The Annals of Statistics
2017, Vol. 45, No. 2, 708–743
DOI: 10.1214/16-AOS1465

A BERNSTEIN-TYPE INEQUALITY FOR SOME MIXING PROCESSES AND DYNAMICAL SYSTEMS WITH AN APPLICATION TO LEARNING

BY HANYUAN HANG AND INGO STEINWART

University of Stuttgart

We establish a Bernstein-type inequality for a class of stochastic processes that includes the classical geometrically ϕ-mixing processes, Rio's generalization of these processes and 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 binary classification, we obtain essentially the same rate as for i.i.d. processes, and for 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 and statistics. For example, [37] used these inequalities to develop a nonasymptotic theory for model selection with applications to variable selection, change points detection and statistical learning. Similarly, the analysis of various methods from nonparametric statistics and machine learning crucially depends on these inequalities; see, for example, [23, 24, 26, 53]. 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, also. For example, generalizations of Bernstein's inequality to α-mixing and ϕ-mixing processes have been

Received March 2015; revised March 2016.

MSC2010 subject classifications. Primary 60E15; secondary 60G10, 37D20, 60F10, 68T05, 62G08, 62M10.

Key words and phrases. Bernstein-type inequalities, mixing processes, dynamical systems, nonparametric classification and regression, support vector machines (SVMs).

found in [12, 41, 42] and [48], respectively. Among many other applications, the Bernstein-type inequality established in [12] was used in [63] to obtain convergence rates for sieve estimates from α-mixing strictly stationary processes in the special case of neural networks. Furthermore, [27] applied the Bernstein-type inequality in [42] 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 [9, 41] 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, [3] obtained the optimal learning rate for ϕ-mixing processes by applying the Bernstein-type inequality established in [48].

Unfortunately, dynamical systems are, in general, not α-mixing, and hence the above-mentioned mixing concepts and Bernstein-type inequalities become invalid. To deal with such nonmixing processes, Rio [43] introduced so-called $\tilde{\phi}$-mixing coefficients, which extend the classical ϕ-mixing coefficients. For dynamical systems with exponentially decreasing, *modified* $\tilde{\phi}$-coefficients, [61] established 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 significantly stronger than Rio's original $\tilde{\phi}$-mixing, so it remains unclear when the Bernstein-type inequality in [61] 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, [38] 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 [27], 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 binary classification, least squares and quantile regression, we can recover the (essentially) optimal rates recently found for the i.i.d. case (see [25]) when the data is generated by certain geometrically $\mathcal{C}$-mixing processes. Finally, we derive learning rates for binary classification on dynamical systems and 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 [52] to learning rates. In this regard, recall that for stochastic dynamical systems, statistical inference for parameter estimation has been widely investigated in a variety of articles; see [40] and the reference therein. For example, for dynamical systems such as shifts of finite type with Gibbs measures and Axiom A attractors with SRB measures, [39] established the consistency of maximum likelihood estimation.

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 inequality to previously established Bernstein-type 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 an oracle inequality for ERM, learning rates for SVMs for binary classification, least squares regression and quantile regression and an oracle inequality for forecasting certain dynamical systems. Numerical experiments are implemented in Section 5. The last section contains the proof of the main result and the remaining proofs for Sections 2 and 4 can be found in the Supplementary Material [28].

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 notation. 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. Furthermore, 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 $f \in L_1(\mu)$. Moreover, 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

$$\|f\|_{\mathcal{C}} := \|f\|_\infty + \|f\| \tag{2.1}$$

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

$$\mathcal{C}(Z) := \{f : Z \to \mathbb{R} | \|f\|_{\mathcal{C}} < \infty\}. \tag{2.2}$$

To prove our Bernstein inequality, we further need to make the following technical assumption on the semi-norm:

$$\|e^f\| \le \|e^f\|_\infty \|f\|, \qquad f \in \mathcal{C}(Z). \tag{2.3}$$

This assumption will be used once in our proof; see (6.14). Moreover, a closer inspection shows that modulo a change in constants our proof still works if we replace the right-hand side of (2.3) by $c \cdot \|e^f\|_\infty \|f\|$, where c is a constant independent of f. Since the examples we are interested in do not need this additional freedom, we decided to omit the details.

If one views the semi-norm as a norm describing aspects of the smoothness of f, then (2.3) can be viewed as an abstract "chain rule". The Examples 2.2–2.4 below illustrate this interpretation.

EXAMPLE 2.1. Let Z be an arbitrary set and $\|f\| = 0$ for all $f : Z \to \mathbb{R}$. Then, it is easy to see that $\|e^f\| = \|f\| = 0$. Hence, (2.3) is satisfied.

EXAMPLE 2.2. Let $Z \subset \mathbb{R}^d$ be an open subset. For a continuously differentiable function $f : Z \to \mathbb{R}$, we write

$$\|f\| := \sum_{i=1}^{d} \left\| \frac{\partial f}{\partial z_i} \right\|_\infty .$$

It is well known that $C^1(Z) := \{f : Z \to \mathbb{R} | f$ continuously differentiable and $\|f\|_\infty + \|f\| < \infty\}$ is a Banach space with respect to the norm $\|\cdot\|_\infty + \|\cdot\|$. Moreover, inequality (2.3) holds for all $f \in C^1(Z)$.

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, inequality (2.3) is also valid for $f \in C_{b,\alpha}(Z)$. As usual, we speak of Lipschitz continuous functions if $\alpha = 1$ and write $\mathrm{Lip}(Z) := C_{b,1}(Z)$.

Before presenting the next example, we mention that throughout this paper the underlying set Z of the bounded variation $\|\cdot\|_{\mathrm{BV}(Z)}$ is always assumed to be bounded if not mentioned otherwise.

EXAMPLE 2.4. 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 $\|f\|_{\mathrm{BV}(Z)}$ is bounded. Denote by $\mathrm{BV}(Z)$ the set of all functions of bounded variation. It is well known that $\mathrm{BV}(Z)$ together with $\|f\|_\infty + \|f\|_{\mathrm{BV}(Z)}$ forms a Banach space. Moreover, we have (2.3), that is, we have for all $f \in \mathcal{C}(Z)$:

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

For high-dimensional cases, the chain rule for functions of bounded variation holds as well; see, for example, the first line of the proof of Theorem 3.96 in [4].

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 on Z, which is defined by $\mu_\chi(B) := \mu(\chi^{-1}(B))$, $B \in \mathcal{B}$.

Let $\mathcal{Z} := (Z_n)_{n\geq 0}$ be an Z-valued stochastic process on $(\Omega, \mathcal{A}, \mu)$, and for $0 \leq i \leq j \leq \infty$, denote by $\mathcal{A}_i^j$ the σ-algebra generated by $(Z_i, \dots, Z_j)$. The process $\mathcal{Z}$ is called *stationary* if $\mu_{(Z_{i_1+i},\dots,Z_{i_n+i})} = \mu_{(Z_{i_1},\dots,Z_{i_n})}$ for all $n, i, i_1, \dots, i_n \geq 1$. 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

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

Several dependency coefficients for $\mathcal{Z}$ can be expressed in terms of the set of such correlations for restricted sets of functions ψ and φ. The following definition, which is taken from [38], 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 (2.1) for some semi-norm $\|\cdot\|$. Then, for $n \geq 0$, we define the $\mathcal{C}$-mixing coefficients by

$$\phi_{\mathcal{C}}(\mathcal{Z}, n) := \sup\{\operatorname{cor}(\psi, h(Z_{k+n})) : k \geq 0, \psi \in B_{L_1(\mathcal{A}_0^k,\mu)}, h \in B_{\mathcal{C}(Z)}\} \tag{2.4}$$

and the time-reversed $\mathcal{C}$-mixing coefficients by

$$\phi_{\mathcal{C},\mathrm{rev}}(\mathcal{Z}, n) := \sup\{\operatorname{cor}(h(Z_k), \varphi) : k \geq 0, h \in B_{\mathcal{C}(Z)}, \varphi \in B_{L_1(\mathcal{A}_{k+n}^\infty,\mu)}\}. \tag{2.5}$$

Let $(d_n)_{n\geq 0}$ be a strictly positive sequence converging to 0. 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

$$d_n := c\exp(-bn^\gamma), \qquad n \geq 1, \tag{2.6}$$

for some constants $c > 0$, $b > 0$, 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

$$\operatorname{cor}(\psi, h(Z_{k+n})) \leq \|\psi\|_{L_1(\mu)} \|h\|_{\mathcal{C}} d_n, \tag{2.7}$$

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

$$\operatorname{cor}(h(Z_k), \varphi) \leq \|h\|_{\mathcal{C}} \|\varphi\|_{L_1(\mu)} d_n. \tag{2.8}$$

REMARK 2.6. If $\|\cdot\| \equiv 0$, we obtain the classical ϕ-mixing coefficients; see Example 2.7. In the new case $\|\cdot\| \not\equiv 0$, the resulting $\mathcal{C}$-norm satisfies $\|f\|_{\mathcal{C}} \geq \|f\|_\infty$ and, therefore, the mixing coefficients admit fewer functions. Inequality (2.3), which in some sense can be viewed as a generalized chain rule (see Examples 2.2–2.4) suggests that the considered functions are "smoother" than the ones in the ϕ-mixing case and, therefore statistical changes of small spatial nature in x do not have such a large impact on $h(x)$, if h is smooth. In other words, even if the trajectory $x_1, \dots, x_n$ stays in a certain region for a while, this does not impact the empirical average $\frac{1}{n}\sum_{i=1}^n h(x_i)$ as much as it would for nonsmooth h.

In the rest of this section we consider examples of (time-reversed) $\mathcal{C}$-mixing processes.

EXAMPLE 2.7. Assume that $\mathcal{Z}$ is a stationary ϕ-mixing process [30] with rate $(d_n)_{n\geq 0}$. By [20], inequality (1.1), we then have

$$\operatorname{cor}(\psi, \varphi) \leq \|\psi\|_{L_1(\mu)} \|\varphi\|_{L_\infty(\mu)} d_n, \qquad n \geq 1, \tag{2.9}$$

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 $\varphi := h(Z_{k+n})$, we then see that (2.7) is satisfied, that is, $\mathcal{Z}$ is $\mathcal{C}$-mixing with rate $(d_n)_{n\geq 0}$. Finally, by similar arguments we can deduce that time-reversed ϕ-mixing processes ([14], Section 3.13) are also time-reversed $\mathcal{C}$-mixing with the same rate. In other words, we have found

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

EXAMPLE 2.8. Let $Z \subset \mathbb{R}$ be an interval. To deal with processes that are not α-mixing [44], Rio [43] introduced the following relaxation of ϕ-mixing coefficients:

$$\begin{aligned} \tilde{\phi}(\mathcal{Z}, n) &:= \sup_{k\geq 0,\, f\in B_{\mathrm{BV}(Z)}} \big\|\mathbb{E}\big(f(Z_{k+n})|\mathcal{A}_0^k\big) - \mathbb{E}f(Z_{k+n})\big\|_\infty \\ &= \sup\big\{\operatorname{cor}\big(\psi, h(Z_{k+n})\big) : k \geq 0, \psi \in B_{L_1(\mathcal{A}_0^k,\mu)}, h \in B_{\mathrm{BV}(Z)}\big\} \end{aligned} \tag{2.10}$$

and an analogous time-reversed coefficient

$$\begin{aligned} \tilde{\phi}_{\mathrm{rev}}(\mathcal{Z}, n) &:= \sup_{k\geq 0,\, f\in B_{\mathrm{BV}(Z)}} \big\|\mathbb{E}\big(f(Z_k)|A_{k+n}^\infty\big) - \mathbb{E}f(Z_k)\big\|_\infty \\ &= \sup\big\{\operatorname{cor}\big(h(Z_k), \varphi\big) : k \geq 0, \varphi \in B_{L_1(\mathcal{A}_{k+n}^\infty,\mu)}, h \in B_{\mathrm{BV}(Z)}\big\}, \end{aligned}$$

where the two identities follow from [22], Lemma 4. In other words, we have

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

Moreover, recall that some uniformly expanding maps are $\tilde{\phi}$-mixing but not α-mixing; see, for example, [21], page 41. The first picture of Figure 1 summarizes the relations between ϕ, $\tilde{\phi}$, and $\phi_{\mathcal{C}}$.

