$

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

$$
\mathbb{E}_P \Big((h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})\Big)^2 \leq \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})^2 \leq B_0 \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}).
$$

Moreover, we find

$$
\begin{aligned}
\|h_{f_0} - h_{\widehat{f}_0}\| &= \|(L \circ f_0 - L \circ f^*_{L,P}) - (L \circ \widehat{f}_0 - L \circ f^*_{L,P})\| \\
&= \|L \circ f_0 - L \circ \widehat{f}_0\| \leq \|L \circ f_0\| + \|L \circ \widehat{f}_0\| \leq 2A_0.
\end{aligned}
$$

Inequality (15) applied to $h := (h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})$ thus shows that for

$$
n \geq n^*_0 \geq \max\left\{\min\left\{m \geq 3 : m^2 \geq \frac{808c(6A_0 + B_0)}{B_0} \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \geq 4\right\}, e^{\frac{3}{b}}\right\},
$$

we have

$$
\mathbb{E}_{D_n}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) \leq \sqrt{\frac{8(\log n)^{\frac{2}{\gamma}} \tau B_0 \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})}{n}} + \frac{8(\log n)^{\frac{2}{\gamma}} B_0 \tau}{3n}
$$

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

$$
\sqrt{8(\log n)^{\frac{2}{\gamma}} n^{-1} \tau B_0 \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})} \leq \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) + 2(\log n)^{\frac{2}{\gamma}} n^{-1} B_0 \tau,
$$

and consequently we have with probability μ not less than $1 - 2e^{-\tau}$ that

$$
\mathbb{E}_{D_n}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) \leq \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) + \frac{14(\log n)^{\frac{2}{\gamma}} B_0 \tau}{3n}. \qquad (68)
$$

In order to bound the remaining term in (67), that is $\mathbb{E}_{D_n} h_{\widehat{f}_0} - \mathbb{E}_P h_{\widehat{f}_0}$, we first observe that the assumed $L(x, y, t) \leq 1$ for all $(x, y) \in X \times Y$ and $t, t' \in [-M, M]$ implies $\|h_{\widehat{f}_0}\|_\infty \leq 1$, and hence we have $\|h_{\widehat{f}_0} - \mathbb{E}_P h_{\widehat{f}_0}\|_\infty \leq 2$. Furthermore, we have

$$
\|h_{f_0}\| = \|L \circ f_0 - L \circ f^*_{L,P}\| \leq \|L \circ f_0\| + \|L \circ f^*_{L,P}\| \leq A_0 + A^*.
$$

Moreover, (31) yields

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

In addition, if $\vartheta \in (0,1]$, the second inequality in Lemma 5.5 implies for $q := \frac{2}{2-\vartheta}$, $q' := \frac{2}{\vartheta}$, $a := ((\log n)^{\frac{2}{\gamma}} n^{-1} 2^{3-\vartheta}\vartheta^\vartheta V\tau)^{1/2}$, and $b := (2\vartheta^{-1}\mathbb{E}_P h_{\widehat{f}_0})^{\vartheta/2}$, that

$$\begin{aligned}\sqrt{\frac{8(\log n)^{\frac{2}{\gamma}} V\tau(\mathbb{E}_P h_{\widehat{f}_0})^\vartheta}{n}} &\leq \left(1-\frac{\vartheta}{2}\right)\left(\frac{2^{3-\vartheta}\vartheta^\vartheta(\log n)^{\frac{2}{\gamma}} V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \mathbb{E}_P h_{\widehat{f}_0} \\ &\leq \left(\frac{8(\log n)^{\frac{2}{\gamma}} V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \mathbb{E}_P h_{\widehat{f}_0}.\end{aligned}$$

Since $\mathbb{E}_P h_{\widehat{f}_0} \geq 0$, this inequality also holds for $\vartheta = 0$, and hence (15) shows that for

$$n \geq n_0^* \geq \max\left\{\min\left\{m \geq 3 : m^2 \geq \frac{808c(3A_0 + 3A^* + 2)}{2} \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \geq 4\right\}, e^{\frac{3}{b}}\right\},$$

we have

$$\mathbb{E}_{D_n} h_{\widehat{f}_0} - \mathbb{E}_P h_{\widehat{f}_0} < \mathbb{E}_P h_{\widehat{f}_0} + \left(\frac{8(\log n)^{\frac{2}{\gamma}} V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{16(\log n)^{\frac{2}{\gamma}}\tau}{3n} \tag{69}$$

with probability μ not less than $1 - 2e^{-\tau}$. By combining this estimate with (68) and (67), we now obtain that with probability μ not less than $1 - 4e^{-\tau}$ we have

$$\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_P h_{f_0} < \mathbb{E}_P h_{f_0} + \left(\frac{8(\log n)^{\frac{2}{\gamma}} V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{16(\log n)^{\frac{2}{\gamma}}\tau}{3n} + \frac{14(\log n)^{\frac{2}{\gamma}} B_0\tau}{3n}, \tag{70}$$

since $1 \leq B_0$, i.e., we have established a bound on the second term in (66).

