Statistical Learning of Kernel-Based Methods for non-i.i.d. Observations

Von der Fakultät Mathematik und Physik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

Vorgelegt von

Hanyuan Hang

aus VR China

Hauptberichter:	Prof. Dr. Ingo Steinwart
Mitberichter:	Prof. Dr. Andreas Christmann
	Prof. Dr. Gilles Blanchard
Tag der mündlichen Prüfung:	24. April 2015

Institut für Stochastik und Anwendungen
der Universität Stuttgart

2015

Abstract

Statistical learning theory has proved itself in many practical applications such as computer vision, speech recognition, bioinformatics, etc. So far, most results in statistical learning theory presume that successive data points are independent of one another. This is mathematically convenient, but clearly not always suitable for non-i.i.d. processes including many time series. For instance, most of the techniques have been developed in ways which have rendered it impossible to apply it immediately to time series forecasting problems. To address these problems, recent work has adapted key results such as the concentration inequalities and the resulting oracle inequalities to the situations where time widely-separated data points are asymptotically independent. Motivated by this, in this thesis, we will establish a new oracle inequality for generic regularized empirical risk minimization algorithms based on a generic form of a Bernstein inequality and use this oracle inequality to derive learning rates from two classes of non-i.i.d. processes called α- and $\mathcal{C}$-mixing processes.

Applying this oracle inequality to α-mixing processes, we derive learning rates for some learning methods such as empirical risk minimization (ERM), least squares support vector machines (LS-SVMs) using given generic kernels, and support vector machines (SVMs) using the Gaussian RBF kernels for both least squares and quantile regression. It turns out that for i.i.d. processes our learning rates for ERM and SVMs with Gaussian kernels match, up to some arbitrarily small extra term in the exponent, the optimal rates, while in the remaining cases our rates are at least close to the optimal rates.

For geometrically $\mathcal{C}$-mixing processes that include the classical geometrically ϕ-mixing processes, Rio's generalization of these processes, as well as many time-discrete dynamical systems, we establish a Bernstein-type inequality of the generic form that coincides with the classical Bernstein inequality for i.i.d. data modulo a logarithmic factor and some constants. Applying the oracle inequality to support vector machines using the Gaussian kernels for both least squares and quantile regression, it turns out that the resulting learning rates match, up to some arbitrarily small extra term in the exponent, the optimal rates for i.i.d. processes.

Zusammenfassung

Die statistische Lerntheorie findet viele praktische Anwendungen, beispielsweise in den Bereichen der Bildverarbeitung, der Spracherkennung oder der Bioinformatik. Bisher setzten die meisten Resultate der statistischen Lerntheorie voraus, dass aufeinanderfolgende Datenpunkte unabhängig voneinander sind. Dies ist aus mathematischer Sicht angenehm, für einige Situationen aber nicht geeignet, wie zum Beispiel für nicht-u.i.v. Prozesse einschließlich vieler Zeitreihen. Weiterhin sind bisher einige Techniken in einer Weise konstruiert worden, die es einem unmöglich machen, diese direkt für Vorhersagen von Zeitreihen anzuwenden. Um diese Probleme anzugehen, haben neueste Arbeiten wichtige Ergebnisse geliefert für den Fall, dass Datenpunkte asymptotisch unabhängig sind. Wie zum Beispiel die Konzentrationsungleichungen und die daraus resultierenden Orakelungleichungen. Dadurch motiviert werden wir in dieser Arbeit eine neue Orakelungleichung für allgemeine, regularisierte empirische Risikominimierungsalgorithmen vorstellen, welche auf einer allgemeinen Form der Bernstein-Ungleichung basiert. Darüber hinaus leiten wir Lernraten für zwei Klassen von nicht-u.i.v. Prozessen her, nämlich, den α- und den $\mathcal{C}$-mischenden Prozess.

Unter Verwendung dieser Orakelungleichung für α-mischende Prozesse leiten wir Lernraten für einige Lernmethoden her, wie zum Beispiel für empirische Risikominimierung (ERM), least square support vector machines (LS-SVMs) mit allgemeinen Kernen und support vector machines (SVMs) mit Gaußkernen für least square and Quantilregression. Es stellt sich heraus, dass für u.i.v. Prozesse, unsere erhaltenen Lernraten für ERM und SVMs mit Gaußkernen – bis zu einem gewissen beliebig kleinen zusätzlichen Term im Exponenten – den optimalen Raten entsprechen, während in den übrigen Fällen unsere Raten zumindest nahe an den optimalen Raten sind.

Für geometrische $\mathcal{C}$-mischende Prozesse, welche klassische geometrische ϕ-mischende Prozesse, Rio's Verallgemeinerungen dieser Prozesse, als auch viele zeitdiskrete dynamischen Systeme enthalten, leiten wir eine Bernstein-Typ-Ungleichung her, welche die allgemeine Form besitzt und mit der klassischen Bernstein-Ungleichung für u.i.v. Daten – modulo eines logarithmischen Faktors und einiger Konstante – übereinstimmt. Unter Verwendung der Orakelungleichung für SVMs mit Gausskernen für least square und Quantilregression stellt sich heraus, dass die resultierenden Lernraten – bis zu einem gewissen beliebig kleinen zusätzlichen Term im Exponenten – den optimalen Raten für u.i.v. Prozesse entsprechen.

Danksagung

Mein besonderer Dank gilt meinem Doktorvater Prof. Dr. Ingo Steinwart für seine wissenschaftliche und moralische Unterstützung, seine Ermutigungen sowie für die Möglichkeit, auf diesem interessanten Gebiet zu promovieren. Außerdem danke ich den Mitberichtern, Prof. Dr. Andreas Christmann und Prof. Dr. Gilles Blanchard, für die von Ihnen aufgebrachte Zeit.

Schönen Dank möchte ich meinem Diplomvater Prof. Dr. Harro Walk für die Empfehlung zur Promotion aussprechen. Außerdem möchte ich mich bei meinen alten und neuen Kollegen am ISA für die schöne Zeit bedanken.

Von ganzem Herzen danke ich meiner Frau Lina und unserem kleinen Sohn Chengrong, die mich immer wieder neue Kraft gaben. Nicht zuletzt danke ich meinen Schwiegereltern und Eltern für ihren uneingeschränkten Beistand und den Rückzugsort, den sie mir jederzeit boten.

Contents

1. Introduction

Statistical Learning Theory (SLT) is a mathematical framework that deals with how machines or mechanisms predict results through a process of learning. Learning, in this case, is thought to be "an alteration of behavior as a result of individual experience. When an organism can perceive and change its behavior, it is said to learn." (Encyclopedia Britannica, Vol. 7, 2007.)

The goal of supervised learning is to find a decision function $f_D : X \to Y$ between a sampled set of input variables $x \in X$ and a predicted series of output values $y \in Y$ from a training data $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ of observations drawn from an unknown distribution P. Every point in the process of training will see a piece of input data matched to a piece of output data, so that the resulting learned function will be able to predict an output from any given future input.

For example, consider a simple classification algorithm which should differentiate two animals, namely foxes and cats, based on certain behaviourial characteristics. This is a classic problem of classification, where the object belongs to a defined finite set of labels y, which are usually denoted by the values -1 and 1. In this case, we aim to find a binary classification algorithm that takes the data D as input and outputs a functional relationship $f_D : X \to \{-1, 1\}$.

In other supervised learning scenarios the output Y can take on a range of continuous values such as the real numbers $\mathbb{R}$. These are often called problems of regression. As an example, in salary prediction the output values are non-discrete in nature. Here, the regression is focused on finding a functional relationship $f_D : X \to Y$ between these variables to enable an accurate prediction of an outcome with any given input values.

To evaluate the quality of a decision function f_D, we often use a loss function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ to measure the difference between estimated and true values. For example, we usually take the binary classification loss and the least square loss for the classification and regression problems, respectively. Given a loss L, the goal of supervised learning is to find an estimator $f_D : X \to \mathbb{R}$ such that its *L-risk*

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

is as small as possible. In other words, the L-risk of f_D ought to be close to the *Bayes risk* that is defined to be the minimal L-risk

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

Now, a learning method, or learning algorithm $\mathcal{L}$, that assigns every data set D to a function f_D, is called *consistent*, if

$$\mathcal{R}_{L,P}(f_D) \xrightarrow{n \to \infty} \mathcal{R}^*_{L,P} \tag{1.1}$$

with probability 1. Moreover, $\mathcal{L}$ is said to be *universally consistent*, if (1.1) holds for all P on $X \times Y$ with, e.g., $\mathcal{R}^*_{L,P} < \infty$.

To describe the speed of the convergence in (1.1), let $c_P > 0$ be a constant and $(\varepsilon_n) \subset (0,1]$ be a decreasing sequence converging to 0. Then $\mathcal{L}$ *learns with rate* (ε_n), if, for all $\tau \in (0,1]$, there exists a constant $c_\tau \in [1,\infty)$ only depending on τ such that, for all $n \geq 1$ and all $\tau \in (0,1]$, the inequality

$$\mathcal{R}_{L,P}(f_D) \leq \mathcal{R}^*_{L,P} + c_P c_\tau \varepsilon_n$$

holds with probability not less that $1-\tau$.

In the literature, consistency and learning rates have already been investigated in a variety of scenarios, see e.g. [39, 49, 32, 98]. In essentially all cases concentration inequalities such as

- **Hoeffding's inequality** [52], which states that for independent random variables $Z_1, \ldots, Z_n : \Omega \to [-B,B]$ with some $B > 0$,

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}_P Z_i) \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2B^2}\right) \tag{1.2}$$

  holds for all $\varepsilon > 0$,

- **Bernstein's inequality** [13], namely that for independent random variables Z_1, $\ldots$, $Z_n : \Omega \to [-B,B]$ with some $B > 0$,

$$P\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(\sigma^2 + \varepsilon B/3)}\right) \tag{1.3}$$

  holds for all $\varepsilon > 0$, providing $\mathbb{E}_P Z_i = 0$ and $\mathbb{E}_P Z_i^2 \leq \sigma^2$ with $\sigma > 0$, for all $i = 1, \ldots, n$,

McDiarmid's inequality [68], and Talagrand's inequality [109, 19] play an important role. Indeed, the analysis of various methods from non-parametric statistics and machine learning crucially depend on these inequalities, see e.g. [39, 40, 49, 98]. For example, if the training samples come from an i.i.d. process, it was shown that an important class of learning methods called support vector machines (SVMs), which will be introduced in detail in Section 2.3, enjoy both universal consistency, see e.g. [93, 125, 94, 28], and good learning rates, see e.g. [25, 104, 15, 62, 105]. 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.

Notice that all concentration inequalities mentioned above assume the data to be generated in an i.i.d. fashion. In fact, this i.i.d. scenario is the predominantly considered scenario in the literature, see e.g. [39, 49, 32, 98] and the references therein. However, in practice, this i.i.d. assumption may be violated due to the nature of the data. Typical examples for this phenomenon are applications from financial predictions, signal processing, system observation and diagnosis, and speech or text recognition, where the observations come from a (suitably pre-processed) time series. Therefore, to understand the behavior of learning methods in such situations, the independence assumption must be weakened, so that various types of stochastic processes including Markov chains and many classical time series models are covered.

A set of natural and widely accepted notions for modelling weak dependencies are classical mixing concepts such as α-, β-, and ϕ-mixing, see e.g. [21, 22, 23], since on the one hand, they quantify the dependence structure in a conceptionally simple way, which is accessible to various types of analysis, while on the other hand they include many of the classical time series models.

Considerable effort has been made to establish concentration inequalities for these mixing processes. For example, [18, 73, 71] established a Bernstein-type inequality for the α-mixing processes, while [123, 69, 74, 75, 76, 90] used the so-called blocking technique to study other concentration properties for β- and ϕ-mixing processes. All these results have been used to analyze particular learning algorithms. For example, [48] studied a classification algorithm based on a regularization scheme in a reproducing kernel Hilbert space and a generic convex loss function for α- and β-mixing processes. For the regression problem, [102] established consistency of SVMs learning from α-mixing processes, while [119, 107, 106, 82] analyzed least squares support vector machines (LS-SVMs) with α-mixing inputs, and [127] established generalization bounds for empirical risk minimization (ERM) when the sampling sequence satisfies an α-mixing condition. Moreover, the Bernstein-type inequality established in [18] was used in [124] to obtain convergence rates for sieve estimates from α-mixing strictly stationary processes in the special case of neural networks. More recently, by applying the Bernstein-type inequality in [73], [99] obtained a general oracle inequality for generic regularized learning algorithms from α-mixing observations, [128] analyzed the generalized performance of empirical risk minimization algorithms with α-mixing samples and [126, 45, 29] considered the regularized learning algorithm associated with the least-square loss and α-mixing observations. Moreover, by employing the Bernstein-type inequality in [71], [11] 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. For the smaller class of β-mixing processes, PAC-learning questions have been investigated in [113], while [64] established consistency of regularized boosting algorithms learning from β-mixing processes. For the even smaller class of ϕ-mixing processes, in the particular case of the least square loss, [2] obtained the optimal learning rate for ϕ-mixing processes by applying the Bernstein-type inequality established in [90].

In this work, one of the main goals is to derive an oracle inequality for a generic class of learning algorithms including ERM and SVMs, which is based on a generic form of Bernstein's inequality. On the technical side, the new oracle inequality is achieved by a refinement of the analysis of [99]. To be more precise, the analysis in [99] partially ignored localization with respect to the regularization term, which we now address by a carefully arranged peeling approach inspired by [98]. As a result, the stochastic error term of our new oracle inequality is always smaller than that of [99]. As far as we know, the best learning rates for LS-SVMs from exponentially α-mixing processes are those derived in [119, 107, 106, 45]. When applied to LS-SVMs from exponentially α-mixing processes, it turns out that our oracle inequality leads to a polynomial learning rate with exponent

$$-\alpha \min\left\{\beta, \frac{\beta}{\beta + p\beta + p}\right\}.$$

This is obviously better than the exponent $-\alpha \min\{\beta, \frac{\beta}{\beta+2p\beta+p}\}$ established by [99] and the exponent $-\frac{\alpha\beta}{2p+1}$ established by [119] and [45]. For sufficiently smooth kernels, our exponent is also better than the exponent $-\frac{2\alpha\beta}{\beta+3}$ derived in [106] as well as the improved exponent $-\frac{3\alpha\beta}{2\beta+4}$ in [107]. We refer to Examples 3.15 and 3.16 for more precise comparisons between all these learning rates.

However, there exist many dynamical systems such as uniformly expanding maps given in [34, p. 41] that are not α-mixing. To deal with such non-mixing processes Rio [84] introduced so-called $\tilde{\phi}$-mixing coefficients, which extend the classical ϕ-mixing coefficients. Unfortunately, the $\tilde{\phi}$-mixing class is still not large enough to cover many commonly considered dynamical systems including uni-modal maps [67, Section 4.2]. To include such dynamical systems, [67] proposed the $\mathcal{C}$-mixing coefficients, which further generalize $\tilde{\phi}$-mixing coefficients.

In contrast to the classical mixing case, so far there are only a few such concentration inequalities known for $\mathcal{C}$-mixing processes. Among these inequalities, it is worth mentioning that for dynamical systems with exponentially decreasing, *modified* $\tilde{\phi}$-coefficients, [116] derived a Bernstein-type inequality that turns out to be the same as the one for i.i.d. processes modulo some logarithmic factor. However, this modification of the $\tilde{\phi}$-coefficients seems to be significantly stronger than Rio's original $\tilde{\phi}$-mixing coefficients, so it remains unclear when the Bernstein-type inequality in [116] is actually applicable. For this reason, the second main goal of this work is to establish a Bernstein-type inequality for stationary geometrically (time-reversed) $\mathcal{C}$-mixing processes $\mathcal{Z} := (Z_n)_{n\geq 0}$ with associated semi-norm $\|\cdot\|$. More precisely, let $A > 0$, $B > 0$, $\sigma \geq 0$, and $h : Z \to [-B, B]$ be such that $\mathbb{E}_P h = 0$, $\|h\| \leq A$, and $\mathbb{E}_P h^2 \leq \sigma^2$. Then we will show that for all $\varepsilon > 0$ and all $n \geq n_0$ with n_0 being some constant depending on A and B,

$$\frac{1}{n}\sum_{i=1}^{n} h(Z_i(\omega)) \geq \sqrt{\frac{8(\log n)^{\frac{2}{\gamma}}\sigma^2\tau}{n}} + \frac{8(\log n)^{\frac{2}{\gamma}}B\tau}{3n} \tag{1.4}$$

holds with probability not less than $1-2e^{-\tau}$. Notice that apart from the constant n_0, the additional logarithmic factor $4(\log n)^{\frac{2}{\gamma}}$, and the constant 2 in front of $e^{-\tau}$, (1.4) coincides with (1.3). Moreover, in this case, our oracle inequality is applicable, since (1.4) is also of the generic form.