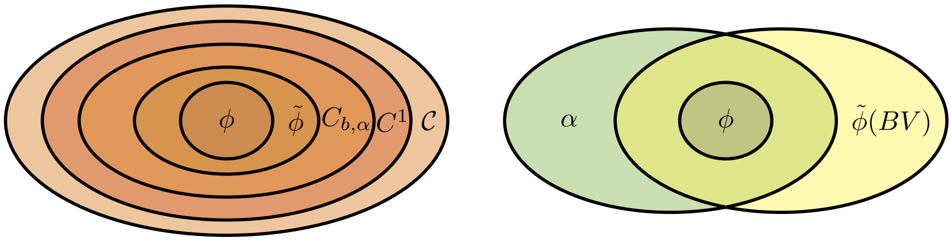

FIG. 1. *Relationship between various mixing processes. Note that $\tilde{\phi}$-mixing equals $\mathcal{C}$-mixing with $\mathcal{C} = \mathrm{BV}$; see Example* 2.8.

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 $\varphi = g(T^{k+n})$ for some suitable, measurable $g : \Omega \to \mathbb{R}$. Let us now suppose that $\|\cdot\|_{\mathcal{C}(\Omega)}$ is defined by (2.1) for some semi-norm $\|\cdot\|$. For $h \in \mathcal{C}(\Omega)$, we then find

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

The next result shows that $\mathcal{T} := (T^n)_{n\geq 0}$ is time-reversed $\mathcal{C}$-mixing even if we only have generic constants $C(h, g)$ in (2.8).

THEOREM 2.9. *Let $(\Omega, \mathcal{A}, \mu, T)$ be a dynamical system, $\|\cdot\|_{\mathcal{C}}$ be defined by* (2.1) *for some semi-norm $\|\cdot\|$, and $(d_n)_{n\geq 0}$ be a strictly positive sequence converging to* 0. *Then the stochastic process $\mathcal{T} := (T^n)_{n\geq 0}$ is time-reversed $\mathcal{C}$-mixing with rate $(cd_n)_{n\geq 0}$ for some constant $c > 0$, if and only if for all $h \in \mathcal{C}(\Omega)$ and all $g \in L_1(\mu)$ there exists a constant $C(h, g) > 0$ such that*

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

It follows from Theorem 2.9 that $\mathcal{T}$ 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 smooth expanding maps on manifolds, piecewise expanding maps, uniformly hyperbolic attractors and nonuniformly hyperbolic uni-modal maps are time-reversed geometrically $\mathcal{C}$-mixing with related spaces $\mathrm{BV}(Z)$ or $C_{b,\alpha}(Z)$; see [60], Propositions 2.7, 3.8, Corollary 4.11 and Theorem 5.15, respectively. Moreover, [38] 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 $\mathrm{BV}(Z)$ or $\mathrm{Lip}(Z)$. Recently, [31] proved that under certain regularity and expansion assumptions on the transformation T (see [31], Conditions (H1)–(H5)), a real valued dynamical system embedded in $\mathbb{R}^2$ turns out to be time-reversed geometrically $\mathcal{C}$-mixing with $\mathcal{C}$ specified as in [31].

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

$$|\mathrm{cor}_{T,n}(h,g)| \le C(h,g)\cdot \exp(-bn^{\gamma}), \qquad n \ge 0, \tag{2.11}$$

for some constants $b > 0$, $\gamma > 0$, and $C(h,g) \ge 0$ depending on h and g. For example, for continuously differentiable h and g, [45, 50] proved (2.11) for two closely related classes of systems, more precisely, Axiom A diffeomorphisms with Gibbs invariant measures and topological Markov chains, which are also known as subshifts of finite type; see also [13]. These results were then extended by [29, 47] 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 behaviour which is characterized by nonuniform hyperbolicity, such as quadratic interval maps (see [32, 62] and the Hénon map [10]), and then extended to chaotic systems with singularities by [34] and specifically to Sinai billiards in a torus by [18, 62]. For some of these extensions, such as smooth expanding dynamics, smooth nonuniformly hyperbolic systems and hyperbolic systems with singularities, we refer to [6] as well. Recently, for h of bounded variation and bounded g, [35] obtained (2.11) for a class of piecewise smooth one-dimensional maps with critical points and singularities. Moreover, [5] has deduced (2.11) 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 inequality.

This section presents our main result, a Bernstein inequality for geometrically (time-reversed) $\mathcal{C}$-mixing process.

THEOREM 3.1. *Let $\mathcal{Z} := (Z_n)_{n\ge 0}$ be a Z-valued stationary geometrically (time-reversed) $\mathcal{C}$-mixing process on $(\Omega, \mathcal{A}, \mu)$ with $\|\cdot\|_{\mathcal{C}}$ be defined by (2.1) for some semi-norm $\|\cdot\|$ satisfying (2.3), and $P := \mu_{Z_0}$. Moreover, let $h : Z \to \mathbb{R}$ be a function such that $h \in \mathcal{C}(Z)$ with $\mathbb{E}_P h = 0$ and assume that there exist some $A > 0$, $B > 0$, and $\sigma \ge 0$ such that $\|h\| \le A$, $\|h\|_\infty \le B$, and $\mathbb{E}_P h^2 \le \sigma^2$. Then, for all $\varepsilon > 0$ and all*

$$n \ge n_0 := \max\left\{\min\left\{m \ge 3 : m^2 \ge \frac{808c(3A+B)}{B} \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \ge 4\right\}, e^{\frac{3}{b}}\right\}, \tag{3.1}$$

we have

$$\mu\left(\frac{1}{n}\sum_{i=1}^{n}h(Z_i)\geq\varepsilon\right)\leq 2\exp\left(-\frac{n\varepsilon^2}{8(\log n)^{\frac{2}{\gamma}}(\sigma^2+\varepsilon B/3)}\right), \tag{3.2}$$

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

$$\mu\left(\frac{1}{n}\sum_{i=1}^{n}h(Z_i)\geq\sqrt{\frac{8(\log n)^{\frac{2}{\gamma}}\sigma^2\tau}{n}}+\frac{8(\log n)^{\frac{2}{\gamma}}B\tau}{3n}\right)\leq 2e^{-\tau}. \tag{3.3}$$

Note that besides the additional logarithmic factor $4(\log n)^{\frac{2}{\gamma}}$ and the constant 2 in front of the exponential, (3.2) 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 Bernstein-type 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. Consider an expanding map T of the interval [0, 1] such that T satisfies Conditions 1–3 in [22], Section 4, and the ergodic measure μ satisfies [22], (4.8). Then [22], Theorem 2, shows that for all separately Lipschitz continuous $f : Z^n \to \mathbb{R}$, and all $\varepsilon\geq 0$, $n\geq 1$, we have

$$\mu\bigl(f(Z_0,\dots,Z_{n-1})-\mathbb{E}f(Z_0,\dots,Z_{n-1})\geq\varepsilon\bigr)\leq\exp\left(-\frac{\varepsilon^2 n}{C}\right), \tag{3.4}$$

where C is some constant only depending on the Lipschitz constants of f. The same result has also been proved by [19], Theorem III.1, for piecewise regular expanding maps. Furthermore, [22], Theorem 2, established (3.4) for causal functions of stationary sequences, iterated random functions, and Markov kernels. Moreover, [15] obtained (3.4) by proving a Devroye inequality in [16] for a large class of nonuniformly hyperbolic dynamical systems including families of piecewise hyperbolic maps, scattering billiards, unimodal and Hénon-like maps. More recently, [17] established (3.4) for Axiom A attractors, Hénon attractors for Benedicks–Carleson parameters, piecewise hyperbolic maps like the Lozi attractor, some billiards with convex scatterers. Notice that, compared to our inequality, almost all the above exponential inequalities hold for more general statistics of the form $f(Z_0,\dots,Z_{n-1})$. This great flexibility, however, is paid by a weaker bound as soon as the variance σ^2 becomes sufficiently small. For the analysis of many learning algorithms, this difference matters since by localization the concentration inequality is applied in situations in which σ^2 depends on n and $\sigma_n^2\to 0$ as $n\to\infty$.

EXAMPLE 3.3. For dynamical systems with exponentially decreasing, modified $\tilde{\phi}$-coefficients (see [61], Condition (3.1)), [61], Theorem 3.1, provides a Bernstein 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 (3.2) is bounded by

$$\exp\left(-\frac{C\varepsilon^2 n}{\sigma^2 + \varepsilon \log f(n)}\right) \tag{3.5}$$

for some constant C independent of n and $f(n)$ being some function monotonically increasing in n. Modulo the factor $\log f(n)$ and the constant C the bound (3.5) is the same as the one for i.i.d. processes. Moreover, if $f(n)$ grows polynomially, cf. [61], Section 3.3, then (3.5) has the same asymptotic behaviour as our bound. However, the required exponential form of Condition (3.1) in [61], that is,

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

for some $c, b > 0$ and all $n \geq 1$, where $\mathbf{Z}_{k+n}^{k+2n-1} := (Z_{k+n}, \dots, Z_{k+2n-1})$ 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

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

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. Therefore, geometrically $\mathcal{C}$-mixing is weaker than Condition (3.1) in [61], or more precisely, processes satisfying Condition (3.1) in [61] are $\tilde{\phi}$-mixing [see (2.10)], which is stronger than geometrically $\mathcal{C}$-mixing; see Figure 1. Moreover, our result holds for all $\gamma > 0$, while [61] 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 $\gamma > 0$, [42], Theorem 4.3, bounds the left-hand side of (3.2) by

$$(1 + 4e^{-2}c)\exp\left(-\frac{3\varepsilon^2 n^{(\gamma)}}{6\sigma^2 + 2\varepsilon B}\right) \qquad \text{with } n^{(\gamma)} \asymp n^{\frac{\gamma}{\gamma+1}} \tag{3.6}$$

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 (2.7) (see, e.g., [14], Example 7.11), and conversely, not every process satisfying (2.7) is necessarily α-mixing; see Figure 1 and the discussion in Section 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.

EXAMPLE 3.5. For an α-mixing stationary sequence of centered and bounded random variables satisfying $\alpha(n) \le \exp(-2cn)$ for some $c > 0$, [41], Theorem 2, bounds the left-hand side of (3.2) by

$$\exp\left(-\frac{C\varepsilon^2 n}{v^2 + \varepsilon B(\log n)^2 + n^{-1}B^2}\right), \tag{3.7}$$

where $C > 0$ is some constant and

$$v^2 := \sigma^2 + 2 \sum_{2 \le i \le n} |\operatorname{cov}(X_1, X_i)|. \tag{3.8}$$

By a covariance inequality for α-mixing processes (see [20], the corollary to Lemma 2.1), we obtain $v^2 \le C_\delta \|X_1\|_{2+\delta}^2$ for an arbitrary $\delta > 0$ and a constant C_δ only depending on δ. If the additional $\delta > 0$ is ignored, (3.7) has therefore the same asymptotic behaviour as our bound. In general, however, the additional δ does influence the asymptotic behaviour. 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 (3.7) instead. Finally, note that in general the bound (3.7) and ours are not comparable; see again Figure 1.

In particular, inequality (3.7) can be applied to geometrically ϕ-mixing processes with $\gamma = 1$. By using the covariance inequality (1.1) for ϕ-mixing processes in [20], we can bound v^2 defined as in (3.8) by $C\sigma^2$ with some constant C independent of n. Modulo the term $n^{-1}B$ in the denominator, the bound (3.7) thus 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 (3.2) by

$$\exp\left(-\frac{n\varepsilon^2}{\tilde{\sigma}^2 + \varepsilon B \log n}\right), \tag{3.9}$$

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

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

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 (3.9) coincides with ours. Again, this bound and our result are not comparable; see Figure 1.