Estimating the Second Stochastic Term. For the third term in (66) let us first consider the case $n/(\log n)^{\frac{2}{\gamma}} < 8(\tau + \varphi(\varepsilon/2)2^p r^p)$. Combining (70) with (66) and using $1 \leq B_0$, $1 \leq V$, and $\mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} - \mathbb{E}_{D_n} h_{\widehat{f}_{D_n,\Upsilon}} \leq 2$, then we find

$$\begin{aligned}&\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\Upsilon}) - \mathcal{R}_{L,P}^* \\ &\leq \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} + \left(\frac{8(\log n)^{\frac{2}{\gamma}} V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{16(\log n)^{\frac{2}{\gamma}}\tau}{3n} + \frac{14(\log n)^{\frac{2}{\gamma}} B_0\tau}{3n} \\ &\quad + (\mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} - \mathbb{E}_{D_n} h_{\widehat{f}_{D_n,\Upsilon}}) + \delta \\ &\leq \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} + \left(\frac{8(\log n)^{\frac{2}{\gamma}} V(\tau + \varphi(\varepsilon/2)2^p r^p)}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{10(\log n)^{\frac{2}{\gamma}} B_0\tau}{n} \\ &\quad + 2\left(\frac{8(\log n)^{\frac{2}{\gamma}} V(\tau + \varphi(\varepsilon/2)2^p r^p)}{n}\right)^{\frac{1}{2-\vartheta}} + \delta \\ &\leq 2\Upsilon(f_0) + 4\mathbb{E}_P h_{f_0} + 3\left(\frac{24(\log n)^{\frac{2}{\gamma}} V(\tau + \varphi(\varepsilon/2)2^p r^p)}{3n}\right)^{\frac{1}{2-\vartheta}} + \frac{10(\log n)^{\frac{2}{\gamma}} B_0\tau}{n} + 2\delta\end{aligned}$$

$$\leq 2\Upsilon(f_0) + 4\mathbb{E}_P h_{f_0} + 4r + 2\delta$$

with probability μ not less than $1 - 4e^{-\tau}$. It thus remains to consider the case $n/(\log n)^{\frac{2}{\gamma}} \geq 8(\tau + \varphi(\varepsilon/2)2^p r^p)$.

Introduction of the Quotients. To establish a non-trivial bound on the term $\mathbb{E}_P h_{\widehat{f}_D} - \mathbb{E}_{D_n} h_{\widehat{f}_D}$ in (66), we define functions

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

For $f \in \mathcal{F}_r$, we have $\|\mathbb{E}_P h_{\widehat{f}} - h_{\widehat{f}}\|_\infty \leq 2$. Furthermore, for $f \in \mathcal{F}_r$ and $k \geq 0$ with $2^k r \leq 1$, by the assumption (29) we find

$$\|h_{\widehat{f}}\| = \|L \circ \widehat{f} - L \circ f^*_{L,P}\| \leq \|L \circ \widehat{f}\| + \|L \circ f^*_{L,P}\| \leq A_{2^k r} + A^* \leq A_1 + A^*.$$

Moreover, for $f \in \mathcal{F}_r$, the variance bound (31) implies

$$\mathbb{E}_P (h_{\widehat{f}} - \mathbb{E}_P h_{\widehat{f}})^2 \leq \mathbb{E}_P h^2_{\widehat{f}} \leq V(\mathbb{E}_P h_{\widehat{f}})^\vartheta \leq V r^\vartheta. \tag{71}$$

Peeling. This part is completely identical to the part *Peeling* on page 135 of our work [23]. Hence we have neglected some steps of the derivations. In case of uncertainty one may refer to [23] for details.

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 $\mathcal{F}_1$:

$$\mathcal{F}_1 \subset \mathcal{F}_r \cup \bigcup_{k=1}^{K+1} \left(\mathcal{F}_{2^k r} \backslash \mathcal{F}_{2^{k-1} r}\right). \tag{72}$$

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

$$\widetilde{\mathcal{C}}_{\varepsilon,r,k} := \bigcup_{l=0}^{k} \overline{C}_{\varepsilon,r,l}, \quad 0 \leq k \leq K+1. \tag{73}$$