This thesis is organized as follows: In Chapter 2, we first present the basic notions of statistical learning. Inter alia, we recall the formal concepts such as loss functions and risks, learning methods, kernels and reproducing kernel Hilbert spaces (RKHSs) as well as their properties. Then, based on a generic form of Bernstein's inequality, we derive an oracle inequality for a generic class of learning algorithms including ERM and SVMs.

In Chapter 3, applying the oracle inequality to learning from α-mixing processes, it turns out that, for ERM, our results cover those in the i.i.d. case. In this sense, our rates for LS-SVMs with Gaussian kernels match essentially the optimal learning rates, while for LS-SVMs with given generic kernel, we only obtain rates that are close to the optimal ones. Moreover, if the α-mixing coefficients decay fast enough, the resulting learning rates for SVMs for both least squares and quantile regression with Gaussian kernels will match the optimal rates for i.i.d. processes up to some arbitrarily small extra term in the exponent.

In Chapter 4, we first prove a Bernstein-type inequality for geometrically $\mathcal{C}$-mixing processes that turns to be of the generic form. Hence, our oracle inequality can also be applied to these processes. Here it turns out that for both least squares and quantile regression using Gaussian kernels, up to several constants associated with some seminorm, we recover the (essentially) optimal rates recently found for the i.i.d. case, see [43], where the data is generated by a geometrically $\mathcal{C}$-mixing process. Moreover, we establish an oracle inequality for the problem of forecasting an unknown dynamical system. This oracle will make it possible to extend the purely asymptotic analysis in [97] to learning rates. Then, in the experiments, we compare the LS-SVMs to some other learning methods such as the polynomial regression, local polynomial regression and multilayer perceptron neural network for learning from some dynamical systems.

In the last chapter, we give a brief summary of this thesis and discuss some open questions and suggestions for refinement.

Finally, we would like to point out that most of the results presented in the thesis have been published or submitted in advance. More precisely, the core techniques of Theorem 2.23 establishing oracle inequality for generic regularized learning algorithms and general exponentially α-mixing observations have been published in [50] in a simplified form. Moreover, the new Bernstein-type inequality for exponentially $\mathcal{C}$-mixing observations with an application to learning can be found in [51].

2. Statistical Learning Theory

In this chapter, we first recall some essential notations that will be used in this thesis. In the second section, we introduce the notion of loss functions and their risks which basically build the basis for the statistical learning. After giving the most important definitions, we discuss properties related to them. Then, Section 2.4 deals with RKHSs and the related reproducing kernels and Section 2.3 is devoted to describe the commonly used learning methods such as ERM and SVMs. In Section 2.5, we recall some concepts that describe the capacity of hypothesis function set. In the last section, based on a generic form of Bernstein's inequality, we establish an oracle inequality for a generic class of learning algorithms including ERM and SVMs by additionally using a rather involved version of the so-called peeling method introduced in [111].

2.1 Preliminaries

Let $(\Omega, \mathcal{A}, \mu)$ be a measurable space and $\mathcal{L}_0(\Omega)$ be the set of all real-valued measurable functions on Ω. For $1 \leq p \leq \infty$, we say that two functions $f, g \in \mathcal{L}_0(\Omega)$ are equivalent if they are equal μ-a.e. The *Lebesgue space* $L_p(\mu)$ consists of equivalence classes of measurable functions $f : \Omega \to \mathbb{R}$ such that the L_p-norm $\|f\|_p$ is finite, where

$$\|f\|_p := \left(\int_\Omega |f|^p \, d\mu \right)^{1/p} \qquad \text{for } 1 \leq p < \infty,$$

and

$$\|f\|_\infty := \operatorname*{ess\,sup}_{x \in \Omega} |f(x)| := \inf\{a \in \mathbb{R} : \mu\{x \in \Omega : f(x) > a\} = 0\}.$$

Moreover, if $\mathcal{A}' \subset \mathcal{A}$ is a sub-σ-algebra, then $L_p(\mathcal{A}', \mu)$ denotes the space of all $\mathcal{A}'$-measurable functions $f \in L_p(\mu)$. It is well-known that, the space $L_p(\mu)$ equipped with the L_p-norm, forms a Banach space.

For a Banach space E, we denote its closed unit ball by B_E. In particular, for the d-dimensional Euclidean space, we write $B_{\ell_2^d}$. For $t \in \mathbb{R}$, $\lfloor t \rfloor$ denotes the largest integer n satisfying $n \leq t$, and similarly, $\lceil t \rceil$ is the smallest integer n satisfying $n \geq t$.

For $1 \leq p \leq \infty$, the *Sobolev space* of order $m \in \mathbb{N}_0$ is defined by

$$W_p^m(\mu) := \left\{ f \in L_p(\mu) : \partial^{(\alpha)} f \in L_p(\mu) \text{ exists for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \leq m \right\}, \qquad (2.1)$$

where $\partial^{(\alpha)}$ is the α-th weak derivative for a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $|\alpha| = \sum_{i=1}^d \alpha_i$. The Sobolev norm is given by

$$\|f\|_{W_p^m(\mu)}^p := \sum_{|\alpha| \leq m} \left\| \partial^{(\alpha)} f \right\|_{L_p(\mu)}^p .$$

If $\Omega \subset \mathbb{R}^d$ and μ is the Lebesgue measure, we write $W_p^m(\Omega) := W_p^m(\mu)$.

Clearly, Sobolev spaces are subspaces of $L_p(\mu)$. To introduce another subspace of $L_p(\mu)$, we first need to recall the modulus of smoothness. To this end, let Ω be a subset of $\mathbb{R}^d$ with non-empty interior, $f \in L_p(\mu)$ for some $p \in (0, \infty]$, and $h = (h_1, \dots, h_d) \in \mathbb{R}^d$. For $s \in \mathbb{N}$, the s-th modulus of smoothness of f is defined by

$$\omega_{r,L_p(\mu)}(f,t) = \sup_{\|h\|_2 \le t} \|\triangle_h^r(f, \,\cdot\,)\|_{L_p(\mu)}\,, \qquad t \ge 0\,,$$

where $\|\cdot\|_2$ denotes the Euclidean norm and the r-th difference $\triangle_h^s(f, \,\cdot\,)$ of f is given by

$$\triangle_h^r(f,x) = \begin{cases} \sum_{j=0}^r \binom{r}{j} (-1)^{r-j} f(x+jh)\,, & \text{if } x, x+h, \dots, x+rh \in \Omega\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

For $1 \le p, q \le \infty$, $\alpha > 0$, $s := \lfloor \alpha \rfloor + 1$, the *Besov space* $B_{p,q}^\alpha(\mu)$ is defined by

$$B_{p,q}^\alpha(\mu) := \left\{ f \in L_p(\mu) : |f|_{B_{p,q}^\alpha(\mu)} < \infty \right\}\,, \tag{2.2}$$

where the semi-norm $|\cdot|_{B_{p,q}^\alpha(\mu)}$ is given by

$$|f|_{B_{p,q}^\alpha(\mu)} := \left(\int_0^\infty \left(t^{-\alpha} \omega_{s,L_p(\mu)}(f,t)\right)^q \frac{dt}{t} \right)^{\frac{1}{q}}\,, \qquad 1 \le q < \infty\,,$$

and

$$|f|_{B_{p,\infty}^\alpha(\mu)} := \sup_{t>0} \left(t^{-\alpha} \omega_{s,L_p(\mu)}(f,t)\right)\,.$$

Note that

$$\|f\|_{B_{p,q}^\alpha(\mu)} := \|f\|_{L_p(\mu)} + |f|_{B_{p,q}^\alpha(\mu)}$$

actually forms a norm of $B_{p,q}^\alpha(\mu)$ for all $q \in [1, \infty]$. Again, if μ is the Lebesgue measure on Ω, we write $B_{p,q}^\alpha(\Omega) := B_{p,q}^\alpha(\mu)$.

In the following X is always a measurable space if not mentioned otherwise and $Y \subset \mathbb{R}$ is always a closed subset. Moreover, metric spaces are always equipped with the Borel σ-algebra, and products of measurable spaces are always equipped with the corresponding product σ-algebra.

2.2 Basic Properties of Losses and Their Risks

As already mentioned in the introduction, the goal of (supervised) statistical learning is to find a function $f : X \to \mathbb{R}$ such that for $(x,y) \in X \times Y$ generated according to the distribution P, the value $f(x)$ is a good prediction of y at x. When we found a function $f : X \to \mathbb{R}$, its quality can be assessed. Now, we introduce some quite well-known concepts of a loss function, see also [98, Definitions 2.1].

Definition 2.1. A function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called a *loss function*, or simply a *loss*, if it is measurable.

We say that a loss $L(x, y, \cdot\,) : \mathbb{R} \to [0, \infty]$ is *convex (or continuous)*, if L is *convex (or continuous)* for all $x \in X$, $y \in Y$.

Given a loss function L and an $f : X \to \mathbb{R}$, we often use the notation $L \circ f$ for the function $(x, y) \mapsto L(x, y, f(x))$. A loss can express the approximation ability of a function f of the response value y of some input value $x \in X$. The smaller the value of $L \circ f$, the better is the prediction of y in x by $f(x)$. Therefore, small values of $L \circ f$ should be considered. Until now, only loss functions for a fixed pair (x, y) have been considered. However, our major goal is to have a small average loss for future unseen observations (x, y). The following definition formalizes the concept of the average quality of the function f, see also [98, Definitions 2.2 and 2.3].

Definition 2.2. Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a probability measure on $X \times Y$. Then, for $f \in \mathcal{L}_0(X)$ the *L-risk* is defined by

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

Moreover, the minimal L-risk

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

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

Note that the above integral over $X \times Y$ always exists because L is non-negative and measurable. Moreover, it is easy to verify that the risk of a convex loss is convex on $\mathcal{L}_0(X)$. However, in general the risk of a continuous loss is not necessarily continuous. In order to ensure this continuity and several others like the Lipschitz continuity, we introduce the following definition, see also [98, Definitions 2.16].

Definition 2.3. A loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called a *Nemitski loss*, if there exist a measurable function $b : X \times Y \to [0, \infty)$ and an increasing function $h : [0, \infty) \to [0, \infty)$ with

$$L(x, y, t) \leq b(x, y) + h(|t|)$$

for all $(x, y, t) \in X \times Y \times \mathbb{R}$. Furthermore, L is called a *Nemitski loss of order* $p \in (0, \infty)$, if there exists a constant $c > 0$ such that

$$L(x, y, t) \leq b(x, y) + c|t|^p$$

for all $(x, y, t) \in X \times Y \times \mathbb{R}$. Besides, for a distribution P on $X \times Y$ with $b \in L_1(P)$ we call L a *P-integrable Nemitski loss.*

Note that for all $f \in L_\infty(P_X)$, a P-integrable Nemitski loss functions L satisfy $\mathcal{R}_{L,P}(f) < \infty$. In particular, we have $\mathcal{R}_{L,P}(0) < \infty$ and $\mathcal{R}^*_{L,P} < \infty$.

Now, we recall the Lipschitz continuity that are satisfied by nearly all commonly used loss functions, see also [98, Definitions 2.18].

Definition 2.4. Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function. We say that L is:

(i) *locally Lipschitz continuous*, if for all $a > 0$ we have

$$|L|_{a,1} := \sup_{\substack{t,t' \in [-a,a] \\ t \neq t'}} \sup_{\substack{x \in X \\ y \in Y}} \frac{L(x,y,t) - L(x,y,t')}{|t - t'|} < \infty. \tag{2.3}$$

(ii) *Lipschitz continuous*, if we have $|L|_1 := \sup_{a>0} |L|_{a,1} < \infty$.

Note that if $Y \subset \mathbb{R}$ is finite and the loss $L : Y \times \mathbb{R} \to [0, \infty)$ is convex, then L is locally Lipschitz continuous by [98, Lemma A.6.5]. Moreover, a locally Lipschitz continuous loss function L is a Nemitski loss, since (2.3) yields

$$L(x,y,t) \leq L(x,y,0) + |L|_{|t|,1}|t|, \qquad (x,y,t) \in X \times Y \times \mathbb{R}.$$

In particular, a locally Lipschitz continuous loss L is a P-integrable Nemitski loss if and only if $\mathcal{R}_{L,P}(0) < \infty$. Moreover, if L is Lipschitz continuous then L is a Nemitski loss of order 1.

The following examples discuss the above mentioned properties of losses which are often used in learning algorithms for classification and regression problems, for more details we refer to [98, Sections 2.3 & 2.4].

Example 2.5. A loss $L : Y \times \mathbb{R} \to [0, \infty)$ is called *margin-based*, if there exists a *representing function* $\varphi : \mathbb{R} \to [0, \infty)$ such that

$$L(y,t) = \varphi(yt), \qquad y \in Y := \{-1, 1\},\ t \in \mathbb{R}.$$

Many commonly used losses in classification algorithms, such as the least squares loss, the (squared) hinge loss and the logistic loss are margin-based, see [98, Examples 2.26-2.29].

Since the representing function φ simplifies the form of a loss function, properties like convexity, continuity or (locally) Lipschitz continuity are easier to check for the representing function instead of the loss function itself. Moreover, convexity of L implies local Lipschitz continuity of L. In addition, L is always a P-integrable Nemitski loss since we have

$$L(y,t) \leq \max\{\varphi(-t), \varphi(t)\}$$

for all $y \in Y$ and all $t \in \mathbb{R}$. Consequently we can easily derive a characterization for L being a P-integrable Nemitski loss of order p. ■

Example 2.6. We say that a loss $L : Y \times \mathbb{R} \to [0, \infty)$ is *distance-based*, if there exists a *representing function* $\psi : \mathbb{R} \to [0, \infty)$ such that $\psi(0) = 0$ and

$$L(y,t) = \psi(y - t), \qquad y \in Y := \mathbb{R},\ t \in \mathbb{R}.$$

Many commonly used losses for regression problems, such as the least squares loss, Huber's insensitive loss, the logistic loss, the ϵ-insensitive loss, or the pinball loss are distance-based, see [98, Examples 2.39-2.42]. Moreover, it is easy to see that L is convex, continuous, or Lipschitz continuous if and only if ψ is. However, in general, the local Lipschitz continuity of ψ does not imply the local Lipschitz continuity of the corresponding distance-based loss function, see again [98, Section 2.4]. ■

Later, we always assume the label set Y to be $[-M, M]$ for some $M > 0$, consequently it is meaningful to present the following concept that enables us to restrict a loss L to $X \times Y \times [-M, M]$, see also [98, Definition 2.22].

Definition 2.7. We say that a loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ can be *clipped* at $M > 0$ if, for all $(x, y, t) \in X \times Y \times \mathbb{R}$, we have

$$L(x, y, \widehat{t}\,) \leq L(x, y, t), \tag{2.4}$$

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

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

With all these preparations we can now summarize assumptions on the loss function L that will be used throughout this thesis.

Assumption 2.8. *The loss function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ can be clipped at some $M > 0$. Moreover, it is both bounded in the sense of $L(x, y, t) \leq 1$ and locally Lipschitz continuous, that is,*

$$|L(x, y, t) - L(x, y, t')| \leq |t - t'|\,, \tag{2.5}$$

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

The following examples illustrate the generality of the made assumptions on L for classification and regression problems:

Example 2.9. Let us first consider the case of binary classification, that is $Y := \{-1, 1\}$. For this learning problem one often uses a convex surrogate for the original discontinuous classification loss $\mathbf{1}_{(-\infty,0]}(y \operatorname{sign} t)$, since the latter may lead to computational infeasible approaches. Typical surrogates L belong to the class of margin-based losses and hence can be clipped, if and only if the representing function φ has a global minimum, see [98, Lemma 2.23]. In particular, the hinge loss, the least squares loss for classification, and the squared hinge loss can be clipped, but the logistic loss for classification and the AdaBoost loss cannot be clipped. On the other hand, [95] established a simple technique, which is similar to inserting a small amount of noise into the labeling process, to construct a clippable modification of an arbitrary convex, margin-based loss. Finally, both the Lipschitz continuity and the boundedness of L can be easily verified for these losses, where for the latter it may be necessary to suitably scale the loss. ■

Example 2.10. Bounded regression is another class of learning problems, where the assumptions made on L are often satisfied. Indeed, if $Y := [-M, M]$ and L is a convex, distance-based loss, then L can be clipped, see again [98, Lemma 2.23]. In particular, the least squares loss and the τ-pinball loss used for quantile regression can be clipped. Again, for both losses, the Lipschitz continuity and the boundedness can be easily enforced by a suitable scaling of the loss. ■

2.3 Learning Methods

Informally, given a data set

$$D_n := \big((X_1, Y_1), \dots, (X_n, Y_n)\big) \in (X \times Y)^n$$

generated from some unknown distribution P on $X \times Y$, the goal of supervised learning is to find a decision function f_D such that $\mathcal{R}_{L,P}(f_D)$ is close to the minimal risk $\mathcal{R}^*_{L,P}$. The following definition will formalize this idea, see e.g. [98, Definition 6.1].