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 geometrically $\mathcal{C}$-mixing processes. More precisely, in Section 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 Section 4.2, we illustrate this oracle inequality by deriving learning rates for SVMs. Finally, in Section 4.3, we derive learning rates for binary classification on dynamical systems and present an oracle inequality for forecasting of dynamical systems.

4.1. *Oracle inequality for CR-ERMs.*

In the following, X always denotes a measurable space if not mentioned otherwise and $Y \subset \mathbb{R}$ always is a closed subset. Recall that in (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 [53], 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

$$L(x, y, \widehat{t}) \le L(x, y, t), \tag{4.1}$$

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 [53], 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 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 [53], Lemma 2.23. In particular, the classical least squares loss as well as the pinball losses 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 and that are satisfied by most examples mentioned above.

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 $L(x, y, t) \le 1$ and locally Lipschitz continuous, that is,

$$\bigl|L(x, y, t) - L\bigl(x, y, t'\bigr)\bigr| \le \bigl|t - t'\bigr|, \tag{4.2}$$

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

Given a loss 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 make the L-risk

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

as small as possible. 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.

Let us now describe the learning algorithms we are interested in. To this end, 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 $\| \cdot \|_\infty$. Hence, for all $\varepsilon > 0$, the covering number $\mathcal{N}(\mathcal{F}, \| \cdot \|_\infty, \varepsilon)$, (see, e.g., [53], Definition 6.19) is always finite. Moreover, we write

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

for a training set of length n that is distributed according to the first n components of the $X \times Y$-valued process $\mathcal{Z} = (Z_i)_{i \geq 1}$. Let D_n be the empirical measure associated to D. The risk of an $f : X \to \mathbb{R}$ with respect to this measure, that is,

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

is called the empirical L-risk.

With these preparations, we can now introduce the class of learning methods [53], Definition 6.1, we are interested in; see also [53], Definition 7.18.

DEFINITION 4.2. 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

$$\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,D_n}(\widehat{f}_{D_n,\Upsilon}) \leq \inf_{f \in \mathcal{F}} \bigl(\Upsilon(f) + \mathcal{R}_{L,D_n}(f)\bigr) + \delta \tag{4.3}$$

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 (4.3) the unclipped loss is considered, and hence CR-ERMs do not necessarily minimize the regularized clipped empirical risk $\Upsilon(\cdot) + \mathcal{R}_{L,D_n}(\cdot)$. 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 (4.3) the clipped function is considered. However, if we have a minimizer of the unclipped regularized risk, then it automatically satisfies (4.3). In particular, ERM decision functions satisfy (4.3) for the regularizer $\Upsilon := 0$ and $\delta := 0$, and SVM decision functions satisfy (4.3) for the regularizer $\Upsilon := \lambda\|\cdot\|_H^2$ and $\delta := 0$. In other words, ERM and SVMs are CR-ERMs.

Before we present the oracle inequality for δ-CR-ERMs, we need to introduce a few more notation. Let $\mathcal{F}$ be a hypothesis set in the sense of Definition 4.2. For

$$r^* := \inf_{f\in\mathcal{F}} \Upsilon(f) + \mathcal{R}_{L,P}(\widehat{f}) - \mathcal{R}_{L,P}^* \tag{4.4}$$

and $r > r^*$, we write

$$\mathcal{F}_r := \{f \in \mathcal{F} : \Upsilon(f) + \mathcal{R}_{L,P}(\widehat{f}) - \mathcal{R}_{L,P}^* \le r\}. \tag{4.5}$$

Then we have $r^* \le 1$, since $L(x, y, 0) \le 1$, $0 \in \mathcal{F}$, and $\Upsilon(0) = 0$. Furthermore, assume that $L \circ \widehat{f} \in \mathcal{C}$ for the considered $\mathcal{C}$ and that there exists a monotonic increasing sequence $(A_r)_{r\in(0,1]}$ in r such that

$$\|L \circ \widehat{f}\| \le A_r \qquad \text{for all } f \in \mathcal{F}_r \text{ and } r \in (0, 1], \tag{4.6}$$

where $\|\cdot\|$ is a semi-norm satisfying (2.3). Because of the Definition (4.5), it is easily to conclude that $\|L \circ \widehat{f}\| \le A_1$ 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

$$\ln \mathcal{N}(\mathcal{F}_r, \|\cdot\|_\infty, \varepsilon) \le \varphi(\varepsilon) r^p. \tag{4.7}$$

Note that there are actually many hypothesis sets satisfying Assumption (4.7), see [27], Section 4, for some examples.

Now the oracle inequality for δ-CR-ERMs reads as follows.

THEOREM 4.3. *Let $\mathcal{Z} := (Z_n)_{n\ge 0}$ be a Z-valued stationary geometrically (time-reversed) $\mathcal{C}$-mixing process on $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\ge 0}$ as in* (2.6), *$\|\cdot\|_{\mathcal{C}}$ be defined by* (2.1) *for some semi-norm $\|\cdot\|$ satisfying* (2.3), *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 $f_{L,P}^*$ satisfying $L \circ f_{L,P}^* \in \mathcal{C}$ and constants $\vartheta \in [0, 1]$ and $V \ge 1$ such that*

$$\mathbb{E}_P\big(L \circ \widehat{f} - L \circ f_{L,P}^*\big)^2 \le V \cdot \big(\mathbb{E}_P\big(L \circ \widehat{f} - L \circ f_{L,P}^*\big)\big)^{\vartheta}, \qquad f \in \mathcal{F}, \tag{4.8}$$

where $\mathcal{F}$ is a hypothesis set with $0 \in \mathcal{F}$. We define r^*, $\mathcal{F}_r$, and A_r by (4.4), (4.5), *and* (4.6), *respectively and assume that* (4.7) *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 with $L \circ f_0 \in \mathcal{C}$ and $L \circ \widehat{f_0} \in \mathcal{C}$, and $A_0, A^* \ge 0$, $B_0 \ge 1$ be constants such that $\|L \circ f_0\| \le A_0$,*

$\|L \circ \widehat{f}_0\| \le A_0$, $\|L \circ f^*_{L,P}\| \le A^*$ *and* $\|L \circ f_0\|_\infty \le B_0$. *Then, for all fixed* $\varepsilon > 0$, $\delta \ge 0$, $\tau \ge 1$, *and*

$$n \ge n_0^* := \max\left\{\min\left\{m \ge 3 : m^2 \ge K \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \ge 4\right\}, e^{\frac{3}{b}}\right\} \tag{4.9}$$

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

$$r \ge \max\left\{\left(\frac{c_V(\log n)^{\frac{2}{\gamma}}(\tau + \varphi(\varepsilon/2)2^p r^p)}{n}\right)^{\frac{1}{2-\vartheta}}, \frac{20(\log n)^{\frac{2}{\gamma}} B_0\tau}{n}, r^*\right\} \tag{4.10}$$

with $c_V := 512(12V + 1)/3$, *every learning method defined by* (4.3) *satisfies with probability* μ *not less than* $1 - 16e^{-\tau}$:

$$\begin{aligned}&\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P}\\ &\quad < 2\Upsilon(f_0) + 4\mathcal{R}_{L,P}(f_0) - 4\mathcal{R}^*_{L,P} + 4r + 5\varepsilon + 2\delta.\end{aligned} \tag{4.11}$$

Let us first discuss the existence of a Bayes decision function $f^*_{L,P}$. For example, if a distance-based loss L is convex and of lower growth $p \in (1, \infty)$ [53], Definition 2.35, then there always exists a Bayes decision function $f^*_{L,P}$ if Assumption 4.1 holds. Indeed, under Assumption 4.1 the set $\mathcal{M} := \{f \in L_p(P_X) : \mathcal{R}_{L,P}(f) \le 1\}$ is nonempty. Moreover, [53], Lemma 2.38, shows that there exists a constant $c_{L,p} > 0$ independent of P such that, for all measurable $f \in \mathcal{M}$, we have

$$\|f\|^p_{L_p(P_X)} \le c_{L,p}\left(\mathcal{R}_{L,P}(f) + \int_X \int_{\mathbb{R}} |y|^p \, dP(x, y) + 1\right) \le c_{L,p}(M^p + 2),$$

since the distribution P is defined on $X \times [-M, M]$. By [53], Theorem A.6.9, we then conclude that there exists a Bayes decision function $f^*_{L,P}$. Moreover, for the least squares loss, the asymmetric least squares loss and many of the distance-based losses including pinball loss and Huber's loss, there exists a Bayes decision function $f^*_{L,P}$. Finally, for margin-based losses, which are used for binary classification and probability estimation, the specific form of the Bayes function is known in several cases; see, for example, [53], Figure 3.1, and the corresponding calculations in [53], Examples 3.6–3.8.

Let us now briefly discuss the variance bound (4.8). For example, if $Y = [-M, M]$ and L is the least squares loss, then it is well known that (4.8) is satisfied for $V := 16M^2$ and $\vartheta = 1$; see, for example, [53], Example 7.3. Moreover, under some assumptions on the distribution P, [54] established a variance bound of the form (4.8) for the pinball loss used for quantile regression. In addition, for the hinge loss, (4.8) is satisfied for $\vartheta := q/(q + 1)$, if Tsybakov's noise assumption [59] holds for q; see [53], Theorem 8.24. Finally, based on [11, 51] 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 also wonder, why the constants A_0 and B_0 are necessary in Theorem 4.3, 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 provides bounds on $\|L \circ f\|$ for specific loss functions. These bounds will be used in the subsequent sections to apply Theorem 4.3.

LEMMA 4.4. *Let $X \subset \mathbb{R}^d$ and $f : X \to \mathbb{R}$ be bounded*:

(i) *Let $Y = (-1 - \varepsilon, 1 + \varepsilon)$ with $\varepsilon > 0$, $Z := X \times Y$, and $f \in \mathrm{BV}(X)$. Then, for the hinge loss L, we have $\|L \circ f\|_{\mathrm{BV}(Z)} \le (1 + \varepsilon)^2 \|f\|_{\mathrm{BV}(X)} + 2(1 + \varepsilon)\|f\|_\infty \mathrm{vol}(X)$.*

(ii) *Let $Y \subset [-M, M]$ with $M > 0$ and $f \in \mathrm{Lip}(X)$. Then, for the least squares loss L, we have $|L \circ f|_1 \le 2\sqrt{2}(M + \|f\|_\infty)(1 + |f|_1)$.*

(iii) *Let $Y \subset [-M, M]$ with $M > 0$ and $f \in \mathrm{Lip}(X)$. Then, for the τ-pinball loss L, we have $|L \circ f|_1 \le \sqrt{2}(1 + |f|_1)$.*

To illustrate the oracle inequality of Theorem 4.3, we will now use it to derive learning rates [53], Lemma 6.5, for some algorithms with observations coming from (time-reversed) $\mathcal{C}$-mixing processes. Our first example, which will later be important for hyper-parameter selection, considers empirical risk minimization over a finite set. Further examples are presented in the following subsections.

EXAMPLE 4.5 (ERM). Let the assumptions on $\mathcal{F}$, L, P and $\mathcal{C}$ in Theorem 4.3 hold. Moreover, assume that $\mathcal{F}$ is finite and $\|f\|_\infty \le M$ and $\|L \circ f\| \le A$ for all $f \in \mathcal{F}$. Then, for accuracy $\delta := 0$, the learning method described by (4.3) is ERM, and Theorem 4.3 shows by some simple estimates that for $n \ge n_0$ as in (4.9) with $K = 1212c(5A + A^* + 1)$, the inequality

$$\begin{aligned}\mathcal{R}_{L,P}(f_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P} \le 8 \inf_{f \in \mathcal{F}} (\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}) \\ + 4\left(\frac{c_V (\log n)^{\frac{2}{\gamma}} (\tau + \ln |\mathcal{F}|)}{3n}\right)^{1/(2-\vartheta)} \\ + \frac{80(\log n)^{\frac{2}{\gamma}} \tau}{n}\end{aligned}$$

holds with probability μ not less than $1 - 16e^{-\tau}$.