Then we have $\overline{C}_{\varepsilon,1} = \bigcup_{k=0}^{K+1} \widetilde{C}_{\varepsilon,r,k}$. Moreover, the cardinality of $\widetilde{\mathcal{C}}_{\varepsilon,r,k}$ can be estimated by

$$|\widetilde{\mathcal{C}}_{\varepsilon,r,k}| \leq (k+1) \exp\left(\varphi(\varepsilon/2) 2^{kp} r^p\right), \quad 0 \leq k \leq K+1. \tag{74}$$

Then, peeling by [23, Theorem 5.2] implies

$$\mu\left(\sup_{f \in \overline{\mathcal{C}}_{\varepsilon,1}} \mathbb{E}_{D_n} g_{f,r} > \frac{1}{4}\right) \leq 2 \sum_{k=1}^{K+1} \mu\left(\sup_{f \in \widetilde{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\widehat{f}} - h_{\widehat{f}}) > 2^{k-3} r\right). \tag{75}$$

Estimating the Error Probabilities on the "Spheres". Our next goal is to estimate all the error probabilities on the right-hand side of (75). By our construction, we have $\widetilde{\mathcal{C}}_{\varepsilon,r,k} \subset \mathcal{F}_{2^k r}$. This, together with (14), (71), the union bound and the estimates of the covering numbers (74), implies that for

$$n \geq n_0^* \geq \max\left\{\min\left\{m \geq 3 : m^2 \geq \frac{808c(3A_1 + 3A^* + 2)}{2} \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \geq 4\right\}, e^{\frac{3}{b}}\right\},$$

we have

$$
\begin{aligned}
&\mu\left(\sup_{f\in\widetilde{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\widehat{f}} - h_{\widehat{f}}) > 2^{k-3}r\right) \\
&\leq 2|\widetilde{\mathcal{C}}_{\varepsilon,r,k}| \exp\left(-\frac{n}{8(\log n)^{\frac{2}{\gamma}}} \cdot \frac{(2^{k-3}r)^2}{V(2^k r)^\vartheta + 2(2^{k-3}r)/3}\right) \\
&\leq 2(k+1)\exp\left(\varphi(\varepsilon/2)2^{kp}r^p\right) \cdot \exp\left(-\frac{n}{8(\log n)^{\frac{2}{\gamma}}} \cdot \frac{3(2^{k-1}r)^2}{96V(2^{k-1}r)^\vartheta + 8(2^{k-1}r)}\right),
\end{aligned}
\tag{76}
$$

since $\vartheta \in [0,1]$. For $k \geq 1$, we denote the right-hand side of this estimate by $p_k(r)$, that is

$$
p_k(r) := 2(k+1)\exp\left(\varphi(\varepsilon/2)2^{kp}r^p\right) \cdot \exp\left(-\frac{n}{8(\log n)^{\frac{2}{\gamma}}} \cdot \frac{3(2^{k-1}r)^2}{96V(2^{k-1}r)^\vartheta + 8(2^{k-1}r)}\right). \tag{77}
$$

Then, as derived in [23], we can obtain

$$
q_k(r) := \frac{p_{k+1}(r)}{p_k(r)} \leq 2\exp\left(\varphi(\varepsilon/2)2^{kp+1}r^p\right) \cdot \exp\left(-\frac{n}{8(\log n)^{\frac{2}{\gamma}}} \cdot \frac{3(2^{k-1}r)^2}{96V(2^{k-1}r)^\vartheta + 8(2^{k-1}r)}\right),
$$

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

$$
\begin{aligned}
q_k(r) &\leq 2\exp\left(\varphi(\varepsilon/2)2^{kp+1}r^p\right) \cdot \exp\left(-\frac{n}{8(\log n)^{\frac{2}{\gamma}}} \cdot \frac{3(2^{k-1}r)^2}{96V(2^{k-1}r)^\vartheta + 8(2^{k-1}r)^\vartheta}\right) \\
&\leq 2\exp\left(2^{(k-1)p} \cdot 4r^p\varphi(\varepsilon/2) - \frac{2^{(k-1)(2-\vartheta)} \cdot 3nr^{2-\vartheta}}{64(12V+1)(\log n)^{\frac{2}{\gamma}}}\right).
\end{aligned}
$$

Since $p \in (0,1]$, $k \geq 1$ and $\vartheta \in [0,1]$, we have $2^{(k-1)p} \leq 2^{(k-1)(2-\vartheta)}$. Then the first assumption in (33), namely,

$$
r \geq \left(\frac{512(12V+1)(\log n)^{\frac{2}{\gamma}}(\tau + \varphi(\varepsilon/2)2^p r^p)}{3n}\right)^{\frac{1}{2-\vartheta}}
$$

implies that $3nr^{2-\vartheta} \geq 512(12V+1)(\log n)^{\frac{2}{\gamma}}\varphi(\varepsilon/2)r^p$. By using $2^{(k-1)(2-\vartheta)} \geq 1$, we find

$$
q_k(r) \leq 2\exp\left(-\frac{3nr^{2-\vartheta}}{128(12V+1)(\log n)^{\frac{2}{\gamma}}}\right).
$$