Definition 2.11. Let X be a set and $Y \subset \mathbb{R}$ be a closed subset. A *learning method* $\mathcal{L}$ on $X \times Y$ maps every set $D_n \in (X \times Y)^n$, $n \geq 1$, to a function $f_D : X \to \mathbb{R}$.

Now a natural question about learning is the consistency which basically describes methods producing decision functions close to the optimum with high probability, provided the training set is sufficiently large, see e.g. [39, Definitions 6.1&6.2], [98, Definition 6.4].

Definition 2.12. Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, P be a distribution on $X \times Y$. A learning method $\mathcal{L}$ is *L-risk consistent* for P if for all $\varepsilon > 0$,

$$\mathcal{R}_{L,P}(f_D) \xrightarrow{n\to\infty} \mathcal{R}^*_{L,P} + \varepsilon$$

holds with probability 1. Moreover, if $\mathcal{L}$ is L-risk consistent for all distributions P on $X \times Y$ with $\mathcal{R}^*_{L,P} < \infty$, it is called *universally L-risk consistent.*

In the i.i.d. case many learning methods are known to be universally consistent, see e.g. [39] for classification methods, [49] for regression methods, and [98] for generic SVMs. For consistent methods, it is natural to ask how fast the convergence above, is. See e.g. [98, Lemma 6.5].

Definition 2.13. Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function, P be a distribution on $X \times Y$, and $\mathcal{L}$ be a learning method on $X \times Y$. Moreover, let $c_P > 0$ be a constant and $(\varepsilon_n) \in (0, 1]$ be a decreasing sequence converging to 0. If, for all $\tau \in (0, 1]$, there exists a constant $c_\tau \in [1, \infty)$ only depending on τ such that, for all $n \geq 1$ and all $\tau \in (0, 1]$,

$$\mathcal{R}_{L,P}(f_D) \leq \mathcal{R}^*_{L,P} + c_P c_\tau \varepsilon_n$$

holds with probability P not less that $1 - \tau$, then $\mathcal{L}$ *learns with rate* (ε_n) *and confidence* $(c_\tau)_{\tau\in(0,1]}$.

Unfortunately, by the no-free-lunch theorem of Devroye [38], we know that in most situations uniform convergence rates are impossible, see [39, Theorem 7.2], and hence learning rates require some assumptions on the underlying distribution P. Again, results in this direction can be found in the above-mentioned books [39, 49, 98].

In the non-i.i.d. case, [77] showed that no uniform consistency is possible if one only assumes that the data generating process $\mathcal{Z}$ is stationary and ergodic. On the other hand, if some further assumptions of the dependence structure of $\mathcal{Z}$ are made, then consistency is possible, see e.g. [102]. The most widely made assumptions in this direction are in terms of so called α-mixing coefficients, which will be introduced in Chapter 3, but these are by no means necessary. Indeed, certain dynamical systems which will be dealed with in

Chapter 4 are not necessarily α-mixing. Fortunately, under some additional assumptions, [97] has established consistency for SVMs. Finally, learning rates are possible for data generated by some mixing process, if one makes additional assumptions on P, see again [102] and the references therein.

In order to introduce our generic learning algorithms, we write $D_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}$, where $\delta_{(X_i,Y_i)}$ denotes the (random) Dirac measure at (X_i, Y_i). In other words, D_n is the empirical measure associated to the data set $D := ((X_1, Y_1), \dots, (X_n, Y_n)) \in (X \times Y)^n$. Finally, the risk of a function $f : X \to \mathbb{R}$ with respect to this measure

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

is called the *empirical L-risk.*

Now we introduce the class of learning methods we are interested in, see also [98, Definition 7.18].

Definition 2.14. 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, $\Upsilon : \mathcal{F} \to [0, \infty)$ with $\Upsilon(0) = 0$. Then, for $\delta \geq 0$, a learning method whose decision functions $f_{D_n,\Upsilon} \in \mathcal{F}$ satisfy

$$\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,D_n}(\widehat{f}_{D_n,\Upsilon}) \leq \inf_{f\in\mathcal{F}} (\Upsilon(f) + \mathcal{R}_{L,D_n}(f)) + \delta \tag{2.6}$$

for all $n \geq 1$ and $D \in (X \times Y)^n$ 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, in (2.6), we consider the clipped function on the left-hand side and the unclipped loss on the right-hand side. Hence, in general, CR-ERMs minimize neither the regularized risk $\Upsilon(\,\cdot\,) + \mathcal{R}_{L,D_n}(\,\cdot\,)$ nor the regularized clipped empirical risk $\Upsilon(\,\cdot\,) + \mathcal{R}_{L,D_n}(\widehat{\cdot})$. Nevertheless, if we have a minimizer of the unclipped regularized risk, then it automatically satisfies (2.6).

In the rest of the section, we briefly introduce two specific CR-ERMs, namely, ERM and SVMs.

2.3.1 Empirical Risk Minimization (ERM)

An important class of learning methods is called *empirical risk minimization (ERM)* whose decision functions f_D satisfy

$$\mathcal{R}_{L,D_n}(f_D) = \inf_{f\in\mathcal{F}} \mathcal{R}_{L,D_n}(f)$$

for all $n \geq 1$ and $D_n \in (X \times Y)^n$ with respect to L and $\mathcal{F}$, see e.g. [39, Section 4.5], [98, Definition 6.16].

Obviously, ERM decision functions satisfy (2.6) for the regularizer $\Upsilon := 0$ and $\delta := 0$. In other words, ERMs are CR-ERMs.

Unfortunately, note that in general such a minimizer f_D does not need to exist. Moreover, it is critical to choose the size of the hypothesis set $\mathcal{F}$, since on the one hand, a too small set $\mathcal{F}$ may cause "underfitting" and on the other hand, a too large set $\mathcal{F}$ may lead to "overfitting".

To avoid these unfavorable phenomena, in the 1990s, V. Vapnik and co-workers [17, 31] have developed a new generation of learning algorithms, support vector machines (SVMs).

2.3.2 Support Vector Machines (SVMs)

Let us recall the kernel-based regularized empirical risk minimiziers, the so-called *support vector machines (SVMs)*, see [98] for details. To this end, let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss function and k be a measurable (reproducing) kernel on X with reproducing kernel Hilbert space (RKHS) H, which are special Hilbert spaces containing functions $f : X \to \mathbb{R}$ and will be introduced in more detail in the Section 2.4. It is well-known, see e.g. [98, Lemma 5.1 and Theorem 5.2], that for all $\lambda > 0$ and all observations D, there exists exactly one element $f_{D_n,\lambda} \in H$ such that

$$f_{D_n,\lambda} = \operatorname*{arg\,min}_{f \in H} \left(\lambda \|f\|_H^2 + \mathcal{R}_{L,D_n}(f) \right) . \tag{2.7}$$

In particular, SVMs using the least-squares loss

$$L(y, t) = (y - t)^2 \tag{2.8}$$

are called *least-squares SVMs (LS-SVMs)*, see [112, 108], while SVMs using the τ-pinball loss

$$L_\tau(y, t) := \psi(y - t) = \begin{cases} -(1 - \tau)(y - t), & \text{if } y - t < 0 \\ \tau(y - t), & \text{if } y - t \geq 0 \end{cases} \tag{2.9}$$

are called *SVMs for quantile regression.*

Note that SVM decision functions (2.7) satisfy (2.6) for the regularizer $\Upsilon := \lambda \| \cdot \|_H^2$ and $\delta := 0$. In other words, SVMs are CR-ERMs. Moreover, Assumption 2.8 implies that

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

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

$$f_{D_n,\lambda} \in \lambda^{-1/2} B_H, \tag{2.10}$$

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

2.4 Kernels and Reproducing Kernel Hilbert Spaces

In this section, we briefly recall some basic properties of kernels and reproducing kernel Hilbert spaces (RKHSs) presented in [98, Chapter 4]. Let us begin with the definition of kernels, see also [98, Definition 4.1].

Definition 2.15. Let X be a non-empty set. Then a function $k : X \times X \to \mathbb{R}$ is called a *kernel* on X, if there exists a Hilbert space H and a map $\Phi : X \to H$ such that, for all $x, x' \in X$, we have

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \ . \tag{2.11}$$

Here, Φ is called a *feature map* and H a *feature space* of k.

Note that the feature space H and the feature map Φ are, in general, not uniquely determined. However, in the following, this problem can be resolved by finding a way of assigning to every kernel a unique feature space and feature map. Namely, we construct a feature space which is in some sense a canonical choice. It is called the reproducing kernel Hilbert space (RKHS) and is defined as follows.

Definition 2.16 (cf. [98, Definition 4.18]). Let $X \neq \emptyset$ and H be a Hilbert function space over X, i.e., a Hilbert space that consists of functions mapping from X into $\mathbb{R}$.

1. A function $k : X \times X \to \mathbb{R}$ is called a *reproducing kernel* of H if $k(\,\cdot\,, x) \in H$ holds for all $x \in X$ and the *reproducing property*

   $$f(x) = \langle f, k(\,\cdot\,, x)\rangle$$

   is satisfied for all $f \in H$ and all $x \in X$.

2. The space H is called a *reproducing kernel Hilbert space (RKHS)* over X if, for all $x \in X$, the Dirac functional $\delta_x : H \to \mathbb{R}$ defined by

   $$\delta_x(f) := f(x)\,, \qquad\qquad f \in H\,,$$

   is continuous.

It is well-known, see e.g. [98, Theorems 4.20 & 4.21], that there exists a one-to-one correspondence between kernels and RKHSs. This means that for every kernel k there exists exactly one RKHS, such that k is a reproducing kernel of H. Conversely, for every RKHS H there exists exactly one reproducing kernel of H which was shown to be indeed a kernel.

To describe the approximation properties of H, we further need the approximation error function

$$A(\lambda) := \inf_{f\in H} \left(\lambda\|f\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\right), \qquad \lambda > 0. \tag{2.12}$$

Moreover, given a distribution P on $X \times Y$, we say that the RKHS H is (L, P)*-rich* if we have

$$\mathcal{R}^*_{L,P,H} := \inf_{f\in H} \mathcal{R}_{L,P}(f) = \mathcal{R}^*_{L,P}, \tag{2.13}$$

i.e. if the Bayes risk can be approximated by functions from H.

Note that if the kernel of H is *universal* in the sense of [92], i.e. X is a compact metric space and H is dense in the space $C(X)$ of continuous functions, the condition (2.13) is satisfied, see e.g. [98, Corollary 5.29]. Moreover, [92] has proved (2.13) under less restrictive assumptions on H and X and established some necessary and sufficient conditions for (L, P)-richness on countable spaces X.

Example 2.17. The *Gaussian RBF kernels* k_σ on X, are defined by

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

for some width $\sigma \in (0, 1]$. We write H_σ for the RKHS of k_σ, which are described in some detail in [101]. [92] has shown that the Gaussian RBF kernels on $\mathbb{R}^d$ are (L, P)-rich for all distributions P on $\mathbb{R}^d \times Y$ and all continuous, P-integrable Nemitski losses L of order $p \in [1, \infty)$. ■

2.5 Covering and Entropy Numbers

In this section, we recall some concepts describing the capacity of hypothesis set $\mathcal{F}$ in Definition 2.14. Assume that we have a hypothesis set $\mathcal{F}$ consisting of bounded measurable functions $f : X \to \mathbb{R}$, which is pre-compact with respect to the supremum norm $\|\cdot\|_\infty$. Since $\mathcal{F}$ can be infinite, we need to recall the following concept, which will enable us to approximate infinite $\mathcal{F}$ by finite subsets, see e.g. [60, 61] and [98, Definition 6.19].

Definition 2.18. Let (T, d) be a metric space and $\varepsilon > 0$. We call $S \subset T$ an ε-net of T if for all $t \in T$ there exists an $s \in S$ with $d(s, t) \leq \varepsilon$. Moreover, the *ε-covering number* of T is defined by

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

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

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

Besides covering numbers, we introduce the following "inverse" concept, which is also frequently used in the literature, see also [98, Definition 6.20].

Definition 2.19. Let (T, d) be a metric space and $i \geq 1$ be an integer. Then the *i-th (dyadic) entropy number* of (T, d) is defined by

$$e_i\,(T, d) := \inf \left\{ \varepsilon > 0 : \exists t_1, \ldots, t_{2^{i-1}} \in T \text{ such that } T \subset \bigcup_{j=1}^{2^{i-1}} B_d(t_j, \varepsilon) \right\},$$

where the convention $\inf \emptyset := \infty$ is used. Moreover, let $S : E \to F$ be a bounded, linear operator between the normed spaces E and F, then $e_i\,(S) := e_i\,(SB_E, \|\cdot\|_F)$.

Indeed, there exits an equivalence between covering and entropy numbers, we refer to [98, Lemma 6.21] for the proof.

Lemma 2.20. *Let (T, d) be a metric space and $a, q > 0$ be constants such that*

$$e_n\,(T, d) \leq a n^{-1/q}, \qquad n \geq 1.$$

Then, for all $\varepsilon > 0$, we have

$$\ln \mathcal{N}(T, d, \varepsilon) \leq \ln(4) \cdot \left(\frac{a}{\varepsilon}\right)^q.$$

In the following, for an RKHS H, we assume that there exist constants $a > 0$ and $p > 0$ such that

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \leq a\varepsilon^{-2p}, \qquad \varepsilon > 0. \tag{2.14}$$

The following example shows that the covering numbers of Gaussian RKHSs are of the form (2.14), for more examples and discussions we refer to [98, Section 6.4].

Example 2.21. Let H_σ be the Gaussian RKHS and P_X be a distribution on X. By [98, Theorem 7.34] we know that, for all $\varepsilon > 0$ and $0 < p < 1$, there exists a constant $c_{\varepsilon,p} \geq 0$ such that

$$e_i\left(\mathrm{id} : H_\sigma \to L_2\left(P_X\right)\right) \leq c_{\varepsilon,p}\sigma^{-\frac{(1-p)(1+\varepsilon)d}{2p}} i^{-\frac{1}{2p}}$$

for all $i \geq 1$. Lemma 2.20 yields then

$$\ln \mathcal{N}(B_{H_\sigma}, \|\cdot\|_\infty, \varepsilon) \leq a_{p,\zeta}\sigma^{-(1-p)(1+\zeta)d}\varepsilon^{-2p}, \qquad \varepsilon > 0, \tag{2.15}$$

for some constants $a_{p,\zeta} > 0$ and $p \in (0,1)$. ■

2.6 An Oracle Inequality for Generic Learning Algorithms

In this section, we present the key result of this thesis, an oracle inequality for a generic class of learning algorithms including ERM and SVMs based on a generic form of a Bernstein inequality.

2.6.1 Bernstein's Inequality of Generic Form

A generic form of Bernstein's inequality for stationary processes can be stated as follows:

Assumption 2.22. *Let $\mathcal{Z} := (Z_i)_{i\geq 1}$ be an $X \times Y$-valued, stationary stochastic process and $P := \mu_{Z_0}$. Furthermore, let $h : X \times Y \to \mathbb{R}$ be a bounded measurable function for which there exist constants $B > 0$ and $\sigma \geq 0$ such that $\mathbb{E}_\mathrm{P} h = 0$, $\mathbb{E}_\mathrm{P} h^2 \leq \sigma^2$, and $\|h\|_\infty \leq B$. Assume that, there exists a constant $\eta \in [0,1]$ such that for all $\varepsilon > 0$ and all $n \geq n_0$ with n_0 independent of ε, we have*

$$\mu\left(\left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^{n} h(Z_i(\omega)) \geq \varepsilon\right\}\right)$$
$$\leq C \exp\left(-\frac{\varepsilon^2 n}{C_\sigma(n)\sigma^2 + C_\eta(n)\sigma^{2\eta} + C_E(n)B^2/n + C_B(n)\varepsilon B}\right), \tag{2.16}$$

where C is a constant independent of n, $C_\eta(n)$ is a constant depending on η, and $C_\sigma(n) \geq 0$, $C_\eta(n) \geq 0$, $C_E(n) \geq 0$, and $C_B(n) \geq 1$ are some constants that may depend on n.