Note that in the i.i.d. case we have $\gamma = \infty$. Therefore, besides some constants and the restrictions on n, the oracle inequality (4.11) is an exact analogue to standard oracle inequality for ERM learning from i.i.d. processes; see, for example, [53], Theorem 7.2.

4.2. *Learning rates for SVMs.* In this section, we apply the developed theory to support vector machines (SVMs) using the hinge loss, the least squares loss, and the pinball loss for learning with observations from (time-reversed) $\mathcal{C}$-mixing processes.

Let us begin by briefly recalling SVMs,; see [53] 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

$$f_{D_n,\lambda} = \underset{f \in H}{\arg\min}\big(\lambda \|f\|_H^2 + \mathcal{R}_{L,D_n}(f)\big). \tag{4.12}$$

In particular, SVMs using the hinge losses are called hinge SVMs, SVMs using the least-squares loss are called least-squares SVMs (LS-SVMs), and SVMs using the τ-pinball loss are called SVMs for quantile regression.

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

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

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 [55]. The entropy numbers for Gaussian kernels (see, e.g., [53], Theorem 6.27), and the equivalence of covering and entropy numbers (see, e.g., [53], Lemma 6.21) yield that

$$\ln \mathcal{N}\big(B_{H_\sigma}, \|\cdot\|_\infty, \varepsilon\big) \le a\sigma^{-d}\varepsilon^{-2p}, \qquad \varepsilon > 0, \tag{4.13}$$

for all $p \in (0, 1)$, where a is a constant only depending on P.

Recall that for SVMs we always have $f_{D_n,\lambda} \in \lambda^{-1/2} B_H$ (see [53], (5.4)) where B_H denotes the closed unit ball of the RKHS H. Consequently, we can choose the hypothesis set as $\mathcal{F} = \lambda^{-1/2} B_{H_\sigma}$. Then (4.5) implies $\mathcal{F}_r \subset r^{1/2}\lambda^{-1/2} B_{H_\sigma}$, and hence we have

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

For the function φ in Theorem 4.3, we can thus choose

$$\varphi(\varepsilon) := a\sigma^{-d}\lambda^{-p}\varepsilon^{-2p}. \tag{4.14}$$

Now we can apply the oracle inequality in Theorem 4.3 to derive the learning rates for the SVMs using Gaussian kernels. In the following examples, $B_{\ell_2^d}$ denotes the closed unit ball of d-dimensional Euclidean space ℓ_2^d.

EXAMPLE 4.6 (Binary classification with Gaussian kernels). Let $X \subset B_{\ell_2^d}$, $Y := \{-1, 1\}$, $Z := X \times Y$, and $\mathcal{Z}$ be a geometrically (time-reversed) $\mathcal{C}(Z)$-mixing process with $\mathcal{C}(Z)$ being $L_\infty(Z)$ or $\mathrm{BV}(Z)$. Moreover, let L be the hinge loss, P be a distribution on $\mathbb{R}^d \times Y$ that has margin-noise exponent $\beta \in (0, \infty)$ and noise exponent $q \in [0, \infty]$ and its marginal distribution on $\mathbb{R}^d$ is concentrated on X. Assume that the Bayes decision function $f^*_{L,P}$ given by $f^*_{L,P}(x) = \mathrm{sign}(P(y = 1|x) - 0.5)$ satisfies $f^*_{L,P} \in \mathcal{C}(X)$. Then for all $\xi > 0$, the Hinge loss SVM using Gaussian RKHS H_σ learns with rate

$$n^{-\frac{\beta(q+1)}{\beta(q+2)+d(q+1)}+\xi}, \tag{4.15}$$

which equals the best known rate for i.i.d. processes; see, for example, [53], Theorem 8.26.

Roughly speaking, the margin-noise exponent measures the amount of noise in the labeling and the concentration of the marginal distribution *in the vicinity of the decision boundary*, while the noise exponent, introduced by Mammen and Tsybakov (see [36, 59]) measures the *total* amount of noise. We refer to [53], Chapter 8, for precise definitions and illustrative examples. In the one-dimensional case, the assumption $f^*_{L,P} \in \mathrm{BV}(X)$ is satisfied, if $f^*_{L,P}$ has only finitely many points of discontinuity. For high-dimensional cases, we refer to [4], Chapter 3. Finally, note that considering smoother classes $\mathcal{C}$ such as $C^1(Z)$ or $\mathrm{Lip}(Z)$ does not make sense for pure binary classification, since the Bayes function $f^*_{L,P}$ is a step function and, therefore, usually not contained in such smooth $\mathcal{C}$.

In the next example, $B^t_{2s,\infty}$ denotes the usual Besov space with the smoothness parameter t; for more details, we refer to [2, 58].

EXAMPLE 4.7 (Least Square Regression with Gaussian kernels). Let $Y := [-M, M]$ for some $M > 0$, $Z := \mathbb{R}^d \times Y$, and $\mathcal{Z}$ be a geometrically (time-reversed) $\mathcal{C}(Z)$-mixing process with $\mathcal{C}(Z)$ being $L_\infty(Z)$, $\mathrm{BV}(Z)$, or $C_{b,\alpha}(Z)$. Moreover, let P be a distribution on Z whose marginal distribution on $\mathbb{R}^d$ is concentrated on $X \subset B_{\ell_2^d}$ and absolutely continuous w.r.t. the Lebesgue measure μ on $\mathbb{R}^d$. We denote the corresponding density $g : \mathbb{R}^d \to [0, \infty)$ and assume $\mu(\partial X) = 0$ and $g \in L_q(\mu)$ for some $q \geq 1$. Moreover, assume that the Bayes decision function $f^*_{L,P} = \mathbb{E}_P(Y|x)$ satisfies $f^*_{L,P} \in L_2(\mu) \cap \mathrm{Lip}(\mathbb{R}^d)$ as well as $f^*_{L,P} \in B^t_{2s,\infty}$ 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

$$\lambda_n = \frac{(\log n)^{\frac{2}{\gamma}}}{n} \quad \text{and} \quad \sigma_n = \left(\frac{(\log n)^{\frac{2}{\gamma}}}{n}\right)^{\frac{1}{2t+d}} \tag{4.16}$$

learns with rate

$$n^{-\frac{2t}{2t+d}+\xi}. \tag{4.17}$$

REMARK 4.8. Note that if $s = \infty$ and $t > 1$, then we have $B^t_{2s,\infty} \subset \mathrm{Lip}(\mathbb{R}^d)$; see, for example, [46], Section 2.1.2. In this case, the assumptions on the Bayes decision function $f^*_{L,P}$ made in Examples 4.7 and 4.9 are identical to that for the i.i.d. case; see [25], Sections 3 and 4. Moreover, as larger values of t lead to smoother functions contained in $B^t_{2s,\infty}$, it is not surprising that the rates become better, the larger t gets.

It turns out that modulo the arbitrarily small $\xi > 0$, the learning rates (4.17) equal the optimal rates for i.i.d. processes; see, for example, [56], Theorem 9, or [26], Theorem 3.2.

To achieve these rates, however, we need to set λ_n and σ_n as in (4.16), which in turn requires us to know γ and t. In practice, we usually do not know these values nor their existence. To obtain the above rates without knowledge about γ and t, we can use the training/validation approach TV-SVM; see, for example, [53], Chapters 6.5, 7.4, 8.2. 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 $\epsilon_n \le n^{-1}$ and $\delta_n \le n^{-\frac{1}{2+d}}$. Furthermore, assume that the cardinalities $|\Lambda_n|$ and $|\Sigma_n|$ grow polynomially in n. For a data set $D := ((x_1, y_1), \dots, (x_n, y_n))$, we define $D_1 := ((x_1, y_1), \dots, (x_m, y_m))$ and $D_2 := ((x_{m+1}, y_{m+1}), \dots, (x_n, y_n))$, where $m := \lfloor \frac{n}{2} \rfloor + 1$ and $n \ge 4$. We will use D_1 as a training set by computing the SVM decision functions

$$f_{D_1,\lambda,\sigma} := \arg\min_{f \in H_\sigma} \lambda \|f\|^2_{H_\sigma} + \mathcal{R}_{L,D_1}(f), \qquad (\lambda, \sigma) \in \Lambda_n \times \Sigma_n$$

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

$$\mathcal{R}_{L,D_2}(\widehat{f}_{D_1,\lambda_{D_2},\sigma_{D_2}}) = \min_{(\lambda,\sigma)\in\Lambda_n\times\Sigma_n} \mathcal{R}_{L,D_2}(\widehat{f}_{D_1,\lambda,\sigma}).$$

Then, analogous to the proof of Theorem 3.3 in [25], by using Theorem 4.3 and Example 4.5, one can show that for all $\xi > 0$, the TV-SVM producing the decision functions $f_{D_1,\lambda_{D_2},\sigma_{D_2}}$ learns with the above learning rates (4.17).

The next example discusses learning rates for SVMs for quantile regression. For more information on such SVMs, we refer to [25, 57].

EXAMPLE 4.9 (Quantile regression with Gaussian kernels). Let $Y := [-1, 1]$, $Z := \mathbb{R}^d \times Y$, and $\mathcal{Z}$ be geometrically (time-reversed) $\mathcal{C}(Z)$-mixing processes with $\mathcal{C}(Z)$ being $L_\infty(Z)$, $\mathrm{BV}(Z)$, or $C_{b,\alpha}(Z)$. Moreover, let P be a distribution on Z and Q be the marginal distribution of P on $\mathbb{R}^d$. Assume that $X := \mathrm{supp}Q \subset B_{\ell_2^d}$ and that for Q-almost all $x \in X$, the conditional probability $P(\cdot|x)$ is absolutely continuous w.r.t. the Lebesgue measure on Y and the conditional densities $h(\cdot, x)$ of $P(\cdot|x)$ are uniformly bounded away from 0 and ∞; see also [25], Example 4.5. Moreover, assume that Q is absolutely continuous w.r.t. the Lebesgue measure on X with associated density $g \in L_u(X)$ for some $u \ge 1$.

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 [54], Theorem 2.8, yields a variance bound of the form

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

for all $f : X \to \mathbb{R}$, where V is a suitable constant and L_τ is the τ-pinball loss; see [53], Example 2.43. Arguments similar to the proof of Example 4.7 shows that the essentially optimal learning rate (4.17) can be achieved as well.

Note that the rate (4.17) is for the excess L_τ-risk, but since [54], 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. *Learning from dynamical systems.* In this section, we consider two learning scenarios in which the observations are generated by some dynamical systems plus some (possible) noise. In the following, Ω denotes a compact subset of $\mathbb{R}^d$, $(\Omega, \mathcal{A}, \mu, T)$ is a dynamical system, and X_0 is an Ω-valued random variable describing the true but unknown state at time 0.

4.3.1. *Binary classification on dynamical systems.* Let us assume that our covariates $\mathcal{X} := (X_i)_{i\geq 0}$ are generated by a dynamical system $(\Omega, \mathcal{A}, \mu, T)$, that is $X_i := T^i$, $i \geq 0$, and that the binary labels Y_i are randomly drawn depending on the location X_i. To model this, let $\mathcal{N} := (\epsilon_i)_{i\geq 0}$ be i.i.d. random variables that are uniformly distributed on $[0,1]$ and independent of X_0. Then we model the label generation by setting $Y_i := \mathrm{sign}(\eta(X_i) - \epsilon_i)$, $i \geq 0$, where $\eta : X \to [0,1]$ is a fixed function. In this case, we have

$$P(Y_i = 1 | X_i = x_i) = P_\epsilon\big(\eta(x_i) - \epsilon_i > 0\big) = P_\epsilon\big(\epsilon_i < \eta(x_i)\big) = \eta(x_i),$$

and thus also $P(Y_i = -1 | X_i = x_i) = 1 - \eta(x_i)$. In other words, we can model arbitrary label distributions by choosing a suitable η.