Moreover, since $\tau \geq 1$, the first assumption in (33) implies also $3nr^{2-\vartheta} \geq 4 \cdot 128(12V+1)(\log n)^{\frac{2}{\gamma}}$. Hence we have $q_k(r) \leq 2e^{-4}$, that is,

$$
p_{k+1}(r) \leq 2e^{-4}p_k(r) \qquad \text{for all } k \geq 1. \tag{78}
$$

Summing all the Error Probabilities. Now, combining (75) with (76), (77), and (78), we obtain

$$
\mu\left(\sup_{f\in\overline{\mathcal{C}}_{\varepsilon,1}} \mathbb{E}_{D_n} g_{f,r} > \frac{1}{4}\right) \leq 2\sum_{k=1}^{K+1} p_k(r) \leq 3p_1(r)
$$

$$= 12 \exp\left(\varphi(\varepsilon/2) 2^p r^p\right) \cdot \exp\left(-\frac{n}{8(\log n)^{\frac{2}{\gamma}}} \cdot \frac{3r^2}{96Vr^\vartheta + 8r}\right)$$
$$\le 12 \exp\left(\varphi(\varepsilon/2) 2^p r^p\right) \cdot \exp\left(-\frac{3nr^{2-\vartheta}}{64(12V+1)(\log n)^{\frac{2}{\gamma}}}\right),$$

where in the last step we used $r \in (0,1]$ and $\vartheta \in [0,1]$. Then once again the first assumption in (33) gives $3nr^{2-\vartheta} \ge 64(12V+1)(\log n)^{\frac{2}{\gamma}}(\tau + \varphi(\varepsilon/2)2^p r^p)$ and a simple transformation thus yields

$$\mu\left(D_n \in (X \times Y)^n : \sup_{f \in \overline{\mathcal{C}}_{\varepsilon,1}} \mathbb{E}_{D_n} g_{f,r} \le \frac{1}{4}\right) \ge 1 - 12e^{-\tau}.$$

The rest of the argument is completely analogous to the proof of [23, Theorem 3.1] and the assertion is proved. □

Proof of Theorem 4.8. For the least-square loss, the variance bound (31) is valid with $\vartheta = 1$, hence the condition (33) is satisfied if

$$r \ge \max\left\{ \left(c_V 2^{1+3p} a\right)^{\frac{1}{1-p}} \sigma^{-\frac{d}{1-p}} \lambda^{-\frac{p}{1-p}} \left(\frac{n}{(\log n)^{\frac{2}{\gamma}}}\right)^{\frac{1}{1-p}} \varepsilon^{-\frac{2p}{1-p}}, \frac{2c_V (\log n)^{\frac{2}{\gamma}}\tau}{n}, \frac{20B_0 (\log n)^{\frac{2}{\gamma}}\tau}{n}, r^* \right\}. \tag{79}$$

Furthermore, [21, Section 2] shows that there exists a constant $c > 0$ such that for all $\sigma \in (0,1]$, there is an $f_0 \in H_\sigma$ with $\|f_0\|_\infty \le c$, $\|f_0\|^2_{H_\sigma} \le c\sigma^{-d}$, and

$$\mathcal{R}_{L,P}(f_0) - \mathcal{R}^*_{L,P} \le c\sigma^{2t}\,.$$

Moreover, [41, Lemma 5.5] shows every function f in H_σ is Lipschitz continuous with

$$|f|_1 \le \sqrt{2}\sigma^{-1}\|f\|_{H_\sigma(X)}\,,$$

and this implies

$$|\widehat{f}_0|_1 \le |f_0|_1 \le \sqrt{2}\sigma^{-1}\|f_0\|_{H_\sigma(X)} \le \sqrt{2}c\sigma^{-1}.$$

Moreover, there exists a constant $C^* < \infty$ such that $|f^*_{L,P}|_1 \le C^*$, since we have assumed that $f^*_{L,P} \in \mathrm{Lip}(\mathbb{R}^d)$. Then, Lemma 4.7 (i) yields

$$4A_0 + A_1 + A^* + 1 = 2\sqrt{2}\,(M + \|f\|_\infty)\left(4 + 4|f_0|_1 + 1 + \sup_{f \in \mathcal{F}_1} |f|_1 + 1 + |f^*_{L,P}|_1 + 1\right) + 1$$
$$\le 2\sqrt{2}\,(M + \|f\|_\infty)\left(7 + 4\sqrt{2}c\sigma^{-1} + \sup_{r \le 1} \sqrt{2}\sigma^{-1}\lambda^{-1/2} r^{1/2} + C^*\right) + 1$$
$$= 2\sqrt{2}\,(M + \|f\|_\infty)\left(7 + 5\sqrt{2}c\sigma^{-1}\lambda^{-1/2} + C^*\right) + 1$$
$$\le 2C\sigma^{-1}\lambda^{-1/2} \le 2Cn$$