For later use, we need to reformulate (2.16). Setting

$$\tau = \frac{\varepsilon^2 n}{C_\sigma(n)\sigma^2 + C_\eta(n)\sigma^{2\eta} + C_E(n)B^2/n + C_B(n)\varepsilon B},$$

with some simple transformations we obtain

$$\mu\Bigg(\Bigg\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^{n} h(Z_i(\omega)) \geq \sqrt{\frac{\tau C_\sigma(n)\sigma^2}{n}} + \sqrt{\frac{\tau C_\eta(n)\sigma^{2\eta}}{n}}$$
$$+ \frac{\sqrt{\tau C_E(n)}B}{n} + \frac{\tau C_B(n)B}{n}\Bigg\}\Bigg) \leq Ce^{-\tau}. \tag{2.17}$$

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

Clearly, the classical Bernstein's inequality satisfies (2.16) with $n_0 = 1$, $C = 1$, $C_\sigma(n) = 2$, $C_\eta(n) = 0$, $\eta \in [0,1]$, $C_E(n) = 0$, and $C_B(n) = 2/3$. Moreover, in the literature, there are actually many Bernstein-type inequalities for non-i.i.d. processes of the generic form (2.17), we refer to Section 3.3 and Section 4.2 for some examples.

2.6.2 An Oracle Inequality

To present the following oracle inequality for learning from stationary stochastic processes for which the generic Bernstein-type inequality (2.16) holds, we need to introduce a few more notations. Let $\mathcal{F}$ be a hypothesis set in the sense of Definition 2.14. For

$$r^* := \inf_{f \in \mathcal{F}} \Upsilon(f) + \mathcal{R}_{L,P}(\widehat{f}\,) - \mathcal{R}^*_{L,P} \tag{2.18}$$

and $r > r^*$, we write

$$\mathcal{F}_r := \left\{ f \in \mathcal{F} : \Upsilon(f) + \mathcal{R}_{L,P}(\widehat{f}\,) - \mathcal{R}^*_{L,P} \leq r \right\}. \tag{2.19}$$

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

$$\ln \mathcal{N}(\mathcal{F}_r, \|\cdot\|_\infty, \varepsilon) \leq \varphi(\varepsilon) r^p \tag{2.20}$$

for all $\varepsilon > 0$, $r > 0$ and a suitable constant $p \in (0,1]$. Note that there are actually many hypothesis sets satisfying Assumption (2.20), see Section 3.4 for some examples.

Theorem 2.23. *Suppose that Assumption 2.22 holds with the constants n_0, $\eta \in [0,1]$, C, $C_\sigma(n) \geq 0$, $C_\eta(n) \geq 0$, $C_E(n) \geq 0$, and $C_B(n) \geq 1$. Furthermore, let L be a loss satisfying Assumption 2.8. Assume that there exists a Bayes decision function $f^*_{L,P}$ and constants $\vartheta \in [0,1]$ and $V \geq 1$ such that*

$$\mathbb{E}_P(L \circ \widehat{f} - L \circ f^*_{L,P})^2 \leq V \cdot \left(\mathbb{E}_P(L \circ \widehat{f} - L \circ f^*_{L,P}) \right)^\vartheta, \qquad f \in \mathcal{F}, \tag{2.21}$$

where $\mathcal{F}$ is a hypothesis set with $0 \in \mathcal{F}$. We define r^* and $\mathcal{F}_r$ by (2.18) and (2.19), respectively and assume that (2.20) is satisfied. Finally let $\Upsilon : \mathcal{F} \to [0, \infty)$ be a regularizer with $\Upsilon(0) = 0$, $f_0 \in \mathcal{F}$ be a fixed function, and $B_0 \geq 1$ be a constant such that $\|L \circ f_0\|_\infty \leq B_0$. Then, for all fixed $\varepsilon > 0$, $\delta \geq 0$, $\tau \geq 1$, $n \geq n_0$, and $r \in (0,1]$ satisfying

$$r \geq \max \left\{ \left(\frac{C_V(n)(\tau + \varphi(\varepsilon/2) 2^p r^p)}{n} \right)^{\frac{1}{2-\vartheta\eta}}, \left(\frac{\tau C_\eta(n) B_0}{n} \right)^{\frac{1}{2-\eta}}, \frac{C_\Sigma(n) B_0 \tau C_B(n)}{n}, r^* \right\}, \tag{2.22}$$

every learning method defined by (2.6) satisfies with probability μ not less than $1 - 8Ce^{-\tau}$:

$$\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P} < 2\Upsilon(f_0) + 4\mathcal{R}_{L,P}(f_0) - 4\mathcal{R}^*_{L,P} + 9r + 5\varepsilon + 2\delta. \tag{2.23}$$

Here the constants $C_V(n)$ and $C_\Sigma(n)$ are defined by

$$C_V(n) := 64(4(C_\sigma(n) + C_\eta(n))V + (C_E(n) + C_B(n))), \tag{2.24}$$

$$C_\Sigma(n) := 16(C_\sigma(n) + \sqrt{C_E(n)} + 1). \tag{2.25}$$

Remark 2.24. Before we illustrate this theorem in the next section with the help of a few examples, let us briefly discuss the variance bound (2.21). For example, if $Y = [-M, M]$ and L is the least squares loss, then it is well-known that (2.21) is satisfied for $V := 16M^2$ and $\vartheta = 1$, see e.g. [98, Example 7.3]. Moreover, under some assumptions on the distribution P, [100] established a variance bound of the form (2.21) for the so-called pinball loss used for quantile regression. In addition, for the hinge loss, (2.21) is satisfied for $\vartheta := q/(q+1)$, if Tsybakov's noise assumption [110, Proposition 1] holds for q, see [98, Theorem 8.24]. Finally, based on [16], [95] established a variance bound with $\vartheta = 1$ for the earlier mentioned clippable modifications of strictly convex, twice continuously differentiable margin-based loss functions.

Remark 2.25. One might wonder, why the constant B_0 is necessary in Theorem 2.23, since apparently it only adds further complexity. However, a closer look reveals that the assumed boundedness of L only guarantees $\|L \circ \widehat{f}\|_\infty \le 1$, while B_0 bounds 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 believe that in general B_0 is necessary. We refer to Examples 3.15, 3.16 and 3.17 for situations, where B_0 is significantly larger than 1.

Remark 2.26. Modulo the parameter η, our oracle inequality match those in the i.i.d. case, if one replaces the number of observations with the "effective number of observations"

$$n_{\text{eff}} := \min\left\{\frac{n}{C_V(n)}, \frac{n}{C_\Sigma(n)C_B(n)}\right\}. \tag{2.26}$$

For the proof of Theorem 2.23, we need the so-called peeling method (see for example [111, Chapter 5.3]). To this end, let $0 < r^* < R < \infty$ and $\Gamma : \mathcal{G} \to [r^*, R)$ be some function on a hypothesis function set $\mathcal{G}$, $r^* = m_0 < m_1 < \cdots < m_{K+1} < m_{K+2} = R$ be a strictly increasing sequence. Then $\mathcal{G}$ can be "peeled off" into

$$\mathcal{G} = \bigcup_{k=1}^{K+2} \mathcal{G}_k, \tag{2.27}$$

where $\mathcal{G}_k$ are the disjoint "spheres"

$$\mathcal{G}_k = \{g \in \mathcal{G} : m_{k-1} \le \Gamma(g) < m_k\}, \qquad k = 1, \dots, K+2.$$

Now we can formulate the peeling as following:

Theorem 2.27. *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $\mathcal{Z} = (\mathcal{Z}_g)_{g\in\mathcal{G}}$ be a stochastic process indexed by $\mathcal{G}$, we have for all $\epsilon \ge 0$ and $r > 0$,*

$$\mu\left(\sup_{g\in\mathcal{G}} \frac{|\mathcal{Z}_g|}{\Gamma(g)+r} > \epsilon\right) \le \sum_{k=1}^{K+2} \mu\left(\sup_{g\in\mathcal{G}, \Gamma(g)<m_k} |\mathcal{Z}_g| > (m_{k-1}+r)\epsilon\right). \tag{2.28}$$

Proof. With the peeling (2.27) we obtain

$$\left\{\sup_{g\in\mathcal{G}} \frac{|\mathcal{Z}_g|}{\Gamma(g)+r} > \epsilon\right\} = \bigcup_{k=1}^{K+2} \left\{\sup_{g\in\mathcal{G}_k} \frac{|\mathcal{Z}_g|}{\Gamma(g)+r} > \epsilon\right\}$$

for all $\epsilon \geq 0$ and $r > 0$. The subadditivity of the measure μ then implies

$$\mu \left(\sup_{g \in \mathcal{G}} \frac{|\mathcal{Z}_g|}{\Gamma(g) + r} > \epsilon \right) \leq \sum_{k=1}^{K+2} \mu \left(\sup_{g \in \mathcal{G}_k} \frac{|\mathcal{Z}_g|}{\Gamma(g) + r} > \epsilon \right)$$
$$\leq \sum_{k=1}^{K+2} \mu \left(\sup_{g \in \mathcal{G}, \Gamma(g) < m_k} |\mathcal{Z}_g| > \epsilon(m_{k-1} + r) \right) . \qquad \square$$

In addition, we will need the following simple and well-known lemma (see e.g. [98, Lemma 7.1]):

Lemma 2.28. *For $q \in (1, \infty)$, define $q' \in (1, \infty)$ by $1/q + 1/q' = 1$. Then, for all $a, b \geq 0$, we have $(qa)^{2/q}(q'b)^{2/q'} \leq (a+b)^2$ and $ab \leq a^q/q + b^{q'}/q'$.*

Since the proof of Theorem 2.23 is rather complicated, we first describe its main steps briefly: First we decompose the regularized excess risk into an approximation error term and two stochstic error terms. The approximation error and the first stochastic error term can be estimated by standard techniques. Similarly, the first step in the estimation of the second error term is a rather standard quotient approach, see e.g. [98, Theorem 7.20], which allows for localization with respect to both the variance and the regularization. Due to the absence of tools from empirical process theory, however, the remaining estimation steps become more involved. To be more precise, we split the "unit ball" of the hypothesis space $\mathcal{F}$ into disjoint "spheres". For each sphere, we then use localized covering numbers and Bernstein's inequality from Assumption 2.22, and the resulting estimates are then combined using the peeling method. This yields a quasi geometric series with rate smaller than 1, if the radius of the innermost ball is sufficiently large. As a result, the estimated error probability on the whole "unit ball" nearly equals the estimated error probability of the innermost "ball", which unsurprisingly leads to a significant improvement compared to [99].

Proof (of Theorem 2.23). Main Decomposition. For $f : X \to \mathbb{R}$ we define $h_f := L \circ f - L \circ f^*_{L,P}$. By the definition of $f_{D_n,\Upsilon}$, we then have

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

and consequently we obtain

$$\begin{aligned} &\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P} \\ &= \Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} \\ &\leq \Upsilon(f_0) + \mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_{D_n} h_{\widehat{f}_{D_n,\Upsilon}} + \mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} + \delta \\ &= (\Upsilon(f_0) + \mathbb{E}_P h_{f_0}) + (\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_P h_{f_0}) + (\mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} - \mathbb{E}_{D_n} h_{\widehat{f}_{D_n,\Upsilon}}) + \delta. \end{aligned} \tag{2.29}$$

Estimating the First Stochastic Term. Let us first bound the term $\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_P h_{f_0}$. To this end, we further split this difference into

$$\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_P h_{f_0} = \Big(\mathbb{E}_{D_n}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) \Big) + (\mathbb{E}_{D_n} h_{\widehat{f}_0} - \mathbb{E}_P h_{\widehat{f}_0}). \tag{2.30}$$

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

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

Inequality (2.17) applied to $h := (h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})$ thus shows that

$$\begin{aligned}&\mathbb{E}_{D_n}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) \\ &\leq \sqrt{\frac{\tau C_\sigma(n) B_0 \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})}{n}} + \sqrt{\frac{\tau C_\eta(n) \Big(B_0 \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) \Big)^\eta}{n}} \\ &\quad + \frac{\sqrt{C_E(n)\tau} B_0}{n} + \frac{\tau C_B(n) B_0}{n}\end{aligned}$$

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

$$\sqrt{n^{-1}\tau C_\sigma(n) B_0 \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})} \leq \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})/2 + n^{-1} C_\sigma(n) B_0 \tau/2.$$

In addition, since $B_0 \geq 1$ and $\eta \in [0, 1]$, the second inequality in Lemma 2.28 implies for $q := \frac{2}{2-\eta}$, $q' := \frac{2}{\eta}$, $a := (\frac{1}{\eta})^{-\frac{\eta}{2}} (n^{-1}\tau C_\eta(n) B_0)^{1/2}$, and $b := (\frac{1}{\eta}\mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}))^{\frac{\eta}{2}}$, that

$$\begin{aligned}&\sqrt{\frac{\tau C_\eta(n) \Big(B_0 \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) \Big)^\eta}{n}} \\ &= \big(n^{-1}\tau C_\eta(n) B_0\big)^{1/2} \Big(\mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) \Big)^{\frac{\eta}{2}} \\ &\leq \left(\frac{2}{2-\eta}\right)^{-1} \left(\frac{1}{\eta}\right)^{-\frac{\eta}{2-\eta}} (n^{-1}\tau C_\eta(n) B_0)^{\frac{1}{2-\eta}} + \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})/2 \\ &\leq \left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}} + \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0})/2.\end{aligned}$$

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

$$\begin{aligned}&\mathbb{E}_{D_n}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) \\ &\leq \mathbb{E}_P(h_{f_0} - h_{\widehat{f}_0}) + \frac{C_\sigma(n) B_0 \tau}{2n} + \left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}} + \frac{\sqrt{C_E(n)\tau} B_0}{n} + \frac{B_0 \tau C_B(n)}{n}.\end{aligned} \tag{2.31}$$

In order to bound the remaining term in (2.30), that is $\mathbb{E}_{D_n} h_{\widehat{f}_0} - \mathbb{E}_P h_{\widehat{f}_0}$, we first observe that (2.5) implies $\|h_{\widehat{f}_0}\|_\infty \leq 1$, and hence we have $\|h_{\widehat{f}_0} - \mathbb{E}_P h_{\widehat{f}_0}\|_\infty \leq 2$. Moreover, (2.21) yields

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

Again, inequality (2.17) applied to $h := h_{\widehat{f}_0} - \mathbb{E}_P h_{\widehat{f}_0}$ thus shows that

$$\mathbb{E}_{D_n}(h_{\widehat{f}_0} - \mathbb{E}_P h_{\widehat{f}_0}) \leq \sqrt{\frac{\tau C_\sigma(n) V(\mathbb{E}_P h_{\widehat{f}_0})^\vartheta}{n}} + \sqrt{\frac{\tau C_\eta(n) \Big(V(\mathbb{E}_P h_{\widehat{f}_0})^\vartheta \Big)^\eta}{n}}$$

$$+\frac{2\sqrt{C_E(n)\tau}}{n}+\frac{2\tau C_B(n)}{n}$$

holds with probability μ not less than $1-Ce^{-\tau}$. If $\vartheta\in(0,1]$, the second inequality in Lemma 2.28 implies for $q:=\frac{2}{2-\vartheta}$, $q':=\frac{2}{\vartheta}$, $a:=(n^{-1}C_\sigma(n)\vartheta^\vartheta V\tau)^{1/2}$, and $b:=(\vartheta^{-1}\mathbb{E}_P h_{\widehat{f_0}})^{\vartheta/2}$, that

$$\begin{aligned}\sqrt{\frac{C_\sigma(n)V\tau(\mathbb{E}_P h_{\widehat{f_0}})^\vartheta}{n}}&\le\left(1-\frac{\vartheta}{2}\right)\left(\frac{C_\sigma(n)\vartheta^\vartheta V\tau}{n}\right)^{\frac{1}{2-\vartheta}}+\mathbb{E}_P h_{\widehat{f_0}}/2\\&\le\left(\frac{C_\sigma(n)V\tau}{n}\right)^{\frac{1}{2-\vartheta}}+\mathbb{E}_P h_{\widehat{f_0}}/2.\end{aligned}$$