Now let ν be the uniform distribution on $[0,1]$ and define the process $\mathcal{Z} = (Z_i)_{i\geq 0}$ on $Z := \Omega \times [-1,1]$ by $Z_i := (X_i, Y_i)$, $i \geq 0$, and write $P := (\mu \otimes \nu)_{Z_0}$. Recall that the Bayes function of the hinge loss for classification is $f^*_{L,P}(x) = \mathrm{sign}(\eta(x) - 0.5)$ and since this is a step function, the best we can hope for is $f^*_{L,P} \in \mathrm{BV}(\Omega)$. Therefore, applying Theorem 4.3 only makes sense if $\mathcal{Z}$ is time-reversed $\mathcal{C}_{\mathcal{Z}}$-mixing with $\mathrm{BV}(Z) \subset \mathcal{C}_{\mathcal{Z}}$. Obviously, for the time-reversed ϕ-mixing case, the process $\mathcal{Z}$ is also time-reversed ϕ-mixing, that is, time-reversed $C_{\mathcal{Z}}$-mixing with $\mathrm{BV}(Z) \subset \mathcal{C}_{\mathcal{Z}} = L_\infty(Z)$. Therefore, we obtain the same rate as (4.15). However, for time-reversed $\mathrm{BV}(Z)$-mixing dynamical systems we unfortunately only know that the process $\mathcal{Z}$ is time-reversed $\mathcal{C}_{\mathcal{Z}}$-mixing with $\mathcal{C}_{\mathcal{Z}} = \mathrm{Lip}(Z) \subset \mathrm{BV}(Z)$ as the next theorem shows.

THEOREM 4.10. *Let $\Omega \subset \mathbb{R}^d$ be compact and $\mathcal{X}$ as above be time-reversed* $\mathrm{BV}(\Omega)$*-mixing with rate (d_n). If* $\max_{\epsilon \in [0,1]} \|\operatorname{sign}(\eta(\cdot) - \epsilon)\|_{\mathrm{BV}(\Omega)} < \infty$, *then $\mathcal{Z}$ is time-reversed* $\mathrm{Lip}(Z)$*-mixing with rate (cd_n) for some constant c.*

Therefore, if $\mathcal{X}$ is only time-reversed $\mathrm{BV}(\Omega)$-mixing, we cannot combine Theorem 4.10 with the oracle inequality in Theorem 4.3 to obtain learning rates for the hinge SVM. One way to resolve this issue is to use the least squares SVM instead, as the following example shows.

EXAMPLE 4.11 (Least squares binary classification on dynamical systems using Gaussian kernels). Assume that the process $\mathcal{Z}$ defined above is geometrically time-reversed $\mathcal{C}$-mixing dynamical system with $\mathcal{C} = \mathrm{BV}(\Omega)$ an L be the least square loss. Moreover, assume that η and P satisfy the assumptions on $f^*_{L,P}$ and the distribution P in Example 4.7, respectively. Then, for all $\xi > 0$, the LS-SVM using Gaussian RKHS H_σ and λ_n, σ_n as in (4.16) learns with rate

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

If the distribution P has some noise exponent $q \in [0, \infty]$, then [8] implies that the classification excess risk of the same SVM converges to zero with rate

$$n^{-\frac{2t(q+1)}{(2t+d)(q+2)}+\xi},$$

see also [53], Theorem 8.29.

4.3.2. *Forecasting of dynamical systems.* Our second scenario is the forecasting problem of dynamical systems considered in [52]. To this end, let $E > 0$ and assume that all observations from the dynamical system $\mathcal{T} := (T^n)_{n\geq 0}$ 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 $(\Theta, \mathcal{B}, \nu)$ which is (stochastically) independent of $\mathcal{T}$. Therefore, the process of all possible observations $(X_n)_{n\geq 0}$ has the form $X_n = T^n(X_0) + \varepsilon_n$. Given an observation of this process at some arbitrary time, our goal is to forecast the next *observable* state. To do so, we will use the training set $\boldsymbol{D} = ((X_0, X_1), \dots, (X_{n-1}, X_n))$ to build a forecaster $f_{\boldsymbol{D}} : \mathbb{R}^d \to \mathbb{R}^d$ of the form $f_{\boldsymbol{D}} := (f_{\boldsymbol{D}^{(1)}}, \dots, f_{\boldsymbol{D}^{(d)}})$ by training separately d different decision functions $f_{\boldsymbol{D}^{(j)}}$ on the training sets

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

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

TABLE 1
$\mathcal{C}$-mixing properties of the observed process (4.18) subject to properties of the underlying dynamical system

Class of $\mathcal{C}_{\mathcal{T}}$	**Smoothness of T**	**Class of $\mathcal{C}_{\mathcal{Z}}$**
L_∞	L_∞	L_∞
BV	BV	Lip
$C_{b,\alpha}$	$C_{b,\alpha}$	Lip
C^1	C^1	C^1

For a fixed $j \in \{1, \dots, d\}$, we write $X := \Omega + [-E, E]^d$, $Y := \pi_j(X)$ and $Z := X \times Y$. Moreover, we define the $X \times Y$-valued process $\mathcal{Z} = (Z_n)_{n\ge0} = (X_n, Y_n)_{n\ge0}$ on $(\Omega \times \Theta, \mathcal{A} \otimes \mathcal{B}, \mu \otimes \nu)$ by

$$X_n := T^n + \varepsilon_n \quad \text{and} \quad Y_n := \pi_j\big(T^{n+1} + \varepsilon_{n+1}\big). \tag{4.18}$$

In addition, we write $P := (\mu \otimes \nu)_{(X_0,Y_0)}$.

The next result shows that for some specific time-reversed $\mathcal{C}$-mixing dynamical systems $\mathcal{T}$, by imposing some restrictions on $\mathcal{E}$, the process $\mathcal{Z}$ is time-reversed $\mathcal{C}_{\mathcal{Z}}$-mixing for some suitable $\mathcal{C}_{\mathcal{Z}}$.

THEOREM 4.12. *Let $\Omega \subset \mathbb{R}^d$ be compact, $\mathcal{T}$ be time-reversed $\mathcal{C}_{\mathcal{T}}$-mixing with rate (d_n), and $\mathcal{Z}$ be defined by (4.18). Moreover, assume that the $[-E, E]^{2d}$-valued process $\mathcal{N} = (N_i)_{i\ge0}$ on $(\Theta, \mathcal{B}, \nu)$ defined by $N_i(\vartheta) = (\varepsilon_i(\vartheta), \varepsilon_{i+1}(\vartheta))$, $i \ge 0$, is time-reversed $\mathcal{C}_{\mathcal{N}}$-mixing with rate (d_n). Then we have*:

(i) *If $T \in$ BV(Ω), $\mathcal{C}_{\mathcal{T}} =$ BV(Ω), and $\mathcal{C}_{\mathcal{N}} =$ BV$([-E, E]^{2d})$, then $\mathcal{Z}$ is time-reversed* Lip(Z)*-mixing with rate (cd_n) for some constant c.*

(ii) *If $T \in C_{b,\alpha}(\Omega)$, $\mathcal{C}_{\mathcal{T}} = C_{b,\alpha}(\Omega)$, and $\mathcal{C}_{\mathcal{N}} = C_{b,\alpha}([-E, E]^{2d})$, then $\mathcal{Z}$ is time-reversed* Lip(Z)*-mixing with rate (cd_n) for some constant c.*

Analogously, other cases of T and $\mathcal{C}_{\mathcal{T}}$ from the literature can be proved. For sake of clarity, we list them in the Table 1.

To formulate the oracle inequality for the forecasting problem, we need to introduce the following concepts. First, for a 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 $\boldsymbol{L}(X_i - \boldsymbol{f}(X_{i-1}))$ gives a value for the discrepancy between the forecast $\boldsymbol{f}(X_{i-1})$ and the observation of the next state X_i. Then the average forecasting performance is given by

$$\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f}) := \iint \boldsymbol{L}\big(T(x) + \varepsilon_1 - \boldsymbol{f}(x + \varepsilon_0)\big)\nu(d\varepsilon)\mu(dx), \tag{4.19}$$

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

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

We say that $\boldsymbol{L}$ 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 $\boldsymbol{L}$ 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

$$\boldsymbol{L}(\boldsymbol{r}) = \psi(r_1) + \cdots + \psi(r_d), \qquad \boldsymbol{r} = (r_1, \dots, r_d) \in \mathbb{R}^d. \tag{4.20}$$

For separable loss $\boldsymbol{L}$, we have $\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f}) = \sum_{j=1}^d \mathcal{R}_{L,P}(f_{\boldsymbol{D}^{(j)}})$ and $\mathcal{R}_{\boldsymbol{L},\boldsymbol{D}_n}(\boldsymbol{f}_{\boldsymbol{D}}) = \sum_{j=1}^d \mathcal{R}_{L,\boldsymbol{D}_n^{(j)}}(f_{\boldsymbol{D}^{(j)}})$, where $\boldsymbol{D}_n$, $\boldsymbol{D}_n^{(j)}$ are the empirical measures associated to $\boldsymbol{D}$, $\boldsymbol{D}^{(j)}$, respectively.

Finally, let $\boldsymbol{L} : \mathbb{R}^d \to [0, \infty)$ be a clippable loss and $\mathcal{F}$ be a hypothesis set, that is, a set of measurable functions $f : X \to \mathbb{R}$, 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 $\delta \ge 0$, a learning method whose decision functions $\boldsymbol{f}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}} \in \mathcal{F}^d$ satisfy

$$\boldsymbol{\Upsilon}(\boldsymbol{f}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}}) + \mathcal{R}_{\boldsymbol{L},\boldsymbol{D}_n}(\widehat{\boldsymbol{f}}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}}) \le \inf_{\boldsymbol{f} \in \mathcal{F}^d} \bigl(\boldsymbol{\Upsilon}(\boldsymbol{f}) + \mathcal{R}_{\boldsymbol{L},\boldsymbol{D}_n}(\boldsymbol{f})\bigr) + \delta \tag{4.21}$$

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

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

THEOREM 4.13. *Let $\Omega \subset \mathbb{R}^d$ be compact and the stationary stochastic process $\mathcal{Z} := (Z_n)_{n \ge 0}$ defined by* (4.18) *be geometrically time-reversed $\mathcal{C}$-mixing. 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 in this subsection and in Theorem* 4.3. *Finally, let $\boldsymbol{\Upsilon} : \mathcal{F}^d \to [0, \infty)$ be a separable regularizer and $\boldsymbol{f}_0 = (f_1^0, \dots, f_d^0)$ be a vector of functions with f_j^0 satisfying the assumptions on f_0 in Theorem* 4.3. *Then, for all fixed $\boldsymbol{f}_0$, $\varepsilon > 0$, $\delta \ge 0$, $\tau \ge 1$, $n \ge n_0$ as in* (4.9), *and $r \in (0, 1]$ satisfying* (4.10), *every learning method defined by* (4.21) *satisfies with probability $\mu \otimes \nu$ not less than $1 - 16de^{-\tau}$*:

$$\begin{aligned} &\boldsymbol{\Upsilon}(\boldsymbol{f}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}}) + \mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\widehat{\boldsymbol{f}}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}}) - \mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}^* \\ &\quad \le 2\boldsymbol{\Upsilon}(\boldsymbol{f}_0) + 4\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f}_0) - 4\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}^* + 4dr + 5d\varepsilon + 2\delta. \end{aligned} \tag{4.22}$$

Again, this general oracle inequality can be applied to SVMs. Here, we only mention the following example.

EXAMPLE 4.14 (Forecasting with LS-SVMs using Gaussian kernels). Let the assumptions of Theorem 4.13 hold with $\mathcal{C}_{\mathcal{Z}}$ being $L_\infty(Z)$, $\mathrm{BV}(Z)$, or $C_{b,\alpha}(Z)$. Moreover, assume that the assumptions on the distribution $P := (\mu \otimes \nu)_{(X_0,Y_0)}$ and the components f_j^* of the Bayes Forecaster $\boldsymbol{f}_{\boldsymbol{L},\boldsymbol{P}}^* = (f_1^*, \dots, f_d^*)$ are satisfied as in Example 4.7. Then, for all $\xi > 0$, the LS-SVMs using Gaussian RKHS H_σ and λ_n, σ_n as in (4.16) learns with rate $n^{-\frac{2t}{2t+d}+\xi}$.