for all $\sigma, \lambda \in (0,1]$ with $\lambda\sigma^2 \ge n^{-2}$, where C is a constant independent of n, λ, and σ. For

$$n \ge \max\left\{2C, \min\left\{m \ge 3 : \frac{m}{(\log m)^{\frac{2}{\gamma}}} \ge 4\right\}, e^{\frac{3}{b}}\right\},$$

the oracle inequality (34) thus implies

$$\begin{aligned}
&\lambda \|f_{D_n,\lambda}\|_{H_\sigma}^2 + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\lambda}) - \mathcal{R}_{L,P}^* \\
&\leq 4\lambda \|f_0\|_{H_\sigma}^2 + 4\mathcal{R}_{L,P}(f_0) - 4\mathcal{R}_{L,P}^* + 4r + 5\varepsilon \\
&\leq C_1 \left(\lambda \sigma^{-d} + \sigma^{2t} + \sigma^{-\frac{d}{1-p}} \lambda^{-\frac{p}{1-p}} \left(\frac{n}{(\log n)^{\frac{2}{\gamma}}} \right)^{\frac{1}{1-p}} \varepsilon^{-\frac{2p}{1-p}} \tau + \varepsilon \right),
\end{aligned}$$

where C_1 is a constant independent of n, λ, σ, τ, and ε. Here, λ, σ, and n need satisfy the additional requirement $\sigma, \lambda \in (0,1]$ with $\lambda\sigma^2 \geq n^{-2}$. Now, optimizing over ε by using [42, Lemma A.1.5], we get

$$\lambda \|f_{D_n,\lambda}\|_{H_\sigma}^2 + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\lambda}) - \mathcal{R}_{L,P}^* \leq C_2 \left(\lambda \sigma^{-d} + \sigma^{2t} + \sigma^{-\frac{d}{1+p}} \lambda^{-\frac{p}{1+p}} \left(\frac{n}{(\log n)^{\frac{2}{\gamma}}} \right)^{\frac{1}{1+p}} \tau \right), \quad (80)$$

where C_2 is a constant independent of n, λ, and σ. By applying [42, Lemma A.1.6], we can optimize the right-hand side of (80) over λ and σ, then we see that for all $\xi > 0$ we can find $p, \zeta \in (0,1)$ sufficiently close to 0 such that the LS-SVM using Gaussian RKHS H_σ and $\lambda_n = n^{-1}$, $\sigma_n = n^{-\frac{1}{2t+d}}$ learns with rate $n^{-\frac{2t}{2t+d}+\xi}$, since the requirement $\lambda_n \sigma_n^2 \geq n^{-2}$ is automatically satisfied by the assumed $t \geq 1$. □

Proof of Theorem 4.10. Theorem 4.6 yields

$$\Upsilon(f_{\boldsymbol{D}_n^{(j)},\Upsilon}) + \mathbb{E}_P h_{\widehat{f}_{\boldsymbol{D}_n^{(j)},\Upsilon}} \leq 2\Upsilon(f_0) + 4\mathbb{E}_P h_{f_0} + 4r + 5\varepsilon + 2\delta$$

with probability $\mu \otimes \nu$ not less than $1 - 16e^{-\tau}$. Using (43) and the definition (42) we then easily obtain the assertion. □

References

[1] R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034, 2008.

[2] P. Alquier and O. Wintenberger. Fast rates in learning with dependent observations. *JMLR: Workshop and Conference Proceedings*, pages 1–15, 2012.

[3] V. Araújo, S. Galatolo, and M. J. Pacifico. Decay of correlations for maps with uniformly contracting fibers and logarithm law for singular hyperbolic attractors. *Math. Z.*, 276(3-4):1001–1048, 2014.

[4] V. Baladi. *Positive Transfer Operators and Decay of Correlations*, volume 16 of *Advanced Series in Nonlinear Dynamics*. World Scientific Publishing Co., Inc., River Edge, NJ, 2000.

[5] V. Baladi. Decay of correlations. In *Smooth ergodic theory and its applications (Seattle, WA, 1999)*, volume 69 of *Proc. Sympos. Pure Math.*, pages 297–325. Amer. Math. Soc., Providence, RI, 2001.