In addition, since $V\ge1$ and $\eta\in[0,1]$, the second inequality in Lemma 2.28 implies for $q:=\frac{2}{2-\vartheta\eta}$, $q':=\frac{2}{\vartheta\eta}$, $a:=(\frac{1}{\vartheta\eta})^{-\frac{\vartheta\eta}{2}}(n^{-1}\tau C_\eta(n)V)^{1/2}$, and $b:=(\frac{1}{\vartheta\eta}\mathbb{E}_P h_{\widehat{f_0}})^{\frac{\vartheta\eta}{2}}$, that

$$\begin{aligned}&\sqrt{\frac{\tau C_\eta(n)\left(V(\mathbb{E}_P h_{\widehat{f_0}})^\vartheta\right)^\eta}{n}}\\&\le\left(n^{-1}\tau C_\eta(n)V\right)^{1/2}\left(\mathbb{E}_P h_{\widehat{f_0}}\right)^{\frac{\vartheta\eta}{2}}\\&\le\left(\frac{2}{2-\vartheta\eta}\right)^{-1}\left(\frac{2}{\vartheta\eta}\right)^{-\frac{\vartheta\eta}{2-\vartheta\eta}}(n^{-1}\tau C_\eta(n)V)^{\frac{1}{2-\vartheta\eta}}+\mathbb{E}_P h_{\widehat{f_0}}/2\\&\le\left(\frac{\tau C_\eta(n)V}{n}\right)^{\frac{1}{2-\vartheta\eta}}+\mathbb{E}_P h_{\widehat{f_0}}/2.\end{aligned}$$

Since $\mathbb{E}_P h_{\widehat{f_0}}\ge0$, these inequalities also holds for $\vartheta=0$, and consequently we have

$$\begin{aligned}&\mathbb{E}_{D_n}h_{\widehat{f_0}}-\mathbb{E}_P h_{\widehat{f_0}}\\&<\mathbb{E}_P h_{\widehat{f_0}}+\left(\frac{C_\sigma(n)V\tau}{n}\right)^{\frac{1}{2-\vartheta}}+\left(\frac{\tau C_\eta(n)V}{n}\right)^{\frac{1}{2-\vartheta\eta}}+\frac{2\sqrt{C_E(n)\tau}}{n}+\frac{2\tau C_B(n)}{n}\end{aligned}\tag{2.32}$$

with probability μ not less than $1-Ce^{-\tau}$. By combining this estimate with (2.31) and (2.30), we now obtain that with probability μ not less than $1-2Ce^{-\tau}$ we have

$$\begin{aligned}&\mathbb{E}_{D_n}h_{f_0}-\mathbb{E}_P h_{f_0}\\&<\mathbb{E}_P h_{f_0}+\left(\frac{C_\sigma(n)V\tau}{n}\right)^{\frac{1}{2-\vartheta}}+\left(\frac{\tau C_\eta(n)V}{n}\right)^{\frac{1}{2-\vartheta\eta}}+\frac{2\sqrt{C_E(n)\tau}}{n}+\frac{2\tau C_B(n)}{n}\\&\quad+\frac{C_\sigma(n)B_0\tau}{2n}+\left(\frac{\tau C_\eta(n)B_0}{n}\right)^{\frac{1}{2-\eta}}+\frac{\sqrt{C_E(n)\tau}B_0}{n}+\frac{B_0\tau C_B(n)}{n}\\&\le\mathbb{E}_P h_{f_0}+\left(\frac{C_\sigma(n)V\tau}{n}\right)^{\frac{1}{2-\vartheta}}+\left(\frac{\tau C_\eta(n)V}{n}\right)^{\frac{1}{2-\vartheta\eta}}\\&\quad+\left(\frac{\tau C_\eta(n)B_0}{n}\right)^{\frac{1}{2-\eta}}+\frac{3(C_\sigma(n)+\sqrt{C_E(n)}+1)B_0\tau C_B(n)}{n},\end{aligned}\tag{2.33}$$

since $B_0, \tau \geq 1$, i.e., we have established a bound on the second term in (2.29).

Estimating the Second Stochastic Term. For the third term in (2.29) let us first consider the case $n < C_V(n)(\tau + \varphi(\varepsilon/2)2^p r^p)$ with $C_V(n)$ defined as in (2.24). Combining (2.33) with (2.29) and using $1 \leq B_0$, $1 \leq V$, $C_\Sigma(n)$ as in (2.25), and $\mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} - \mathbb{E}_{D_n} h_{\widehat{f}_{D_n,\Upsilon}} \leq 2$, then we find

$$\begin{aligned}
&\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P} \\
&\leq \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} + \left(\frac{C_\sigma(n) V \tau}{n}\right)^{\frac{1}{2-\vartheta}} + \left(\frac{\tau C_\eta(n) V}{n}\right)^{\frac{1}{2-\vartheta\eta}} \\
&\quad + \left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}} + \frac{3(C_\sigma(n) + \sqrt{C_E(n)} + 1) B_0 \tau C_B(n)}{n} \\
&\quad + (\mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} - \mathbb{E}_{D_n} h_{\widehat{f}_{D_n,\Upsilon}}) + \delta \\
&\leq \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} + \left(\frac{C_\sigma(n) V \tau}{n}\right)^{\frac{1}{2-\vartheta}} + \left(\frac{C_V(\tau + \varphi(\varepsilon/2)2^p r^p)}{n}\right)^{\frac{1}{2-\vartheta\eta}} \\
&\quad + \left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}} + \frac{3(C_\sigma(n) + \sqrt{C_E(n)} + 1) B_0 \tau C_B(n)}{n} \\
&\quad + 2\left(\frac{C_V(\tau + \varphi(\varepsilon/2)2^p r^p)}{n}\right)^{\frac{1}{2-\vartheta\eta}} + \delta \\
&\leq \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} + \left(\frac{C_\sigma(n) V \tau}{n}\right)^{\frac{1}{2-\vartheta}} + 3\left(\frac{C_V(\tau + \varphi(\varepsilon/2)2^p r^p)}{n}\right)^{\frac{1}{2-\vartheta\eta}} \\
&\quad + \left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}} + \frac{3(C_\sigma(n) + \sqrt{C_E(n)} + 1) B_0 \tau C_B(n)}{n} + \delta
\end{aligned}$$

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

Introduction of the Quotients. To establish a non-trivial bound on the term $\mathbb{E}_P h_{\widehat{f}_D} - \mathbb{E}_{D_n} h_{\widehat{f}_D}$ in (2.29), we define functions

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

For $f \in \mathcal{F}$, we have $\|\mathbb{E}_P h_{\widehat{f}} - h_{\widehat{f}}\|_\infty \leq 2$. Moreover, for $f \in \mathcal{F}_r$, the variance bound (2.21) implies

$$\mathbb{E}_P(h_{\widehat{f}} - \mathbb{E}_P h_{\widehat{f}})^2 \leq \mathbb{E}_P h^2_{\widehat{f}} \leq V(\mathbb{E}_P h_{\widehat{f}})^\vartheta \leq V r^\vartheta. \tag{2.34}$$

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

$$\mathcal{F}_1 \subset \mathcal{F}_r \cup \bigcup_{k=1}^{K+1} (\mathcal{F}_{2^k r} \backslash \mathcal{F}_{2^{k-1} r}). \tag{2.35}$$

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

$$\widetilde{\mathcal{C}}_{\varepsilon,r,k} := \bigcup_{l=0}^{k} \overline{C}_{\varepsilon,r,l}, \qquad 0 \leq k \leq K+1, \tag{2.36}$$

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

$$\mathcal{N}(A, \|\cdot\|_\infty, \varepsilon) \leq \mathcal{N}(B, \|\cdot\|_\infty, \varepsilon/2). \tag{2.37}$$

By using (2.37) for $\mathcal{F}_{2^k r} \backslash \mathcal{F}_{2^{k-1} r} \subset \mathcal{F}_{2^k r}$ we can estimate the cardinality of $\widetilde{\mathcal{C}}_{\varepsilon,r,k}$ by

$$\begin{aligned}
|\widetilde{\mathcal{C}}_{\varepsilon,r,k}| &= \left| \bigcup_{l=0}^{k} \overline{C}_{\varepsilon,r,l} \right| \\
&\leq \sum_{l=0}^{k} |\overline{C}_{\varepsilon,r,l}| \\
&= \sum_{l=0}^{k} \mathcal{N}(\mathcal{F}_{2^k r} \backslash \mathcal{F}_{2^{k-1} r}, \|\cdot\|_\infty, \varepsilon) \\
&\leq \sum_{l=0}^{k} \mathcal{N}(\mathcal{F}_{2^k r}, \|\cdot\|_\infty, \varepsilon/2) \\
&\leq \sum_{l=0}^{k} \exp\left(\varphi(\varepsilon/2)(2^l r)^p\right) \\
&\leq (k+1) \exp\left(\varphi(\varepsilon/2) 2^{kp} r^p\right), \qquad 0 \leq k \leq K+1.
\end{aligned} \tag{2.38}$$

Peeling by Theorem 5.2 with $\mathcal{Z}_f := \mathbb{E}_{D_n}(\mathbb{E}_P h_{\widehat{f}} - h_{\widehat{f}})$, $\Gamma(f) := \Upsilon(f) + \mathbb{E}_P h_{\widehat{f}}$ and

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

by using $\epsilon = \frac{1}{4} > 0$ imply

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

$$\begin{aligned}
&\leq \mu \left(\sup_{f \in \widetilde{\mathcal{C}}_{\varepsilon,r,1}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\widehat{f}} - h_{\widehat{f}}) > \frac{1}{4} r \right) \\
&\quad + \sum_{k=1}^{K+1} \mu \left(\sup_{f \in \widetilde{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\widehat{f}} - h_{\widehat{f}}) > \frac{1}{4} 2^{k-1} r \right) \\
&\leq 2 \sum_{k=1}^{K+1} \mu \left(\sup_{f \in \widetilde{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\widehat{f}} - h_{\widehat{f}}) > 2^{k-3} r \right). \qquad (2.39)
\end{aligned}$$

Estimating the Error Probabilities on the “Spheres”. Now we estimate all the error probabilities in (2.39). By using the inequality (2.16) with (2.34) and the union bound, we obtain

$$\begin{aligned}
&\mu \left(\sup_{f \in \widetilde{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\widehat{f}} - h_{\widehat{f}}) > 2^{k-3} r \right) \\
&\leq C |\widetilde{\mathcal{C}}_{\varepsilon,r,k}| \exp \left(- \frac{(2^{k-3} r)^2 n}{C_\sigma(n) V (2^k r)^\vartheta + C_\eta(n) \left(V (2^k r)^\vartheta \right)^\eta + 4 C_E(n)/n + 2 C_B(n)(2^{k-3} r)} \right).
\end{aligned}$$

Our assumption $2^k r \leq 1$, $0 \leq k \leq K$ together with the last assumption in (2.22), namely,

$$r \geq \frac{16(C_\sigma(n) + \sqrt{C_E(n)} + 1) B_0 \tau C_B(n)}{n} \geq \frac{1}{n} \qquad \text{for } n \geq n_0,$$

implies that

$$\begin{aligned}
&C_\sigma(n) V (2^k r)^\vartheta + C_\eta(n) \left(V (2^k r)^\vartheta \right)^\eta + 4 C_E(n)/n + 2 C_B(n)(2^{k-3} r) \\
&\leq C_\sigma(n) V (2^k r)^{\vartheta\eta} + C_\eta(n) V (2^k r)^{\vartheta\eta} + 2 C_E(n)(2^{k-3} r) + 2 C_B(n)(2^{k-3} r) \\
&= (C_\sigma(n) + C_\eta(n)) V (2^k r)^{\vartheta\eta} + 2(C_E(n) + C_B(n))(2^{k-3} r),
\end{aligned}$$

since $\vartheta \in [0,1]$. With the estimate of the covering numbers (2.38) we obtain

$$\begin{aligned}
&\mu \left(\sup_{f \in \widetilde{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\widehat{f}} - h_{\widehat{f}}) > 2^{k-3} r \right) \\
&\leq C |\widetilde{\mathcal{C}}_{\varepsilon,r,k}| \exp \left(- \frac{(2^{k-3} r)^2 n}{(C_\sigma(n) + C_\eta(n)) V (2^k r)^{\vartheta\eta} + 2(C_E(n) + C_B(n))(2^{k-3} r)} \right) \\
&\leq C \cdot (k+1) \exp \left(\varphi(\varepsilon/2) 2^{kp} r^p \right) \cdot \\
&\quad \cdot \exp \left(- \frac{(2^{k-1} r)^2 n}{32 (C_\sigma(n) + C_\eta(n)) V (2^{k-1} r)^{\vartheta\eta} + 8(C_E(n) + C_B(n))(2^{k-1} r)} \right).
\end{aligned}$$

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

$$\begin{aligned}
p_k(r) &:= C \cdot (k+1) \exp \left(\varphi(\varepsilon/2) 2^{kp} r^p \right) \cdot \\
&\quad \cdot \exp \left(- \frac{(2^{k-1} r)^2 n}{32 (C_\sigma(n) + C_\eta(n)) V (2^{k-1} r)^{\vartheta\eta} + 8(C_E(n) + C_B(n))(2^{k-1} r)} \right).
\end{aligned}$$

Then we have

$$q_k(r) := \frac{p_{k+1}(r)}{p_k(r)}$$

$$\begin{aligned}
&\leq \frac{k+2}{k+1} \cdot \exp\left(\varphi(\varepsilon/2)(2^{k+1}r)^p - \varphi(\varepsilon/2)(2^k r)^p\right) \cdot \\
&\quad \cdot \exp\Bigg(- \frac{2^2(2^{k-1}r)^2 n}{32(C_\sigma(n)+C_\eta(n))V \cdot 2(2^{k-1}r)^\vartheta + 8(C_E(n)+C_B(n)) \cdot 2(2^{k-1}r)} \\
&\qquad\qquad + \frac{(2^{k-1}r)^2 n}{32(C_\sigma(n)+C_\eta(n))V(2^{k-1}r)^{\vartheta\eta} + 8(C_E(n)+C_B(n))(2^{k-1}r)} \Bigg) \\
&\leq 2\exp\left(\varphi(\varepsilon/2)2^{kp+1}r^p\right) \cdot \\
&\quad \cdot \exp\left(- \frac{(2^{k-1}r)^2 n}{32(C_\sigma(n)+C_\eta(n))V(2^{k-1}r)^{\vartheta\eta} + 8(C_E(n)+C_B(n))(2^{k-1}r)} \right),
\end{aligned}$$

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

$$\begin{aligned}
q_k(r) &\leq 2\exp\left(\varphi(\varepsilon/2)2^{kp+1}r^p\right) \cdot \\
&\quad \cdot \exp\left(- \frac{(2^{k-1}r)^2 n}{32(C_\sigma(n)+C_\eta(n))V(2^{k-1}r)^{\vartheta\eta} + 8(C_E(n)+C_B(n))(2^{k-1}r)} \right) \\
&\leq 2\exp\Bigg(2^{(k-1)p} \cdot 4r^p \varphi(\varepsilon/2) \\
&\qquad\qquad - 2^{(k-1)(2-\vartheta\eta)} \cdot \frac{r^{2-\vartheta\eta} n}{32(C_\sigma(n)+C_\eta(n))V + 8(C_E(n)+C_B(n))} \Bigg).
\end{aligned}$$

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

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

The second assumption in (2.22), namely,

$$r \geq \left(\frac{64(4(C_\sigma(n)+C_\eta(n))V + (C_E(n)+C_B(n)))(\tau + \varphi(\varepsilon/2)2^p r^p)}{n} \right)^{\frac{1}{2-\vartheta\eta}}$$

implies that

$$r \geq \left(\frac{64(4(C_\sigma(n)+C_\eta(n))V + (C_E(n)+C_B(n)))\varphi(\varepsilon/2) r^p}{n} \right)^{\frac{1}{2-\vartheta\eta}}$$

or equivalently that

$$4r^p\varphi(\varepsilon/2) \leq \frac{1}{2} \cdot \frac{r^{2-\vartheta\eta} n}{32(C_\sigma(n)+C_\eta(n))V + 8(C_E(n)+C_B(n))},$$

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

$$q_k(r) \leq 2\exp\left(-\frac{1}{2} \cdot \frac{r^{2-\vartheta\eta} n}{32(C_\sigma(n)+C_\eta(n))V + 8(C_E(n)+C_B(n))} \right).$$