5. Experiments.

Until now, our investigation has mainly been theoretical. Unfortunately all our results contain some constants that may or may not influence our bounds for moderately sized data sets. It is therefore reasonable to complement our investigation by some empirical simulations.

EXAMPLE 5.1 (Empirical deviation from the mean for the logistic map). It is well known that the logistic map $T(x) = 4 \cdot x \cdot (1-x)$, $x \in (0,1)$, is geometrically $\mathcal{C}$-mixing with $\mathcal{C} = \mathrm{Lip}(0,1)$; see, for example, [38], Theorem 4.7. Moreover, it has the unique invariant Lebesgue density $f(x) = (\pi\sqrt{x(1-x)})^{-1}$, $x \in (0,1)$; see, for example, [33], Example 4.1.2. With start values from the corresponding distribution, we generate n samples $x_1, \dots, x_n$ from the logistic map. For a comparison, we further generate data $x_1, \dots, x_n$ from the i.i.d. process that belongs to the density f. Then we investigate the deviation

$$\epsilon_n := \left| \frac{1}{n} \sum_{i=1}^{n} h(x_i) - \mathbb{E}h(x_i) \right| \tag{5.1}$$

for three different functions h, namely $h_1 = \mathrm{id}_{(0,1)}$,

$$h_2 = (e-1) \cdot \mathbf{1}_{(0,1/3)} + (e^{3x} - 1) \cdot \mathbf{1}_{[1/3,2/3)} + (e^2 - 1) \cdot \mathbf{1}_{[2/3,1)},$$

$$h_3 = \frac{3}{2} \cdot \mathbf{1}_{(0,1/4)} + \frac{9}{4} \cdot \mathbf{1}_{[1/4,1/2)} + \frac{15}{4} \cdot \mathbf{1}_{(1/2,3/4)} + \frac{9}{2} \cdot \mathbf{1}_{[3/4,1)}.$$

Notice that we have $h_1 \in C^1(0,1)$, $h_2 \in \mathrm{Lip}(0,1)$, while $h_3 \in \mathrm{BV}(0,1)$. We repeat the experiments 100 times for each n. In Figure 2, we see that in all these cases, for large samples, the mean error (5.1) of the dynamical system and the i.i.d. process have similar behaviour, but the deviations converge slightly faster in the i.i.d. case.

5.1. *Binary classification for the Gauss map on* (0, 1).

The Gauss map $T(x) = (1/x) \bmod 1$, $x \in (0,1)$, is known to be geometrically $\mathcal{C}$-mixing with $\mathcal{C} = \mathrm{BV}$; see, for example, [38] or [6], Chapter 3, Theorem 4.4. Moreover, it has the unique invariant Lebesgue density $f(x) = 1/(\log 2 \cdot (1+x))$, $x \in (0,1)$; see again [38]. With start values from the distribution with density f, we generate $n = 30{,}000$ samples $(\tilde{x}_i)$ from the Gauss map and add i.i.d. Gaussian noises

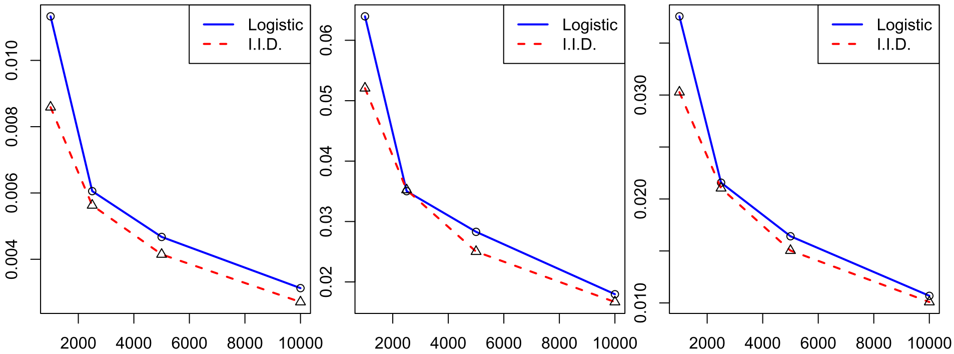

FIG. 2. *The deviation ϵ_n in* (5.1) *for the functions h_1* (*left*), *h_2* (*middle*) *and h_3* (*right*) *and different sample sizes $n = 1000, 2500, 5000, 10{,}000$.*

$\varepsilon_i \sim N(0, 0.03^2)$ to them. These noisy samples $x_i = \tilde{x}_i + \varepsilon_i$ are then labeled as in Section 4.3.1 using the function

$$\eta(x) := \big(\sin(6\pi x) + 1\big)/2.$$

Now our goals are to (a) classify the data sets with SVMs using the hinge loss and (b) estimate the function $2\eta - 1$ using the least square loss. We repeat the experiments 10 times. Every time, the last $n_{\text{test}} = 20{,}000$ data points are used for testing and the first n_{train} data points are used for training including hyper-parameter selection, where the candidate values for λ and σ are taken from a geometrically spaced 10 by 10 grid using 5-fold cross validation. The resulting test errors are box-plotted for both SVMs in Figure 3. In the pictures, we can see that the test er-

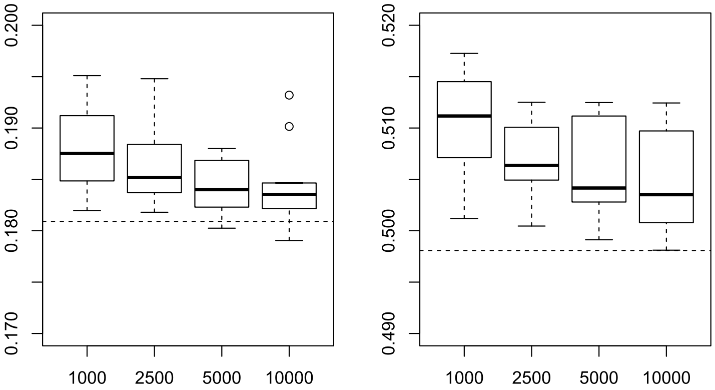

FIG. 3. *Gauss map*: *Box-plots of test errors for an SVM with the hinge loss* (*left*) *and the least squares loss* (*right*). *For the hinge loss SVM*, *the test error is calculated with the help of the classification loss*, *since for classification*, *one is actually interested in misclassification rate*, *which is measured by the classification loss*. *The dotted lines indicate the Bayes error.*

rors are close to the Bayes errors. Moreover, as the training sample size increases, the test errors indicate a decreasing trend.

5.2. *Binary classification on Shub's Solenoid.* Here, we consider a solid torus $\mathbb{T}^2$ in $\mathbb{R}^3$ whose points are represented by means of coordinates (θ, r, s), where the angle $\theta \in [-\pi, \pi]$, r and s are real numbers between -1 and 1 such that $r^2 + s^2 \leq 1$. For ε_1 and ε_2 satisfying $0 < \varepsilon_2 < \varepsilon_1 < 1/2$, Shub's Solenoid [49], Example 4.9, is defined as the mapping $T : \mathbb{T}^2 \to \mathbb{T}^2$ with

$$T(\theta, r, s) = \bigl((2\theta) \bmod (2\pi), \varepsilon_1 \cos\theta + \varepsilon_2 r, \varepsilon_1 \sin\theta + \varepsilon_2 s\bigr).$$

As an archetype for Axiom A diffeomorphisms or, uniformly hyperbolic systems, Shub's Solenoid is known to be geometrically time-reversed $\mathcal{C}$-mixing with $\mathcal{C} = C_{b,\alpha}$; see [6], Chapter 4, and [7].

In our simulation, we select $\varepsilon_1 = 0.25$ and $\varepsilon_2 = 0.125$. With start values from the uniform distribution on $\mathbb{T}^2$, we generate $n = 30{,}000$ samples by iteration, add i.i.d. Gaussian noise $N(0, 0.1^2)$ to them, and transform the data points from the polar coordinates into Euclidean coordinates. Then we do two classification problems with the same set-up as in the previous example using

$$\eta_1(x, y, z) := -\frac{x^3}{81} - \frac{y}{6} - \frac{2z^2}{27} + 1,$$

$$\eta_2(x, y, z) := \exp\biggl(-\frac{x}{6} - \frac{y^2}{9} - \frac{z^2}{36}\biggr) - \frac{1}{2}.$$

The results are reported in Figure 4.

6. Proof of the key result.

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

LEMMA 6.1. *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 $f_n := f(Z_n)$. Finally, assume that we have natural numbers k and l satisfying*

$$2l \cdot \|f\|_{\mathcal{C}} \cdot d_k \leq \|f\|_{L_1(P)}. \tag{6.1}$$

Then we have

$$\mathbb{E}_\mu \prod_{j=0}^{l} f_{jk} \leq 2\|f\|_{L_1(P)}^{l+1}. \tag{6.2}$$

PROOF. The proof will be provided for two cases: $\mathcal{C}$-mixing and time-reversed $\mathcal{C}$-mixing, separately.

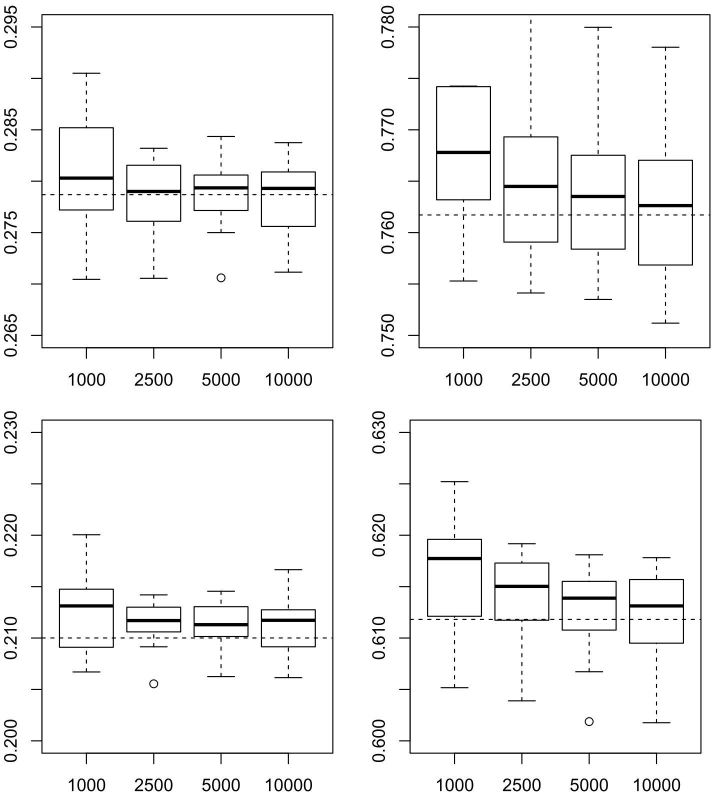

FIG. 4. *Shub's Solenoid*: *Box-plots of test errors for an SVM with the hinge loss* (*left*) *and least squares loss* (*right*). *For the hinge loss SVM*, *the test error is again calculated with the help of the classification loss. The upper two plots report the errors for* η_1 *and the lower two plots report the errors for* η_2. *The dotted lines indicate the Bayes error.*

(i) Suppose that the correlation inequality (2.7) holds. Obviously, the case $f = 0$ P-a.s. is trivial. For $f \neq 0$, we define

$$D_l := \left| \mathbb{E}_\mu \prod_{j=0}^{l} f_{jk} - \prod_{j=0}^{l} \mathbb{E}_\mu f_{jk} \right|. \tag{6.3}$$

Then we have

$$\begin{aligned} D_l \le{} & \left| \mathbb{E}_\mu \left(\left(\prod_{j=0}^{l-1} f_{jk} \right) f_{lk} \right) - \left(\mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk} \right) \mathbb{E}_\mu f_{lk} \right| \\ & + \left| \left(\mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk} \right) \mathbb{E}_\mu f_{lk} - \prod_{j=0}^{l} \mathbb{E}_\mu f_{jk} \right| =: D_l^{(1)} + D_l^{(2)}. \end{aligned}$$