[6] P. Bálint and I. Melbourne. Decay of correlations and invariance principles for dispersing billiards with cusps, and related planar billiard flows. *J. Stat. Phys.*, 133(3):435–447, 2008.

[7] D. Belomestny. Spectral estimation of the Lévy density in partially observed affine models. *Stochastic Process. Appl.*, 121(6):1217–1244, 2011.

[8] M. Benedicks and L.-S. Young. Markov extensions and decay of correlations for certain Hénon maps. *Astérisque*, (261):xi, 13–56, 2000. Géométrie complexe et systèmes dynamiques (Orsay, 1995).

[9] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.*, 4(5):861–894, 2004.

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

[11] R. Bowen. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*. Lecture Notes in Mathematics, Vol. 470. Springer-Verlag, Berlin-New York, 1975.

[12] R. C. Bradley. *Introduction to Strong Mixing Conditions. Vol. 1*. Kendrick Press, Heber City, UT, 2007.

[13] N. Chernov. Decay of correlations and dispersing billiards. *J. Statist. Phys.*, 94(3-4):513–556, 1999.

[14] N. Chernov and H.-K. Zhang. Billiards with polynomial mixing rates. *Nonlinearity*, 18(4):1527–1553, 2005.

[15] P. Collet, S. Martinez, and B. Schmitt. Exponential inequalities for dynamical measures of expanding maps of the interval. *Probab. Theory Related Fields*, 123(3):301–322, 2002.

[16] Y. A. Davydov. Convergence of distributions generated by stationary stochastic processes. *Theory Probab. Appl.*, 13(4):691–696, 1968.

[17] J. Dedecker, P. Doukhan, G. Lang, J. R. León R., S. Louhichi, and C. Prieur. *Weak dependence: with examples and applications*, volume 190 of *Lecture Notes in Statistics*. Springer, New York, 2007.

[18] J. Dedecker and C. Prieur. New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields*, 132(2):203–236, 2005.

[19] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[20] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, 2001.

[21] M. Eberts and I. Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Stat.*, 7:1–42, 2013.

[22] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[23] H. Hang and I. Steinwart. Fast learning from α-mixing observations. *J. Multivariate Anal.*, 127:184–199, 2014.

[24] F. Hofbauer and G. Keller. Ergodic properties of invariant measures for piecewise monotonic transformations. *Math. Z.*, 180(1):119–140, 1982.

[25] I. A. Ibragimov. Some limit theorems for stationary processes. *Theory Probab. Appl.*, 7(4):349–382, 1962.

[26] R. S. Kallabis and M. H. Neumann. An exponential inequality under weak dependence. *Bernoulli*, 12(2):333–350, 2006.

[27] G. Keller and T. Nowicki. Spectral theory, zeta functions and the distribution of periodic points for Collet-Eckmann maps. *Comm. Math. Phys.*, 149(1):31–69, 1992.

[28] C. Liverani. Decay of correlations. *Ann. of Math. (2)*, 142(2):239–301, 1995.

[29] S. Luzzatto and I. Melbourne. Statistical properties and decay of correlations for interval maps with critical points and singularities. *Comm. Math. Phys.*, 320(1):21–35, 2013.

[30] R. Markarian. Billiards with polynomial decay of correlations. *Ergodic Theory Dynam. Systems*, 24(1):177–197, 2004.

[31] V. Maume-Deschamps. Exponential inequalities and functional estimations for weak dependent data; applications to dynamical systems. *Stoch. Dyn.*, 6(4):535–560, 2006.

[32] F. Merlevède, M. Peligrad, and E. Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, volume 5 of *Inst. Math. Stat. Collect.*, pages 273–292. Inst. Math. Statist., Beachwood, OH, 2009.

[33] D. S. Modha and E. Masry. Minimum complexity regression estimation with weakly dependent observations. *IEEE Trans. Inform. Theory*, 42(6, part 2):2133–2145, 1996.

[34] E. Rio. Sur le théorème de Berry-Esseen pour les suites faiblement dépendantes. *Probab. Theory Related Fields*, 104(2):255–282, 1996.

[35] M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U. S. A.*, 42:43–47, 1956.

[36] D. Ruelle. A measure associated with axiom-A attractors. *Amer. J. Math.*, 98(3):619–654, 1976.

[37] M. Rychlik. Bounded variation and invariant measures. *Studia Math.*, 76(1):69–80, 1983.

[38] P.-M. Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000.

[39] J. G. Sinaĭ. Gibbs measures in ergodic theory. *Russ. Math. Surveys*, 27:21–69, 1972.

[40] I. Steinwart. Two oracle inequalities for regularized boosting classifiers. *Stat. Interface*, 2(3):271–284, 2009.

[41] I. Steinwart and M. Anghel. Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. *Ann. Statist.*, 37(2):841–875, 2009.