Moreover, since $\tau \geq 1$, the second assumption in (2.22) implies also

$$r \geq \left(\frac{64(4(C_\sigma(n)+C_\eta(n))V + (C_E(n)+C_B(n)))}{n} \right)^{\frac{1}{2-\vartheta\eta}}$$

or equivalently that

$$\frac{1}{2}\cdot\frac{r^{2-\vartheta\eta}n}{32(C_\sigma(n)+C_\eta(n))V+8(C_E(n)+C_B(n))}\geq 4,$$

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

Summing all the Error Probabilities. From the discussion above we have

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

Then once again the second assumption in (2.22) gives

$$r\geq\left(\frac{(32(C_\sigma(n)+C_\eta(n))V+8(C_E(n)+C_B(n)))(\tau+\varphi(\varepsilon/2)2^pr^p)}{n}\right)^{\frac{1}{2-\vartheta\eta}}$$

and a simple transformation thus yields

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

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

$$\mathbb{E}_Ph_{\widehat{f}}-\mathbb{E}_{D_n}h_{\widehat{f}}\leq\frac{1}{4}\left(\Upsilon(f)+\mathbb{E}_Ph_{\widehat{f}}+r\right)\tag{2.40}$$

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

$$|h_{\widehat{f}_{D_n}}(x,y)-h_{\widehat{f}_{D_n,\Upsilon}}(x,y)|\leq|\widehat{f}_{D_n}(x)-\widehat{f}_{D_n,\Upsilon}(x)|\leq|f_{D_n}(x)-f_{D_n,\Upsilon}(x)|\leq\varepsilon\tag{2.41}$$

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

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

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

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

consequently, we always have

$$\Upsilon(f_{D_n}) + \mathbb{E}_P h_{\widehat{f}_{D_n}} \le 2 \left(\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} \right) + r. \tag{2.42}$$

Combining (2.41) with (2.40) and (2.42), we obtain

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

with probability μ not less than $1 - 6Ce^{-\tau}$. By combining this estimate with (2.29) and (2.33), we then obtain that

$$\begin{aligned} \Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} \le \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} &+ \left(\frac{C_\sigma(n) V \tau}{n} \right)^{\frac{1}{2-\vartheta}} + \left(\frac{\tau C_\eta(n) V}{n} \right)^{\frac{1}{2-\vartheta\eta}} \\ &+ \left(\frac{\tau C_\eta(n) B_0}{n} \right)^{\frac{1}{2-\eta}} + \frac{3(C_\sigma(n) + \sqrt{C_E(n)} + 1) B_0 \tau C_B(n)}{n} + \delta \\ &+ \frac{\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}}}{2} + \frac{5}{2}\varepsilon + \frac{1}{2} r \end{aligned}$$

holds with probability μ not less than $1 - 8Ce^{-\tau}$. From the assumptions in (2.22) follows that

$$\begin{aligned} \Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} \le \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} &+ r + r + r + r + \delta \\ &+ \frac{\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}}}{2} + \frac{5}{2}\varepsilon + \frac{1}{2} r \end{aligned}$$

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

$$\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}} \le 2\Upsilon(f_0) + 4\mathbb{E}_P h_{f_0} + 9r + 5\varepsilon + 2\delta,$$

i.e. we have shown the assertion. □

3. Learning from α-mixing Processes

In this chapter, we study the learning performance from geometrically α-mixing processes. More precisely, in Section 3.1, we recall some of the classical mixing concepts including α-, β-, and ϕ-mixing coefficients and give some concrete examples of α-mixing processes. Then, in Section 3.3, with the help of some covariance inequalities from Section 3.2, we show that some existing Bernstein-type inequalities for α-mixing processes in the literature are of the generic form. In the last section, we apply the oracle inequality established in Section 2.6 to derive learning rates for some α-mixing processes and several learning methods such as ERM, LS-SVMs using given generic kernels, and SVMs using the Gaussian RBF kernels for both least squares and quantile regression.

3.1 α-mixing Processes

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

Furthermore, let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $\mathcal{Z} := (Z_i)_{i\geq 1}$ be an $Z := X \times Y$-valued stochastic process on $(\Omega, \mathcal{A}, \mu)$, $\mathcal{A}_1^i$ and $\mathcal{A}_{i+n}^\infty$ be the σ-algebras generated by $(Z_1, \ldots, Z_i)$ and $(Z_{i+n}, Z_{i+n+1}, \ldots)$, respectively. $\mathcal{Z}$ is said to be stationary, if the $(X \times Y)^n$-valued random variables $(Z_{i_1}, \ldots, Z_{i_n})$ and $(Z_{i_1+i}, \ldots, Z_{i_n+i})$ have the same distribution for all n, i, $i_1, \ldots, i_n \geq 1$. In this case, we always write $P := \mu_{Z_0}$.

To estimate the correlation between the σ-algebras $\mathcal{A}_1^i$ and $\mathcal{A}_{i+n}^\infty$, various mixing coefficients have been proposed and used in the literature:

$$\alpha(\mathcal{Z}, n) := \sup_{A\in\mathcal{A}_1^i, B\in\mathcal{A}_{i+n}^\infty} |\mu(A \cap B) - \mu(A)\mu(B)|, \tag{3.1}$$

$$\beta(\mathcal{Z}, n) := \frac{1}{2} \sup_{\substack{K,J\geq 1,\\ (A_k)_{1\leq k\leq K}\in\mathcal{A}_1^i,\\ (B_j)_{1\leq j\leq J}\in\mathcal{A}_{i+n}^\infty}} \sum_{k=1}^{K}\sum_{j=1}^{J} |\mu(A_k \cap B_j) - \mu(A_k)\mu(B_j)|, \tag{3.2}$$

$$\phi(\mathcal{Z}, n) := \sup_{A\in\mathcal{A}_1^i, B\in\mathcal{A}_{i+n}^\infty} |\mu(B) - \mu(B|A)|, \tag{3.3}$$

$$\phi_{\text{rev}}(\mathcal{Z}, n) := \sup_{A\in\mathcal{A}_{i+n}^\infty, B\in\mathcal{A}_1^i} |\mu(B) - \mu(B|A)|. \tag{3.4}$$

In the definition of β-coefficient, the supremum is taken over all measurable partitions $(A_k)_{1\leq k\leq K}$, $(B_j)_{1\leq j\leq J}$ of Ω.

Recall that the α-coefficient (3.1) was introduced by Rosenblatt [85], while the β-mixing coefficient (3.2) was introduced by [114, 115], and was attributed there to Kol-

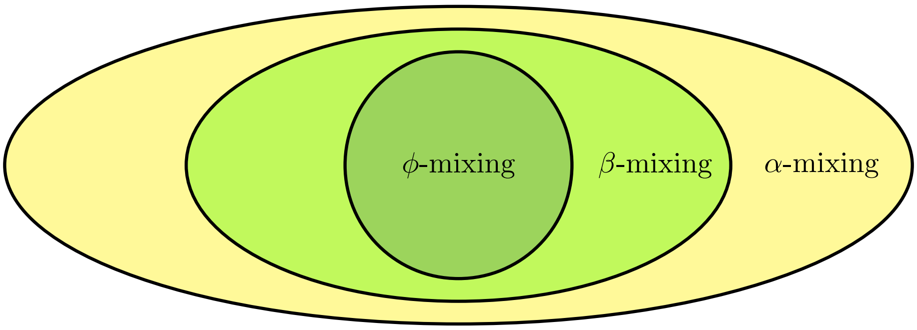


Figure 3.1: Relationship between α-, β-, and ϕ-mixing processes

mogorov. Moreover, Ibragimov [54] introduced the ϕ-coefficient, see also [56]. An extensive and thorough account on mixing concepts including β- and (time-reversed) ϕ-mixing is also provided by [21, 22, 23].

Definition 3.1. A stochastic process $\mathcal{Z} = (Z_i)_{i\geq 1}$ is called *α-mixing*, if the α-mixing coefficients satisfy

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

Similarly one can define β- and (time-reversed) ϕ-mixing sequences. Moreover, $\mathcal{Z}$ is called *geometrically α-mixing*, if we have

$$\alpha(\mathcal{Z}, n) \leq c \exp(-bn^{\gamma}), \qquad n \geq 1, \tag{3.5}$$

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

Note the asymmetry in the definition of (3.3) and (3.4). There exist stationary, countable-state Markov chains that are ϕ-mixing but not time-reversed ϕ-mixing, see e.g. [83] or [86, pp. 213-214].

It is well-known, see e.g. [21, p. 186], that these coefficients satisfy

$$2\alpha(\mathcal{Z}, n) \leq \beta(\mathcal{Z}, n) \leq \phi_{(\mathrm{rev})}(\mathcal{Z}, n). \tag{3.6}$$

By (3.6) we know that the β- and (time-reversed) ϕ-mixing sequences are also α-mixing, see Figure 3.1. Therefore, in the following sections, we will only consider the α-mixing processes.

Now let us briefly present some connections between mixing conditions and specific processes.

Example 3.2 (i.i.d. processes). As a trivial example, i.i.d. processes satisfy (3.5) with $\gamma = \infty$. ■

Example 3.3 (Markov chains). Suppose $\mathcal{Z} = (Z_i)_{i\geq 1}$ is a stationary Markov chain. Rosenblatt [86] proves that that $\mathcal{Z}$ is α-mixing if and only if it is uniformly pure nondeterministic and gives equivalent conditions for α-mixing in terms of the transition operator and the invariant probability measure. Moreover, Athreya and Pantula [6] established the α-mixing property for a wide class of Harris-recurrent Markov chains.

In the case where $\mathcal{Z}$ has countable (but not necessarily finite) state space and is irreducible and aperiodic, it satisfies β-mixing, but the mixing rate can be arbitrarily slow, see e.g. [59]. In the case where $\mathcal{Z}$ has real (but not necessarily countable) state space,

(i) Harris recurrence and aperiodicity together are equivalent to β-mixing, see [81, 80, 72], and [22, Theorem 21.6],

(ii) the geometric ergodicity condition (3.5) in [24] is equivalent to β-mixing with at least exponentially fast mixing rate, see e.g. [78, 79],

(iii) one particular version of Doeblin's condition [24, Section 3.2] is equivalent to ϕ-mixing and the mixing rate will then be at least exponentially fast, see [22, Theorem 21.23].

For this and other information on classical mixing conditions for Markov chains, see e.g. [21, Chapters 7], [22, Chapters 21], [86, Chapter 7], and [42, 72]. ■

Example 3.4 (Stationary Gaussian processes). For stationary Gaussian sequences $\mathcal{Z} = (Z_i)_{i\geq 1}$, Rozanov [87] has shown that $\mathcal{Z}$ is α-mixing, provided the spectral density function of the process [24, Section 6.1] exists everywhere and is continuous and non-vanishing over $[-\pi, \pi]$. Moreover, Ibragimov and Linnik [55] proved that $\mathcal{Z}$ is α-mixing if it has a continuous spectral density that is bounded away from 0. It is worth mentioning that Ibragimov and Rozanov [56] give characterizations of various mixing conditions in terms of properties of spectral density functions. For some further closely related information on stationary Gaussian sequences, see e.g. [24, Section 7], [21, Chapter 9], and [23, Chapter 27]. ■

Example 3.5 (Dynamical systems). Many dynamical systems have mixing properties. For certain stationary finite-state stochastic processes built on piecewise expanding mappings of the unit interval onto itself, the β-mixing condition holds with at least exponentially fast mixing rate (3.5). The same is true for some dynamical systems perturbed by dynamic noise, see e.g. [113, Chapter 3.5]. For more details on the mixing properties of these and other dynamical systems, see also [36]. ■

Example 3.6 (Linear and related processes). There is a large literature on mixing properties of stationary linear processes including stationary ARMA processes, non-causal linear processes, linear random fields, and some other related processes such as bilinear, ARCH, or GARCH models. For example, using the result derived in [6], Athreya and Pantula [6, 7] obtained a set of sufficient conditions to guarantee the α-mixing property for AR and ARMA processes. Furthermore, based on [117], Withers [118] proved that the exponentially α-mixing rate (3.5) holds for certain linear processes including some ARMA processes with $\gamma = 1$. Moreover, several time series models such as GARCH processes, which are often used to describe, e.g. financial data, satisfy (3.5) under natural conditions [44, Chapter 2.6.1]. For details on mixing properties of these and other related processes, see also [42, Chapter 2]. ■

3.2 Covariance Inequalities

In this section, we present some covariance inequalities for mixing processes which will be used in subsequent sections. For the sake of brevity, we omit the proofs. We begin with a covariance inequality for α-mixing processes derived in [14].

Proposition 3.7 (Billingsley). *Let $\mathcal{Z} = (Z_i)_{i\geq 1}$ be an α-mixing process on $(\Omega, \mathcal{A}, \mu)$. Moreover, let ξ be a bounded $\mathcal{A}_1^i$-measurable random variable and η be a bounded $\mathcal{A}_{i+n}^\infty$-measurable random variable, then we have*

$$|\mathrm{cov}(\xi, \eta)| \leq 4\alpha(\mathcal{Z}, n)\|\xi\|_\infty \|\eta\|_\infty \tag{3.7}$$

Based on the above covariance inequality in L_∞-norm, we can establish the following covariance inequality in L_p-norm, see e.g. [33].

Proposition 3.8 (Davydov). *Let $\mathcal{Z} = (Z_i)_{i\geq 1}$ be an α-mixing process on $(\Omega, \mathcal{A}, \mu)$. Moreover, for $p, q, r \geq 1$ with $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$, let $\xi \in L_p(\mu)$ be $\mathcal{A}_1^i$-measurable and $\eta \in L_q(\mu)$ be $\mathcal{A}_{i+n}^\infty$-measurable, then we have*

$$|\mathrm{cov}(\xi, \eta)| \leq 10\alpha(\mathcal{Z}, n)^{\frac{1}{r}} \|\xi\|_p \|\eta\|_q. \tag{3.8}$$

Finally, we recall the covariance inequality for ϕ-mixing processes derived in [33].

Proposition 3.9 (Davydov). *Let $\mathcal{Z} = (Z_i)_{i\geq 1}$ be a ϕ-mixing process on $(\Omega, \mathcal{A}, \mu)$. Moreover, for $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, let $\xi \in L_p(\mu)$ be $\mathcal{A}_1^i$-measurable and $\eta \in L_q(\mu)$ be $\mathcal{A}_{i+n}^\infty$-measurable, then we have*

$$|\mathrm{cov}(\xi, \eta)| \leq 2\phi(\mathcal{Z}, n)\|\xi\|_p \|\eta\|_q. \tag{3.9}$$

3.3 Bernstein-type Inequalities for α-mixing Processes

In this section, we recall some Bernstein-type inequalities for α-mixing processes that are of the general form (2.16) and give their "effective number of observations"as specified in Remark 2.26.