Since the stochastic process $\mathcal{Z}$ is stationary, the decay of correlations (2.7) together with $\psi := \prod_{j=0}^{l-1} f_{jk}$, $h := f$, and the assumption $f \geq 0$ yields

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

Moreover, for the second term, we find

$$\begin{aligned} D_l^{(2)} &= \left| \left(\mathbb{E}_\mu \prod_{j=0}^{l-1} f_{jk} \right) \mathbb{E}_\mu f_{lk} - \left(\prod_{j=0}^{l-1} \mathbb{E}_\mu f_{jk} \right) \mathbb{E}_\mu f_{lk} \right| \\ &= \|f\|_{L_1(P)} D_{l-1}. \end{aligned}$$

These estimates together imply that

$$\begin{aligned} D_l = D_l^{(1)} + D_l^{(2)} &\leq (D_{l-1} + \|f\|_{L_1(P)}^l) \|f\|_{\mathcal{C}} d_k + \|f\|_{L_1(P)} D_{l-1} \\ &= (\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}} d_k) D_{l-1} + \|f\|_{\mathcal{C}} \|f\|_{L_1(P)}^l d_k. \end{aligned} \tag{6.4}$$

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

$$D_l \leq \|f\|_{L_1(P)} \bigl((\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}} d_k)^l - \|f\|_{L_1(P)}^l \bigr). \tag{6.5}$$

When $l = 1$, (6.5) is true because of (2.7). Now let $l \geq 1$ be given and suppose (6.5) is true for l. Then (6.4) and (6.5) imply

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

Thus, (6.5) holds for $l + 1$, and the proof of the induction step is complete. By the principle of induction, (6.5) is thus true for all $l \geq 1$.

Using the binomial formula, we obtain

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

For $i = 0, \dots, l$ we now set $a_i := \binom{l}{i}\|f\|_{L_1(P)}^{l-i}(\|f\|_{\mathcal{C}} d_k)^i$. Then the assumption (6.1) implies for $i = 0, \dots, l-1$

$$\frac{a_{i+1}}{a_i} = \frac{\binom{l}{i+1}\|f\|_{L_1(P)}^{l-i-1}(\|f\|_{\mathcal{C}} d_k)^{i+1}}{\binom{l}{i}\|f\|_{L_1(P)}^{l-i}(\|f\|_{\mathcal{C}} d_k)^i} = \frac{\frac{l!}{(i+1)!(l-i-1)!}}{\frac{l!}{i!(l-i)!}} \frac{\|f\|_{\mathcal{C}} d_k}{\|f\|_{L_1(P)}}$$
$$= \frac{l-i}{i+1} \frac{\|f\|_{\mathcal{C}} d_k}{\|f\|_{L_1(P)}} \le l \cdot \frac{\|f\|_{\mathcal{C}}}{\|f\|_{L_1(P)}} \cdot d_k \le \frac{1}{2}.$$

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

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

This implies

$$D_l \le \|f\|_{L_1(P)} \left(\sum_{i=0}^{l} a_i - \|f\|_{L_1(P)}^l\right) \le \|f\|_{L_1(P)} \bigl(2a_0 - \|f\|_{L_1(P)}^l\bigr)$$
$$= \|f\|_{L_1(P)} \bigl(2\|f\|_{L_1(P)}^l - \|f\|_{L_1(P)}^l\bigr) = \|f\|_{L_1(P)}^{l+1}.$$

Using the definition of D_l, we thus obtain the assertion (6.2).

(ii) Suppose that the correlation inequality (2.8) holds. Again, the case $f = 0$ P-a.s. is trivial. For $f \ne 0$, we estimate D_l defined as in (6.3) in a slightly different way from above:

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

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

To prove Theorem 3.1, we need to introduce some notation. In the following, for $t \in \mathbb{R}$, $\lfloor t \rfloor$ is the largest integer n satisfying $n \le t$, and similarly, $\lceil t \rceil$ is the smallest integer n satisfying $n \ge t$. We now introduce a "lattice method". To this end, we partition the set $\{1, 2, \dots, n\}$ into k sublattices. Each sublattice 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. Define I_i, the indexes of terms in the ith sublattice, as

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

Note that the number of the terms $|I_i|$ equals $l+1$ for $1 \le i \le r$, and equals l for $r+1 \le i \le k$. Furthermore, for $j = 1, \dots, n$, we write $h_j := h(Z_j)$, and for $i = 1, 2, \dots, k$, we define the ith sublattice sum as

$$g_i := \sum_{j \in I_i} h_j = \sum_{j \in I_i} h(Z_j) \tag{6.6}$$

such that

$$S_n = \sum_{j=1}^{n} h_j = \sum_{i=1}^{k} g_i. \tag{6.7}$$

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

$$p_i := \frac{|I_i|}{n}. \tag{6.8}$$

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

LEMMA 6.2. *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* (6.6) *and* (6.7), *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. The convexity of the exponential function together with $\sum_{i=1}^{k} p_i = 1$, (6.7), and (6.8) yields

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

LEMMA 6.3. *Let $\mathcal{Z} := (Z_n)_{n \ge 0}$ be a Z-valued stationary* (*time-reversed*) *$\mathcal{C}$-mixing process on the probability space $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n \ge 0}$, and $P := \mu_{Z_0}$. Moreover, for $h : Z \to [0, \infty)$, we write $h_n := h(Z_n)$. Finally, let k and l be defined as above. Then, for all $t > 0$ satisfying*

$$e^{\frac{t}{|I_i|} h} \in \mathcal{C}(Z) \quad \text{and} \quad 2l \cdot \left\| e^{\frac{t}{|I_i|} h} \right\|_{\mathcal{C}} \cdot d_k \le \left\| e^{\frac{t}{|I_i|} h} \right\|_{L_1(P)}, \tag{6.9}$$

we have

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

PROOF. The ith sublattice sum g_i in (6.6) depends only on h_{i+jk} with j ranging from 0 through $|I_i|-1$. Since $\mathcal{Z}$ is stationary, Lemma 6.1 with $f := \exp(\frac{t}{|I_i|}h)$ then yields

$$\mathbb{E}_\mu \exp\biggl(t\frac{g_i}{|I_i|}\biggr) = \mathbb{E}_\mu \exp\Biggl(\frac{t}{|I_i|}\sum_{j=0}^{|I_i|-1} h_{i+jk}\Biggr) = \mathbb{E}_\mu \exp\Biggl(\frac{t}{|I_i|}\sum_{j=0}^{|I_i|-1} h_{jk}\Biggr)$$

$$= \mathbb{E}_\mu \prod_{j=0}^{|I_i|-1} \exp\biggl(\frac{t}{|I_i|}h_{jk}\biggr) \le 2\biggl(\mathbb{E}_P \exp\biggl(t\frac{h}{|I_i|}\biggr)\biggr)^{|I_i|}. \qquad \square$$

LEMMA 6.4. *Let $\mathcal{Z} := (Z_n)_{n\ge 0}$ be a Z-valued stationary (time-reversed) $\mathcal{C}$-mixing process on the probability space $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\ge 0}$, and $P := \mu_{Z_0}$. Moreover, for $h : Z \to [0, \infty)$, we write $h_n := h(Z_n)$ and suppose that $\mathbb{E}_P h = 0$, $\|h\| \le A$, $\|h\|_\infty \le B$, and $\mathbb{E}_P h^2 \le \sigma^2$ for some $A > 0$, $B > 0$ and $\sigma \ge 0$. Finally, let k and l be defined as above. Then, for all $i = 1, \dots, k$, and all $t > 0$ satisfying $0 < t < 3l/B$ and (6.9), we have*

$$\mathbb{E}_\mu \exp\biggl(t\frac{g_i}{|I_i|}\biggr) \le 2\exp\biggl(\frac{t^2\sigma^2}{2(l - tB/3)}\biggr). \tag{6.10}$$

PROOF. Because of $\|h\|_\infty \le B$ and $2 \cdot 3^{j-2} \le j!$, we obtain

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

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

$$\begin{aligned}\biggl(\mathbb{E}_P \exp\biggl(t\frac{h}{|I_i|}\biggr)\biggr)^{|I_i|} &\le \biggl(1 + \frac{1}{2}\biggl(\frac{t}{|I_i|}\biggr)^2 \sigma^2 \frac{1}{1 - tB/(3|I_i|)}\biggr)^{|I_i|} \\ &\le \biggl(\exp\biggl(\frac{1}{2}\biggl(\frac{t}{|I_i|}\biggr)^2 \sigma^2 \frac{1}{1 - tB/(3|I_i|)}\biggr)\biggr)^{|I_i|} \\ &= \exp\biggl(\frac{t^2\sigma^2}{2(|I_i| - tB/3)}\biggr) \le \exp\biggl(\frac{t^2\sigma^2}{2(l - tB/3)}\biggr),\end{aligned}$$

since the assumed $tB/(3l) < 1$ implies $tB/(3|I_i|) < 1$. Lemma 6.3 then yields the assertion (6.10). □

PROOF OF THEOREM 3.1. For k and l as above, we define

$$t := \frac{l\varepsilon}{\sigma^2 + \varepsilon B/3}. \tag{6.11}$$

Then we have

$$\frac{t}{|I_i|} \le \frac{t}{l} = \frac{\varepsilon}{\sigma^2 + \varepsilon B/3} \le \frac{\varepsilon}{\varepsilon B/3} = \frac{3}{B}. \tag{6.12}$$

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

$$\left\|\exp\left(\frac{t}{|I_i|}h\right)\right\|_\infty \le \exp\left(\frac{3}{B}\cdot B\right) = e^3. \tag{6.13}$$

Then the assumption (2.3) together with the bounds (6.13) and (6.12) implies

$$\left\|\exp\left(\frac{t}{|I_i|}h\right)\right\| \le \left\|\exp\left(\frac{t}{|I_i|}h\right)\right\|_\infty \left\|\frac{t}{|I_i|}h\right\| \le e^3 \cdot \frac{t}{|I_i|}\|h\| \le \frac{3e^3 A}{B}. \tag{6.14}$$

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

$$\left\|\exp\left(\frac{t}{|I_i|}h\right)\right\|_{L_1(P)} = \mathbb{E}_P \exp\left(\frac{t}{|I_i|}h\right) \ge \exp\left(\frac{3}{B}\cdot(-B)\right) = e^{-3}. \tag{6.15}$$

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

$$l = \frac{n-r}{k} \ge \frac{n}{k} - 1 \ge \frac{1}{2}\frac{n}{(\log n)^{\frac{2}{\gamma}}} - 1 \ge \frac{1}{4}\frac{n}{(\log n)^{\frac{2}{\gamma}}}, \tag{6.16}$$

since we have $n \ge 4(\log n)^{\frac{2}{\gamma}}$ for $n \ge n_0$. Now, by (6.13), (6.14), (6.15), (2.6) and (3.1) we obtain

$$\begin{aligned} l \cdot \frac{\|e^{\frac{t}{|I_i|}h}\|_C}{\|e^{\frac{t}{|I_i|}h}\|_{L_1(P)}} \cdot d_k &\le l \cdot \frac{\|e^{\frac{t}{|I_i|}h}\|_\infty + \|e^{\frac{t}{|I_i|}h}\|}{\|e^{\frac{t}{|I_i|}h}\|_{L_1(P)}} \cdot c \cdot \exp(-bk^\gamma) \\ &\le n \cdot \frac{e^3 + \frac{3e^3 A}{B}}{e^{-3}} \cdot c \cdot \exp(-b(\log n)^2) \\ &\le n \cdot \frac{404c(3A+B)}{B} \cdot \exp\left(-b\log n \cdot \frac{3}{b}\right) \\ &\le n \cdot \frac{n^2}{2} \cdot n^{-3} = \frac{1}{2}, \end{aligned} \tag{6.17}$$

that is, the assumption (6.9) is valid.