[42] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008.

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

[44] I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory*, 52(10):4635–4643, 2006.

[45] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In S. Dasgupta and A. Klivans, editors, *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93. 2009.

[46] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[47] O. Wintenberger. Deviation inequalities for sums of weakly dependent time series. *Electron. Commun. Probab.*, 15:489–503, 2010.

[48] L.-S. Young. Statistical properties of dynamical systems with some hyperbolicity. *Ann. of Math. (2)*, 147(3):585–650, 1998.

[49] L.-S. Young. Recurrence times and rates of mixing. *Israel J. Math.*, 110:153–188, 1999.

[50] J. Zhang. Sieve estimates via neural network for strong mixing processes. *Stat. Inference Stoch. Process.*, 7(2):115–135, 2004.

Hanyuan Hang
Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
E-Mail: Hanyuan.Hang@mathematik.uni-stuttgart.de

Ingo Steinwart
Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
E-Mail: steinwart@mathematik.uni-stuttgart.de
WWW: http://www.isa.uni-stuttgart.de/Steinwart/

Erschienene Preprints ab Nummer 2012-001

Komplette Liste: `http://www.mathematik.uni-stuttgart.de/preprints`

2015-016 *Hang, H.; Steinwart, I.:* A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning

2015-015 *Steinwart, I.:* Representation of Quasi-Monotone Functionals by Families of Separating Hyperplanes

2015-014 *Muhammad, F.; Steinwart, I.:* An SVM-like Approach for Expectile Regression

2015-013 *Nava-Yazdani, E.:* Splines and geometric mean for data in geodesic spaces

2015-012 *Kimmerle, W.; Köster, I.:* Sylow Numbers from Character Tables and Group Rings

2015-011 *Györfi, L.; Walk, H.:* On the asymptotic normality of an estimate of a regression functional

2015-010 *Gorodski, C, Kollross, A.:* Some remarks on polar actions

2015-009 *Apprich, C.; Höllig, K.; Hörner, J.; Reif, U.:* Collocation with WEB–Splines

2015-008 *Kabil, B.; Rodrigues, M.:* Spectral Validation of the Whitham Equations for Periodic Waves of Lattice Dynamical Systems

2015-007 *Kollross, A.:* Hyperpolar actions on reducible symmetric spaces

2015-006 *Schmid, J.; Griesemer, M.:* Well-posedness of Non-autonomous Linear Evolution Equations in Uniformly Convex Spaces

2015-005 *Hinrichs, A.; Markhasin, L.; Oettershagen, J.; Ullrich, T.:* Optimal quasi-Monte Carlo rules on higher order digital nets for the numerical integration of multivariate periodic functions

2015-004 *Kutter, M.; Rohde, C.; Sändig, A.-M.:* Well-Posedness of a Two Scale Model for Liquid Phase Epitaxy with Elasticity

2015-003 *Rossi, E.; Schleper, V.:* Convergence of a numerical scheme for a mixed hyperbolic-parabolic system in two space dimensions

2015-002 *Döring, M.; Györfi, L.; Walk, H.:* Exact rate of convergence of kernel-based classification rule

2015-001 *Kohler, M.; Müller, F.; Walk, H.:* Estimation of a regression function corresponding to latent variables

2014-021 *Neusser, J.; Rohde, C.; Schleper, V.:* Relaxed Navier-Stokes-Korteweg Equations for Compressible Two-Phase Flow with Phase Transition

2014-020 *Kabil, B.; Rohde, C.:* Persistence of undercompressive phase boundaries for isothermal Euler equations including configurational forces and surface tension

2014-019 *Bilyk, D.; Markhasin, L.:* BMO and exponential Orlicz space estimates of the discrepancy function in arbitrary dimension

2014-018 *Schmid, J.:* Well-posedness of non-autonomous linear evolution equations for generators whose commutators are scalar

2014-017 *Margolis, L.:* A Sylow theorem for the integral group ring of $PSL(2, q)$

2014-016 *Rybak, I.; Magiera, J.; Helmig, R.; Rohde, C.:* Multirate time integration for coupled saturated/unsaturated porous medium and free flow systems

2014-015 *Gaspoz, F.D.; Heine, C.-J.; Siebert, K.G.:* Optimal Grading of the Newest Vertex Bisection and H^1-Stability of the L_2-Projection

2014-014 *Kohler, M.; Krzyżak, A.; Walk, H.:* Nonparametric recursive quantile estimation

2014-013 *Kohler, M.; Krzyżak, A.; Tent, R.; Walk, H.:* Nonparametric quantile estimation using importance sampling

2014-012 *Györfi, L.; Ottucsák, G.; Walk, H.:* The growth optimal investment strategy is secure, too.