Example 3.10. For stationary geometrically α-mixing processes satisfying (3.5), that is,

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

for some constants $b > 0$, $c \geq 0$, and $\gamma > 0$, [73, Theorem 4.3] bounds the left-hand side of (2.16) by

$$(1 + 4e^{-2}c)\exp\left(-\frac{3\varepsilon^2 n^{(\gamma)}}{6\sigma^2 + 2\varepsilon B}\right) \tag{3.10}$$

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

$$n^{(\gamma)} := \left\lfloor n\left\lceil\left(\frac{8n}{b}\right)^{\frac{1}{\gamma+1}}\right\rceil^{-1}\right\rfloor.$$

Observe that $\lceil t \rceil \leq 2t$ for all $t \geq 1$ and $\lfloor t \rfloor \geq t/2$ for all $t \geq 2$. From this it is easy to conclude that, for all n satisfying $n \geq n_0 := \max\{b/8, 2^{2+5/\gamma} b^{-1/\gamma}\}$, we have

$$n^{(\gamma)} \geq 2^{-\frac{2\gamma+5}{\gamma+1}} b^{\frac{1}{\gamma+1}} n^{\frac{\gamma}{\gamma+1}}.$$

Hence, the left-hand side of (2.16) can also be bounded by

$$(1 + 4e^{-2}c) \exp\left(-\frac{\varepsilon^2 n^{\frac{\gamma}{\gamma+1}}}{c_\sigma \sigma^2 + \varepsilon c_B B}\right) \tag{3.11}$$

with $c_\sigma := (\frac{8^{2+\gamma}}{b})^{1/1+\gamma}$, and $c_B := c_\sigma/3$. It is not difficult to see that this bound is of the general form (2.16) with $n_0 = \max\{b/8, 2^{2+5/\gamma} b^{-1/\gamma}\}$, $C = 1 + 4e^{-2}c$, $C_\sigma(n) = c_\sigma n^{\frac{1}{\gamma+1}}$, $C_\eta(n) = 0$, $C_E(n) = 0$, and $C_B(n) = c_B n^{\frac{1}{\gamma+1}} \geq \frac{8}{3}$. Hence, we have $n_{\text{eff}} = n^{\frac{\gamma}{\gamma+1}}$. ■

Example 3.11. If $\mathcal{Z}$ is a stationary geometrically α-mixing processes satisfying (3.5) with $\gamma \geq 1$, [71, Theorem 2] established a bound for the left-hand side of (2.16) of the form

$$C_c \exp\left(-\frac{C_b \varepsilon^2 n}{v^2 + B^2/n + \varepsilon B (\log n)^2}\right), \tag{3.12}$$

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

$$v^2 := \sigma^2 + 2 \sum_{2 \leq i \leq n} |\text{cov}(h(X_1), h(X_i))|. \tag{3.13}$$

For any $\zeta > 0$, by using Davydov's covariance inequality (3.9) with $p = q = 2 + \zeta$ and $r = \frac{2+\zeta}{\zeta}$, we obtain for $i \geq 2$,

$$\begin{aligned}
\text{cov}(h(Z_1), h(Z_i)) &\leq 8 \|h(Z_1)\|_{2+\zeta} \|h(Z_i)\|_{2+\zeta} \alpha(\mathcal{Z}, i-1)^{\frac{\zeta}{2+\zeta}} \\
&\leq 8 \left(\mathbb{E}_{\mathrm{P}} h^{2+\zeta}\right)^{\frac{2}{2+\zeta}} \left(c e^{-b(i-1)}\right)^{\frac{\zeta}{2+\zeta}} \\
&\leq 8 c^{\frac{\zeta}{2+\zeta}} B^{\frac{2\zeta}{2+\zeta}} \left(\sigma^2\right)^{\frac{2}{2+\zeta}} \exp\left(-\frac{b\zeta}{2+\zeta}(i-1)\right)
\end{aligned}$$

and consequently we have

$$\begin{aligned}
v^2 &\leq \sigma^2 + 16 c^{\frac{\zeta}{2+\zeta}} B^{\frac{2\zeta}{2+\zeta}} \left(\sigma^2\right)^{\frac{2}{2+\zeta}} \sum_{2 \leq i \leq n} \exp\left(-\frac{b\zeta}{2+\zeta}(i-1)\right) \\
&\leq \sigma^2 + 16 c^{\frac{\zeta}{2+\zeta}} B^{\frac{2\zeta}{2+\zeta}} \left(\sigma^2\right)^{\frac{2}{2+\zeta}} \sum_{i=1}^{\infty} \exp\left(-\frac{b\zeta i}{2+\zeta}\right).
\end{aligned}$$

Hence, the probability bound (3.12) can be reformulated as

$$C_c \exp\left(-\frac{C_b \varepsilon^2 n}{\sigma^2 + C_\zeta \sigma^{\frac{4}{2+\zeta}} + B^2/n + \varepsilon B (\log n)^2}\right), \tag{3.14}$$

with

$$C_\zeta := 16c^{\frac{\zeta}{2+\zeta}} B^{\frac{2\zeta}{2+\zeta}} \sum_{i=1}^{\infty} \exp\left(-\frac{b\zeta i}{2+\zeta}\right). \tag{3.15}$$

It is easily seen that for $n \geq n_0 := \max\{2, e^{\sqrt{C_b}}\}$, (3.14) is also of the general form (2.16) with $C = C_c$, $C_\sigma(n) = 1/C_b$, $C_\eta(n) = C_\zeta/C_b$, $\eta = \frac{2}{2+\zeta}$, $C_E(n) = 1/C_b$, and $C_B(n) = (\log n)^2/C_b \geq 1$. Thus, we have $n_{\text{eff}} = n/(\log n)^2$.

In particular, Inequality (3.12) is valid for geometrically ϕ-mixing processes with $\gamma \geq 1$. By using the covariance inequality (1.1) for ϕ-mixing processes in [33], we can bound v^2 defined as in (3.13) by $\tilde{C}\sigma^2$ with some constant $\tilde{C}$ independent of n. Consequently the probability bound in (3.14) becomes

$$C_c \exp\left(-\frac{C_b \varepsilon^2 n}{\tilde{C}\sigma^2 + B^2/n + \varepsilon B(\log n)^2}\right), \tag{3.16}$$

which is of the general form (2.16) with $C = C_c$, $C_\sigma(n) = 0$, $C_\eta(n) = \tilde{C}/C_b$, $\eta = 1$, $C_E(n) = 0$, and $C_B(n) = (\log n)^2/C_b$. Hence, we have $n_{\text{eff}} = n$. ■

Example 3.12. For stationary, geometrically α-mixing Markov chains with centered and bounded random variables, [1] bounds the left-hand side of (2.16) by

$$\exp\left(-\frac{n\varepsilon^2}{\tilde{\sigma}^2 + \varepsilon B \log n}\right), \tag{3.17}$$

where $\tilde{\sigma}^2 = \lim_{n\to\infty} \frac{1}{n}\mathrm{Var}\sum_{i=1}^n h(X_i)$. Similar arguments as in Example 3.11 implies that, for an arbitrary $\zeta > 0$, we have

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

where C_ζ is defined as in (3.15). Consequently we obtain the bound

$$\exp\left(-\frac{n\varepsilon^2}{\sigma^2 + C_\zeta\sigma^{\frac{4}{2+\zeta}} + \varepsilon B \log n}\right), \tag{3.18}$$

which is also of the general form (2.16) with $n_0 = 3$, $C = 1$, $C_\sigma(n) = 1$, $C_\eta(n) = \tilde{C}_\zeta$, $\eta = \frac{2}{2+\zeta}$, $C_E(n) = 0$, and $C_B(n) = \log n \geq 1$. Therefore, we have $n_{\text{eff}} = n/\log n$. ■

Example 3.13. For a ϕ-mixing processes $\mathcal{Z}$, [90] provides a bound for the left-hand side of (2.16) of the form

$$\exp\left(-\frac{\varepsilon^2 n}{8C_\phi(4\sigma^2 + \varepsilon B)}\right), \tag{3.19}$$

where $C_\phi := \sum_{k=1}^{\infty} \sqrt{\phi(\mathcal{Z}, k)}$. It is of the general form (2.16) with $C = 1$, $C_\sigma(n) = 0$, $C_\eta(n) = 32C_\phi$, $\eta = 1$, $C_E(n) = 0$, and $C_B(n) = 8C_\phi$. Therefore, we have $n_{\text{eff}} = n$. ■

3.4 Learning Rates for α-mixing Processes

In this section, we use the oracle inequality from Section 2.6.2 to establish learning rates for α-mixing processes and some algorithms including ERM over finite sets and SVMs using either a given generic kernel or a Gaussian kernel with varying widths. Let us now present the first example, that is empirical risk minimization over a finite set.

Example 3.14 (ERM). Let the hypothesis set $\mathcal{F}$ be finite with $0 \in \mathcal{F}$ and $\Upsilon(f) = 0$ for all $f \in \mathcal{F}$. Moreover, assume that $\|f\|_\infty \le M$ for all $f \in \mathcal{F}$. Then, for accuracy $\delta := 0$, the learning method described by (2.6) is ERM, and Theorem 2.23 shows by some simple estimates that

$$\begin{aligned}&\mathcal{R}_{L,P}(f_{D_n,\Upsilon}) - \mathcal{R}_{L,P}^*\\ &< 4 \inf_{f\in\mathcal{F}} \left(\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^*\right) + 9\left(\frac{\tau + \ln|\mathcal{F}|}{n_{\mathrm{eff}}}\right)^{\frac{1}{2-\vartheta\eta}} + 9\left(\frac{\tau B_0}{n_{\mathrm{eff}}}\right)^{\frac{1}{2-\eta}} + \frac{9\tau B_0}{n_{\mathrm{eff}}}\end{aligned}$$

hold with probability μ not less than $1 - 8Ce^{-\tau}$. Note that in the i.i.d. case we have $n_{\mathrm{eff}} = n$ and $\eta = 1$. Besides constants, the oracle inequality (2.23) is thus an exact analogue to standard oracle inequality for ERM learning from i.i.d. processes, see e.g. [98, Theorem 7.2]. ■

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

Example 3.15 (Generic Kernels). Let $(X, \mathcal{X})$ be a measurable space, $Y = [-1, 1]$, and $\mathcal{Z}$ and P as above. Furthermore, let L be the least-squares loss and H be an RKHS over X such that the closed unit ball B_H of H satisfies

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \le a\varepsilon^{-2p}, \qquad \varepsilon > 0, \tag{3.20}$$

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

Because of Assumption (2.8), we only have to consider the hypothesis set $\mathcal{F} = \lambda^{-1/2}B_H$. Then, (2.19) implies that $\mathcal{F}_r \subset r^{1/2}\lambda^{-1/2}B_H$ and consequently we find

$$\ln \mathcal{N}(\mathcal{F}_r, \|\cdot\|_\infty, \varepsilon) \le a\lambda^{-p}\varepsilon^{-2p}r^p. \tag{3.21}$$

Thus, we can define $\varphi(\varepsilon) := a\lambda^{-p}\varepsilon^{-2p}$. For the least square loss the variance bound (2.21) is valid with $\vartheta = 1$, hence the condition (2.22) is satisfied if

$$\begin{aligned}r \ge \max\Bigg\{ &\left(2^{3p}a\right)^{\frac{1}{2-\eta-p}} \lambda^{-\frac{p}{2-\eta-p}} \left(\frac{C_V(n)}{n}\right)^{\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}},\\ &\left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}}, \frac{C_\Sigma(n) B_0 \tau C_B(n)}{n}, r^* \Bigg\},\end{aligned} \tag{3.22}$$

Therefore, let r be the sum of the terms on the right-hand side. In addition, assume that the approximation error function satisfies $A(\lambda) \le c\lambda^\beta$ for some $c > 0$, $\beta \in (0, 1]$, and all $\lambda > 0$. Since for large n, the last term in (3.22) is dominated by the others, the oracle inequality (2.23) becomes

$$\lambda\|f_{D_n,\lambda}\|_H^2 + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\Upsilon}) - \mathcal{R}_{L,P}^*$$

$$
\begin{aligned}
&\leq 4\lambda \|f_{P,\lambda}\|_H^2 + 4\mathcal{R}_{L,P}(f_{P,\lambda}) - 4\mathcal{R}_{L,P}^* + 9r + 5\varepsilon \\
&\leq C\Bigg(\lambda^\beta + \lambda^{-\frac{p}{2-\eta-p}} \left(\frac{C_V(n)}{n}\right)^{\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}} \\
&\qquad\qquad\qquad + \left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}} + \frac{C_\Sigma(n) B_0 \tau C_B(n)}{n} + \varepsilon\Bigg) \\
&\leq C\left(\lambda^\beta + \lambda^{-\frac{p}{2-\eta-p}} n_{\text{eff}}^{-\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}} + 2\left(\frac{\tau B_0}{n_{\text{eff}}}\right)^{\frac{1}{2-\eta}} + \varepsilon\right) \\
&\leq C\left(\lambda^\beta + \lambda^{-\frac{p}{2-\eta-p}} n_{\text{eff}}^{-\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}} + 2\lambda^{\beta-1} n_{\text{eff}}^{-1}\tau + \varepsilon\right),
\end{aligned}
$$

where $f_{P,\lambda}$ is the function at which the infimum in (2.12) is attained and C is a constant independent of n, λ, τ, and ε. Now optimizing over ε, we then see by [98, Lemma A.1.7] that the LS-SVM using $\lambda_n := n_{\text{eff}}^{-\rho/\beta}$ learns with rate $n_{\text{eff}}^{-\rho}$, where

$$
\rho := \min\left\{\beta, \frac{\beta}{\beta + p\beta + p}\right\}. \tag{3.23}
$$

■

In particular, for geometrically α-mixing processes, we obtain the learning rate $n^{-\alpha\rho}$, where $\alpha := \frac{\gamma}{\gamma+1}$ and ρ as in (3.23). Let us compare this rate with the ones previously established for LS-SVMs in the literature. For example, [99] proved a rate of the form

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

under exactly the same assumptions. Since $\beta > 0$ and $p > 0$, our new rate is always better than that of [99]. In addition, [45] generalized the rates of [99] to regularization terms of the form $\lambda\|\cdot\|_H^q$ with $q \in (0, 2]$. The resulting rates are again always slower than the ones we established in this work. For the standard regularization term, that is $q = 2$, [119] established the rate

$$
n^{-\frac{\alpha\beta}{2p+1}},
$$

which is always slower than ours, too. Finally, in the case $p = 1$, [106] established the rate

$$
n^{-\frac{2\alpha\beta}{\beta+3}},
$$

which was subsequently improved to

$$
n^{-\frac{3\alpha\beta}{2\beta+4}}
$$

in [107]. The latter rate is worse than ours, if and only if $(1+\beta)(1+3p) \leq 5$. In particular, for $p \in (0, 1/2]$ we always get better rates. Furthermore, the analysis of [106, 107] is restricted to LS-SVMs, while our results hold for rather generic learning algorithms.

Example 3.16 (Smooth Kernels). Let $X \subset \mathbb{R}^d$ be a compact subset, $Y = [-1, 1]$, and $\mathcal{Z}$ and P as above. Furthermore, let L be the least-squares loss and $H = W^m(X)$ be a Sobolev space with smoothness $m > d/2$. Then it is well-known, see e.g. [103] or [98, Theorem 6.26], that

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \leq a\varepsilon^{-2p}, \qquad \varepsilon > 0, \tag{3.24}$$

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

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

for the same p, see [70] and [103, Corollary 3]. Consequently, we can bound $B_0 \leq \lambda^{(\beta-1)p}$ as in [103]. Moreover, the assumption on the approximation error function is satisfied for $\beta := s/m$, whenever $f^*_{L,P} \in W^s(X)$ and $s \in (0, m]$, see e.g. [103]. Therefore, the resulting learning rate is

$$n_{\text{eff}}^{-\frac{2s}{2s+d+ds/m}}. \tag{3.25}$$

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

Again, for geometrically α-mixing processes, the rate (3.25) becomes

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

where $\alpha := \frac{\gamma}{\gamma+1}$. Comparing this rate with the one from [107], it turns out that their rate is worse than ours, if $m \geq \frac{1}{16}(2s + 3d + \sqrt{4s^2 + 108sd + 9d^2})$. Note that by the constraint $s \leq m$, the latter is always satisfied for $m \geq d$.

In the next example for LS-SVMs we will consider the Gaussian RBF kernels k_σ on X.

Example 3.17 (Gaussian Kernels). Let $Y := [-M, M]$ for $M > 0$, and P be a distribution on $\mathbb{R}^d \times Y$ such that $X := \text{supp}P_X \subset B_{\ell_2^d}$ is a bounded domain with $\mu(\partial X) = 0$. Furthermore, let P_X be absolutely continuous w.r.t. the Lebesgue measure μ on X with associated density $g : \mathbb{R}^d \to \mathbb{R}$ such that $g \in L_q(X)$ for some $q \geq 1$. Moreover, let $f^*_{L,P} : \mathbb{R}^d \to \mathbb{R}$ be a Bayes decision function such that $f^*_{L,P} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ as well as $f^*_{L,P} \in B^t_{2s,\infty}$ for $t \geq 1$ and $s \geq 1$ with $\frac{1}{q} + \frac{1}{s} = 1$. Here, $B^t_{2s,\infty}$ denotes the Besov space with the smoothness parameter t, see also [43, Section 2].