Summarizing, the value of t defined as in (6.11) satisfies $0 < t < 3l/B$ and the assumption (6.9). In other words, all the requirements on t in Lemma 6.4 are satisfied. Now, for this t, by using Markov's inequality, Lemmas 6.2 and 6.4, we obtain for any $\varepsilon > 0$,

$$\begin{aligned} P\left(\frac{S_n}{n} > \varepsilon\right) &\le \exp(-t\varepsilon)\mathbb{E}_\mu \exp\left(t\frac{S_n}{n}\right) \\ &\le \exp(-t\varepsilon)\sum_{i=1}^k p_i \mathbb{E}_\mu \exp\left(t\frac{g_i}{|I_i|}\right) \\ &\le \exp(-t\varepsilon)\cdot 2\exp\left(\frac{t^2\sigma^2}{2(l-tB/3)}\right)\sum_{i=1}^k p_i \\ &= 2\exp\left(-t\varepsilon + \frac{t^2\sigma^2}{2(l-tB/3)}\right). \end{aligned} \tag{6.18}$$

Substituting the definition of t into the exponent of inequality (6.18) and then using the estimate (6.16), we get

$$\begin{aligned} -t\varepsilon + \frac{t^2\sigma^2}{2(l-tB/3)} &= -\frac{l\varepsilon^2}{\sigma^2+\varepsilon B/3} + \frac{l^2\varepsilon^2}{(\sigma^2+\varepsilon B/3)^2}\cdot\frac{\sigma^2}{2(l-\frac{l\varepsilon B/3}{\sigma^2+\varepsilon B/3})} \\ &= \frac{-l\varepsilon^2}{2(\sigma^2+\varepsilon B/3)} \le -\frac{n\varepsilon^2}{8(\log n)^{\frac{2}{\gamma}}(\sigma^2+\varepsilon B/3)}. \end{aligned}$$

Thus, inequality (3.2) is proved. Setting $\tau := \frac{n\varepsilon^2}{8(\log n)^{2/\gamma}(\sigma^2+\varepsilon B/3)}$ in (3.2), simple transformations and estimations then yield inequality (3.3). □

SUPPLEMENTARY MATERIAL

Supplement to "A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning" (DOI: 10.1214/16-AOS1465SUPP; .pdf). The supplement [28] contains an Appendix, in which we provide the proofs for Sections 2 and 4.

REFERENCES

[1] ADAMCZAK, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** 1000–1034. MR2424985

[2] ADAMS, R. A. and FOURNIER, J. J. F. (2003). *Sobolev Spaces*, 2nd ed. *Pure and Applied Mathematics* (*Amsterdam*) **140**. Elsevier/Academic Press, Amsterdam. MR2424078

[3] ALQUIER, P., LI, X. and WINTENBERGER, O. (2013). Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modeling* **1** 65–93.

[4] AMBROSIO, L., FUSCO, N. and PALLARA, D. (2000). *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Univ. Press, New York. MR1857292

[5] ARAÚJO, V., GALATOLO, S. and PACIFICO, M. J. (2014). Decay of correlations for maps with uniformly contracting fibers and logarithm law for singular hyperbolic attractors. *Math. Z.* **276** 1001–1048. MR3175169

[6] BALADI, V. (2000). *Positive Transfer Operators and Decay of Correlations*. *Advanced Series in Nonlinear Dynamics* **16**. World Scientific, River Edge, NJ. MR1793194

[7] BALADI, V. (2001). Decay of correlations. In *Smooth Ergodic Theory and Its Applications* (*Seattle*, *WA*, 1999). *Proc. Sympos. Pure Math.* **69** 297–325. Amer. Math. Soc., Providence, RI. MR1858537

[8] BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032

[9] BELOMESTNY, D. (2011). Spectral estimation of the Lévy density in partially observed affine models. *Stochastic Process. Appl.* **121** 1217–1244. MR2794974

[10] BENEDICKS, M. and YOUNG, L.-S. (2000). Markov extensions and decay of correlations for certain Hénon maps. *Astérisque* **261** 13–56. MR1755436

[11] BLANCHARD, G., LUGOSI, G. and VAYATIS, N. (2004). On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.* **4** 861–894. MR2076000

[12] BOSQ, D. (1993). Bernstein-type large deviations inequalities for partial sums of strong mixing processes. *Statistics* **24** 59–70. MR1238263

[13] BOWEN, R. (1975). *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*. *Lecture Notes in Mathematics* **470**. Springer, Berlin. MR0442989

[14] BRADLEY, R. C. (2007). *Introduction to Strong Mixing Conditions*. *Vol.* 1. Kendrick Press, Heber City, UT. MR2325294

[15] CHAZOTTES, J.-R., COLLET, P. and SCHMITT, B. (2005). Statistical consequences of the Devroye inequality for processes. Applications to a class of non-uniformly hyperbolic dynamical systems. *Nonlinearity* **18** 2341–2364. MR2165706

[16] CHAZOTTES, J.-R., COLLET, P. and SCHMITT, B. (2005). Devroye inequality for a class of non-uniformly hyperbolic dynamical systems. *Nonlinearity* **18** 2323–2340. MR2166315

[17] CHAZOTTES, J.-R. and GOUËZEL, S. (2012). Optimal concentration inequalities for dynamical systems. *Comm. Math. Phys.* **316** 843–889. MR2993935

[18] CHERNOV, N. (1999). Decay of correlations and dispersing billiards. *J. Stat. Phys.* **94** 513–556. MR1675363

[19] COLLET, P., MARTINEZ, S. and SCHMITT, B. (2002). Exponential inequalities for dynamical measures of expanding maps of the interval. *Probab. Theory Related Fields* **123** 301–322. MR1918536

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

[21] DEDECKER, J., DOUKHAN, P., LANG, G., LEÓN, J. R., LOUHICHI, S. and PRIEUR, C. (2007). *Weak Dependence*: *With Examples and Applications*. *Lecture Notes in Statistics* **190**. Springer, New York. MR2338725

[22] DEDECKER, J. and PRIEUR, C. (2005). New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields* **132** 203–236. MR2199291

[23] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093

[24] DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation*. Springer, New York. MR1843146

[25] EBERTS, M. and STEINWART, I. (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Stat.* **7** 1–42. MR3020412

[26] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York. MR1920390

[27] HANG, H. and STEINWART, I. (2014). Fast learning from α-mixing observations. *J. Multivariate Anal.* **127** 184–199. MR3188886

[28] HANG, H. and STEINWART, I. (2016). Supplement to “A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning.” DOI:10.1214/16-AOS1465SUPP.

[29] HOFBAUER, F. and KELLER, G. (1982). Ergodic properties of invariant measures for piecewise monotonic transformations. *Math. Z.* **180** 119–140. MR0656227

[30] IBRAGIMOV, I. A. (1962). Some limit theorems for stationary processes. *Theory Probab. Appl.* **7** 349–382.

[31] JAGER, L., MAES, J. and NINET, A. (2015). Exponential decay of correlations for a real-valued dynamical system embedded in $\mathbb{R}_2$. *C. R. Math. Acad. Sci. Paris* **353** 1041–1045. MR3419857

[32] KELLER, G. and NOWICKI, T. (1992). Spectral theory, zeta functions and the distribution of periodic points for Collet–Eckmann maps. *Comm. Math. Phys.* **149** 31–69. MR1182410

[33] LASOTA, A. and MACKEY, M. C. (1985). *Probabilistic Properties of Deterministic Systems*. Cambridge Univ. Press, Cambridge. MR0832868

[34] LIVERANI, C. (1995). Decay of correlations. *Ann. of Math.* (2) **142** 239–301. MR1343323

[35] LUZZATTO, S. and MELBOURNE, I. (2013). Statistical properties and decay of correlations for interval maps with critical points and singularities. *Comm. Math. Phys.* **320** 21–35. MR3046988

[36] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618

[37] MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

[38] MAUME-DESCHAMPS, V. (2006). Exponential inequalities and functional estimations for weak dependent data; applications to dynamical systems. *Stoch. Dyn.* **6** 535–560. MR2285515

[39] MCGOFF, K., MUKHERJEE, S., NOBEL, A. and PILLAI, N. (2015). Consistency of maximum likelihood estimation for some dynamical systems. *Ann. Statist.* **43** 1–29. MR3285598

[40] MCGOFF, K., MUKHERJEE, S. and PILLAI, N. S. (2012). Statistical inference for dynamical systems: A review. Preprint. Available at arXiv:1204.6265.

[41] MERLEVÈDE, F., PELIGRAD, M. and RIO, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V*: *The Luminy Volume*. *Inst. Math. Stat. Collect.* **5** 273–292. IMS, Beachwood, OH. MR2797953

[42] MODHA, D. S. and MASRY, E. (1996). Minimum complexity regression estimation with weakly dependent observations. *IEEE Trans. Inform. Theory* **42** 2133–2145. MR1447519

[43] RIO, E. (1996). Sur le théorème de Berry–Esseen pour les suites faiblement dépendantes. *Probab. Theory Related Fields* **104** 255–282. MR1373378

[44] ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. USA* **42** 43–47. MR0074711

[45] RUELLE, D. (1976). A measure associated with axiom-A attractors. *Amer. J. Math.* **98** 619–654. MR0415683

[46] RUNST, T. and SICKEL, W. (1996). *Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations*. *De Gruyter Series in Nonlinear Analysis and Applications* **3**. de Gruyter, Berlin. MR1419319

[47] RYCHLIK, M. (1983). Bounded variation and invariant measures. *Studia Math.* **76** 69–80. MR0728198

[48] SAMSON, P.-M. (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.* **28** 416–461. MR1756011

[49] SHUB, M. (1987). *Global Stability of Dynamical Systems*. Springer, New York. MR0869255

[50] SINAĬ, J. G. (1972). Gibbs measures in ergodic theory. *Russ. Math. Surveys* **27** 21–69.

[51] STEINWART, I. (2009). Two oracle inequalities for regularized boosting classifiers. *Stat. Interface* **2** 271–284. MR2540086

[52] STEINWART, I. and ANGHEL, M. (2009). Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. *Ann. Statist.* **37** 841–875. MR2502653

[53] STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer, New York. MR2450103

[54] STEINWART, I. and CHRISTMANN, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli* **17** 211–225. MR2797989

[55] STEINWART, I., HUSH, D. and SCOVEL, C. (2006). An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory* **52** 4635–4643. MR2300845

[56] STEINWART, I., HUSH, D. and SCOVEL, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of the* 22*nd Annual Conference on Learning Theory* (S. Dasgupta and A. Klivans, eds.) 79–93. Available at http://www.cs.mcgill.ca/~colt2009/papers/038.pdf.

[57] TAKEUCHI, I., LE, Q. V., SEARS, T. D. and SMOLA, A. J. (2006). Nonparametric quantile estimation. *J. Mach. Learn. Res.* **7** 1231–1264. MR2274404

[58] TRIEBEL, H. (2010). *Theory of Function Spaces*. Birkhäuser/Springer, Basel. MR3024598

[59] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

[60] VIANA, M. (1997). *Stochastic Dynamics of Deterministic Systems* **21**. IMPA, Brazil.

[61] WINTENBERGER, O. (2010). Deviation inequalities for sums of weakly dependent time series. *Electron. Commun. Probab.* **15** 489–503. MR2733373

[62] YOUNG, L.-S. (1998). Statistical properties of dynamical systems with some hyperbolicity. *Ann. of Math.* (2) **147** 585–650. MR1637655

[63] ZHANG, J. (2004). Sieve estimates via neural network for strong mixing processes. *Stat. Inference Stoch. Process.* **7** 115–135. MR2061181

INSTITUTE FOR STOCHASTICS AND APPLICATIONS
FACULTY 8: MATHEMATICS AND PHYSICS
UNIVERSITY OF STUTTGART
D-70569 STUTTGART
GERMANY
E-MAIL: hanghn@mathematik.uni-stuttgart.de
ingo.steinwart@mathematik.uni-stuttgart.de