2014-011 *Györfi, L.; Walk, H.:* Strongly consistent detection for nonparametric hypotheses

2014-010 *Köster, I.:* Finite Groups with Sylow numbers $\{q^x, a, b\}$

2014-009 *Kahnert, D.:* Hausdorff Dimension of Rings

2014-008 *Steinwart, I.:* Measuring the Capacity of Sets of Functions in the Analysis of ERM

2014-007 *Steinwart, I.:* Convergence Types and Rates in Generic Karhunen-Loève Expansions with Applications to Sample Path Properties

2014-006 *Steinwart, I.; Pasin, C.; Williamson, R.; Zhang, S.:* Elicitation and Identification of Properties

2014-005 *Schmid, J.; Griesemer, M.:* Integration of Non-Autonomous Linear Evolution Equations

2014-004 *Markhasin, L.:* L_2- and $S^r_{p,q}B$-discrepancy of (order 2) digital nets

2014-003 *Markhasin, L.:* Discrepancy and integration in function spaces with dominating mixed smoothness

2014-002 *Eberts, M.; Steinwart, I.:* Optimal Learning Rates for Localized SVMs

2014-001 *Giesselmann, J.:* A relative entropy approach to convergence of a low order approximation to a nonlinear elasticity model with viscosity and capillarity

2013-016 *Steinwart, I.:* Fully Adaptive Density-Based Clustering

2013-015 *Steinwart, I.:* Some Remarks on the Statistical Analysis of SVMs and Related Methods

2013-014 *Rohde, C.; Zeiler, C.:* A Relaxation Riemann Solver for Compressible Two-Phase Flow with Phase Transition and Surface Tension

2013-013 *Moroianu, A.; Semmelmann, U.:* Generalized Killling spinors on Einstein manifolds

2013-012 *Moroianu, A.; Semmelmann, U.:* Generalized Killing Spinors on Spheres

2013-011 *Kohls, K; Rösch, A.; Siebert, K.G.:* Convergence of Adaptive Finite Elements for Control Constrained Optimal Control Problems

2013-010 *Corli, A.; Rohde, C.; Schleper, V.:* Parabolic Approximations of Diffusive-Dispersive Equations

2013-009 *Nava-Yazdani, E.; Polthier, K.:* De Casteljau's Algorithm on Manifolds

2013-008 *Bächle, A.; Margolis, L.:* Rational conjugacy of torsion units in integral group rings of non-solvable groups

2013-007 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups over composition algebras

2013-006 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups, semifields, and translation planes

2013-005 *Eck, C.; Kutter, M.; Sändig, A.-M.; Rohde, C.:* A Two Scale Model for Liquid Phase Epitaxy with Elasticity: An Iterative Procedure

2013-004 *Griesemer, M.; Wellig, D.:* The Strong-Coupling Polaron in Electromagnetic Fields

2013-003 *Kabil, B.; Rohde, C.:* The Influence of Surface Tension and Configurational Forces on the Stability of Liquid-Vapor Interfaces

2013-002 *Devroye, L.; Ferrario, P.G.; Györfi, L.; Walk, H.:* Strong universal consistent estimate of the minimum mean squared error

2013-001 *Kohls, K.; Rösch, A.; Siebert, K.G.:* A Posteriori Error Analysis of Optimal Control Problems with Control Constraints

2012-013 *Diaz Ramos, J.C.; Dominguez Vázquez, M.; Kollross, A.:* Polar actions on complex hyperbolic spaces

2012-012 *Moroianu; A.; Semmelmann, U.:* Weakly complex homogeneous spaces

2012-011 *Moroianu; A.; Semmelmann, U.:* Invariant four-forms and symmetric pairs

2012-010 *Hamilton, M.J.D.:* The closure of the symplectic cone of elliptic surfaces

2012-009 *Hamilton, M.J.D.:* Iterated fibre sums of algebraic Lefschetz fibrations

2012-008 *Hamilton, M.J.D.:* The minimal genus problem for elliptic surfaces

2012-007 *Ferrario, P.:* Partitioning estimation of local variance based on nearest neighbors under censoring

2012-006 *Stroppel, M.:* Buttons, Holes and Loops of String: Lacing the Doily

2012-005 *Hantsch, F.:* Existence of Minimizers in Restricted Hartree-Fock Theory

2012-004 *Grundhöfer, T.; Stroppel, M.; Van Maldeghem, H.:* Unitals admitting all translations

2012-003 *Hamilton, M.J.D.:* Representing homology classes by symplectic surfaces

2012-002 *Hamilton, M.J.D.:* On certain exotic 4-manifolds of Akhmedov and Park

2012-001 *Jentsch, T.:* Parallel submanifolds of the real 2-Grassmannian