Similarly as above, since $\mathcal{F} = \lambda^{-1/2} B_{H_\sigma}$ and $\mathcal{F}_r \subset r^{1/2}\lambda^{-1/2} B_{H_\sigma}$, the covering number (2.15) implies

$$\ln \mathcal{N}(\mathcal{F}_r, \|\cdot\|_\infty, \varepsilon) \leq a_{p,\zeta} \sigma^{-(1-p)(1+\zeta)d} \lambda^{-p} \varepsilon^{-2p} r^p, \tag{3.26}$$

and thus we can define

$$\varphi(\varepsilon) := a_{p,\zeta} \sigma^{-(1-p)(1+\zeta)d} \lambda^{-p} \varepsilon^{-2p}. \tag{3.27}$$

Since the variance bound (2.21) is valid with $\vartheta = 1$ for the least-square loss, the condition (2.22) is satisfied if

$$r \geq \max\Biggl\{ \left(2^{3p} a_{p,\zeta}\right)^{\frac{1}{2-\eta-p}} \sigma^{-\frac{(1+\zeta)(1-p)d}{2-\eta-p}} \lambda^{-\frac{p}{2-\eta-p}} \left(\frac{C_V(n)}{n}\right)^{\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}}, \left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}}, \frac{C_\Sigma(n) B_0 \tau C_B(n)}{n}, r^* \Biggr\}, \tag{3.28}$$

Moreover, [43, Section 2] shows that there exists a constant $c > 0$ such that for all $\lambda > 0$ and all $\sigma \in (0, 1]$, there is an $f_0 \in H_\sigma$ with $\|f_0\|_\infty \leq c$ and

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

Again, since for large n, the last term in (3.22) is dominated by the others, the oracle inequality (2.23) becomes

$$\begin{aligned}
&\lambda \|f_{D_n,\lambda}\|_H^2 + \mathcal{R}_{L,P}(\widehat{f}_{D_n,\Upsilon}) - \mathcal{R}_{L,P}^* \\
&\leq 4\lambda \|f_{P,\lambda}\|_H^2 + 4\mathcal{R}_{L,P}(f_{P,\lambda}) - 4\mathcal{R}_{L,P}^* + 9r + 5\varepsilon \\
&\leq C\Biggl(\lambda\sigma^{-d} + \sigma^{2t} + \sigma^{-\frac{(1+\zeta)(1-p)d}{2-\eta-p}} \lambda^{-\frac{p}{2-\eta-p}} \left(\frac{C_V(n)}{n}\right)^{\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}} \\
&\qquad\qquad + \left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}} + \frac{C_\Sigma(n) B_0 \tau C_B(n)}{n} + \varepsilon \Biggr) \\
&\leq C\left(\lambda\sigma^{-d} + \sigma^{2t} + \lambda^{-\frac{p}{2-\eta-p}} n_{\text{eff}}^{-\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}} + 2\left(\frac{\tau B_0}{n_{\text{eff}}}\right)^{\frac{1}{2-\eta}} + \varepsilon \right) \\
&\leq C\left(\lambda\sigma^{-d} + \sigma^{2t} + \lambda^{-\frac{p}{2-\eta-p}} n_{\text{eff}}^{-\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}} + \varepsilon \right),
\end{aligned}$$

if n is large enough. Here C is a constant independent of n, λ, σ, τ, and ε. Again, optimizing over ε together with some standard techniques, see [98, Lemma A.1.7], we then see that for all $\xi > 0$ we can find $\zeta, p \in (0, 1)$ sufficiently close to 0 such that the LS-SVM using Gaussian RKHS H_σ and

$$\lambda_n = n_{\text{eff}}^{-1} \quad \text{and} \quad \sigma_n = n_{\text{eff}}^{-\frac{1}{2t+d}}, \tag{3.29}$$

learns with rate

$$n_{\text{eff}}^{-\frac{2t}{2t+d}+\xi}. \tag{3.30}$$

In the i.i.d. case we have $n_{\text{eff}} = n$, and hence the learning rate (3.30) becomes

$$n^{-\frac{2t}{2t+d}+\xi}. \tag{3.31}$$

Recall that modulo the arbitrarily small $\xi > 0$ these learning rates are essentially optimal, see e.g. [103, Theorem 13] or [49, Theorem 3.2]. ■

Particularly, for geometrically α-mixing processes, the rate (3.30) becomes

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

where $\alpha := \frac{\gamma}{\gamma+1}$. This rate is optimal up to the factor α and ξ in the exponent.

Analogously, for geometrically α-mixing processes satisfying (3.5) with $\gamma \geq 1$, geometrically α-mixing Markov chains, and geometrically ϕ-mixing processes, we actually obtain the essentially optimal learning rate (3.31).

To achieve these rates, however, we need to set λ_n and σ_n as in (3.29), which in turn requires us to know n_{eff} and t. Since in practice we usually do not know these values nor their existence, we can use the training/validation approach TV-SVM, see e.g. [98, Chapters 6.5, 7.4, 8.2], to achieve the same rates adaptively, i.e. without knowing n_{eff} and t. To this end, let $\Lambda := (\Lambda_n)$ and $\Sigma := (\Sigma_n)$ be sequences of finite subsets $\Lambda_n, \Sigma_n \subset (0,1]$ such that Λ_n is an ϵ_n-net of $(0,1]$ and Σ_n is an δ_n-net of $(0,1]$ with $\epsilon_n \leq n^{-1}$ and $\delta_n \leq n^{-\frac{1}{2+d}}$. Furthermore, assume that the cardinalities $|\Lambda_n|$ and $|\Sigma_n|$ grow polynomially in n. For a data set $D := ((x_1, y_1), \ldots, (x_n, y_n))$, we define

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

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

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

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

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

Then, analogous to the proof of Theorem 3.3 in [43] we can show that for all $\xi > 0$, the TV-SVM producing the decision functions $f_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\sigma_{\mathrm{D}_2}}$ with the above learning rates (3.30).

In the last example we will briefly discuss learning rates for SVMs for quantile regression. Let $X \subset \mathbb{R}^d$, $Y := [-1, 1]$, recall that the goal of quantile regression is to estimate the conditional τ-quantile, i.e. the set valued function

$$F_{\tau,P}^*(x) := \{t \in \mathbb{R} : P(Y \leq t|x) \geq \tau \text{ and } P(Y \geq t|x) \geq 1 - \tau\},$$

where $\tau \in (0,1)$ is a fixed constant. In the following example, we assume that $F_{\tau,P}^*$ consists of singletons, i.e. there exists an $f_{\tau,P}^* : X \to [-1,1]$, such that $F_{\tau,P}^*(x) = \{f_{\tau,P}^*(x)\}$ for P_X-almost all $x \in X$. To estimate the conditional τ-quantile function $f_{\tau,P}^*$, we use the so-called τ-pinball loss (2.9). It is well-known that the conditional τ-quantile function is, modulo P_X-zero sets, the only function that minimizes the L_τ-risk, that is $\mathcal{R}_{L_\tau,P}^* = \mathcal{R}_{L_\tau,P}(f_{\tau,P}^*)$.

Example 3.18 (Quantile Regression with Gaussian Kernels). Let P be a distribution on $X \times Y$ such that $\operatorname{supp} P_X \subset B_{\ell_2^d}$ and P_X is absolutely continuous with respect to the Lebesgue measure μ. Assume that the corresponding conditional density $h(\,\cdot\,, x) := \frac{dP(\,\cdot\,|x)}{d\mu|_Y}$ is uniformly bounded, that is, $h(y, x) \leq b$ for Lebesgue-almost all $y \in Y$. Then, for $p = \infty$, P has a τ-quantile of upper p-average type $q = 2$ with $\varphi(x) := b$, see

[43, Definition 4.4]. Furthermore, if we assume that, for P_X-almost all $x \in X$, the density $h(\,\cdot\,, x)$ is bounded away from 0, i.e., $h(y, x) \geq \hat{b}$ for some $0 < \hat{b} \leq b$ for Lebesgue-almost all $y \in Y$, then, for $p = \infty$, P also has a τ-quantile of lower p-average type $q = 2$ with $\kappa(x) := 2\hat{b}$, see [43, Definition 4.2]. Then for the τ-pinball loss L_τ, [100, Theorem 2.8] yields a variance bound of the form

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

for all $f : X \to \mathbb{R}$, where $V \geq 2$ is a suitable constant. Moreover, let P_X be absolutely continuous w.r.t. the Lebesgue measure on X with associated density $g \in L_u(X)$ for some $u \geq 1$ and for $\tau \in (0, 1)$, let $f^*_{\tau,P} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $f^*_{\tau,P} \in B^t_{2s,\infty}$ for some $t \geq 1$ and $s \geq 1$ such that $\frac{1}{s} + \frac{1}{u} = 1$.

For the covering numbers of the Gaussian kernels (2.15) we define $\varphi(\varepsilon)$ as in (3.27). Consequently, Condition (2.22) is satisfied if

$$\begin{aligned} r \geq \max\Bigg\{ & \left(2^{3p} a_{p,\zeta}\right)^{\frac{1}{2-\eta-p}} \sigma^{-\frac{(1+\zeta)(1-p)d}{2-\eta-p}} \lambda^{-\frac{p}{2-\eta-p}} \left(\frac{C_V(n)}{n}\right)^{\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}}, \\ & \qquad \left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}}, \frac{C_\Sigma(n) B_0 \tau C_B(n)}{n}, r^* \Bigg\}, \end{aligned} \tag{3.32}$$

Moreover, [43, Section 4] shows that there exists a constant $c > 0$ such that for all $\lambda > 0$ and all $\sigma \in (0, 1]$, there is an $f_0 \in H_\sigma$ with $\|f_0\|_\infty \leq c$ and

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

Again, since for large n, the last term in (3.32) is dominated by the others, the oracle inequality (2.23) becomes

$$\begin{aligned} & \lambda \|f_{D_n,\lambda}\|^2_H + \mathcal{R}_{L_\tau,P}(\widehat{f}_{D_n,\Upsilon}) - \mathcal{R}^*_{L_\tau,P} \\ & \leq 4\lambda \|f_{P,\lambda}\|^2_H + 4\mathcal{R}_{L,P}(f_{P,\lambda}) - 4\mathcal{R}^*_{L,P} + 9r + 5\varepsilon \\ & \leq C\Bigg(\lambda\sigma^{-d} + \sigma^{2t} + \sigma^{-\frac{(1+\zeta)(1-p)d}{2-\eta-p}} \lambda^{-\frac{p}{2-\eta-p}} \left(\frac{C_V(n)}{n}\right)^{\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}} \\ & \qquad\qquad + \left(\frac{\tau C_\eta(n) B_0}{n}\right)^{\frac{1}{2-\eta}} + \frac{C_\Sigma(n) B_0 \tau C_B(n)}{n} + \varepsilon \Bigg) \\ & \leq C\left(\lambda\sigma^{-d} + \sigma^{2t} + \lambda^{-\frac{p}{2-\eta-p}} n_{\text{eff}}^{-\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}} + 2\left(\frac{\tau B_0}{n_{\text{eff}}}\right)^{\frac{1}{2-\eta}} + \varepsilon \right) \\ & \leq C\left(\lambda\sigma^{-d} + \sigma^{2t} + \lambda^{-\frac{p}{2-\eta-p}} n_{\text{eff}}^{-\frac{1}{2-\eta-p}} \varepsilon^{-\frac{2p}{2-\eta-p}} + \varepsilon \right). \end{aligned}$$

where C is a constant independent of n, λ, σ, τ, and ε. Again, for every $\xi > 0$ we can then find $\zeta, p \in (0, 1)$ sufficiently close to 0 such that the SVM for quantile regression using Gaussian RKHS H_σ and

$$\lambda_n = n_{\text{eff}}^{-1} \quad \text{and} \quad \sigma_n = n_{\text{eff}}^{-\frac{1}{2t+d}}\,, \tag{3.33}$$

learns with rate

$$n_{\text{eff}}^{-\frac{2t}{2t+d}+\xi}. \tag{3.34}$$

■

Note that these rates is for the excess L_τ-risk, but since [100, Theorem 2.7] shows

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

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 for various α-mixing processes and adaptivity can be discussed along the lines of LS-SVMs.

4. Learning from $\mathcal{C}$-mixing Processes

In the previous chapter we derived the learning rates for α-mixing processes. However, as in Chapter 1, there exist many dynamical systems that are not α-mixing. To include such dynamical systems, [67] proposed the $\mathcal{C}$-mixing coefficients, which also generalize the classical ϕ-mixing coefficients.

In this chapter, we investigate the learning performance from geometrically $\mathcal{C}$-mixing processes. More precisely, in Section 4.1, 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 chapter, a Bernstein-type inequality for geometrically (time-reversed) $\mathcal{C}$-mixing processes will be formulated in Section 4.2. There, we also compare our new Bernstein-type inequality to previously established concentration inequalities. Since our Bernstein-type inequality is of the generic form, the oracle inequality established in Section 2.6 can also be applied to derive learning rates for SVMs and $\mathcal{C}$-mixing processes including certain dynamical systems.

4.1 $\mathcal{C}$-mixing Processes

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

Let us begin by introducing some notations. 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

$$\|f\|_{\mathcal{C}} := \|f\|_\infty + \|f\| \tag{4.1}$$

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

$$\mathcal{C}(Z) := \bigl\{f : Z \to \mathbb{R} \bigm| \|f\|_{\mathcal{C}} < \infty\bigr\}. \tag{4.2}$$

Throughout this chapter, we only consider the semi-norms $\|\cdot\|$ in (4.1) that satisfy the inequality

$$\bigl\|e^f\bigr\| \leq \bigl\|e^f\bigr\|_\infty \|f\| \tag{4.3}$$

for all $f \in \mathcal{C}(Z)$. We are mostly interested in the following examples of semi-norms satisfying (4.3).

Example 4.1. Let Z be an arbitrary set and suppose that we have $\|f\| = 0$ for all $f : Z \to \mathbb{R}$. Then, it is obviously to see that $\|e^f\| = \|f\| = 0$. Hence, (4.3) is satisfied. ■

Example 4.2. Let $Z \subset \mathbb{R}$ be an interval. A function $f : Z \to \mathbb{R}$ is said to have bounded variation on Z if its total variation $\|f\|_{BV(Z)}$ is bounded. Denote by $BV(Z)$ the set of all functions of bounded variation. It is well-known that $BV(Z)$ together with $\|f\|_\infty + \|f\|_{BV(Z)}$ forms a Banach space. Moreover, we have (4.3), i.e. we have for all $f \in \mathcal{C}(Z)$:

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

■

Proof (of Example 4.2). Consider the collection Π of ordered $n+1$-ples of points $z_0 < z_1 < \ldots < z_n \in Z$, where n is an arbitrary natural number. The total variation of a function $f : I \to \mathbb{R}$ is given by

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

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

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

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

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

for all collections Π. Taking the supremum we get $\|e^f\|_{BV} \leq \|e^f\|_\infty \|f\|_{BV}$, i.e. (4.3) is satisfied. □

Example 4.3. 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. Given a function $f \in C_{b,\alpha}(Z)$, we assume that $f(z) \geq f(z')$. Again, by using $|1 - e^t| \leq |t|$, $t \leq 0$, we obtain

$$\begin{aligned} \left|e^{f(z)} - e^{f(z')}\right| &= e^{f(z)} \left|1 - e^{f(z') - f(z)}\right| \\ &\leq \left\|e^f\right\|_\infty |f(z') - f(z)| \end{aligned}$$

$$\leq \big\|e^f\big\|_\infty |f|_\alpha |z-z'|^\alpha .$$

By interchanging the roles of $f(z)$ and $f(z')$ we find the same estimate in the case of $f(z') \geq f(z)$. Consequently we obtain $\|e^f\| \leq \|e^f\|_\infty |f|_\alpha$, i.e. (4.3) is satisfied.

As usual, we speak of Lipschitz continuous functions if $\alpha = 1$ and write $\mathrm{Lip}(Z) := C_{b,1}(Z)$. ■

Example 4.4. Let $Z \subset \mathbb{R}^d$ be an open subset. For a continuously differentiable function $f : Z \to \mathbb{R}$ we write

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

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

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

for all $f \in C^1(Z)$, i.e. (4.3) is satisfied. ■

To define certain dependency coefficients for $\mathcal{Z}$, we denote, for $\psi, \varphi \in L_1(\mu)$ satisfying $\psi\varphi \in L_1(\mu)$ the correlation of ψ and φ by

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

Several dependency coefficients for $\mathcal{Z}$ can be expressed by imposing restrictions on ψ and φ. The following definition, which is taken from [67], introduces the restrictions on ψ and φ we consider throughout this chapter.

Definition 4.5. 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 (4.1) for some semi-norm $\|\cdot\|$. Then, for $n \geq 0$, we define:

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

$$\phi_\mathcal{C}(\mathcal{Z}, n) := \sup\big\{\mathrm{cor}(\psi, h \circ Z_{k+n}) : k \geq 0, \psi \in B_{L_1(\mathcal{A}_0^k,\mu)}, h \in B_{\mathcal{C}(Z)}\big\} \tag{4.4}$$

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

$$\phi_{\mathcal{C},\mathrm{rev}}(\mathcal{Z}, n) := \sup\big\{\mathrm{cor}(h \circ Z_k, \varphi) : k \geq 0, h \in B_{\mathcal{C}(Z)}, \varphi \in B_{L_1(\mathcal{A}_{k+n}^\infty,\mu)}\big\}. \tag{4.5}$$

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

$$d_n := c\exp\big(-bn^\gamma\big), \qquad n \geq 1, \tag{4.6}$$

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

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

$$\operatorname{cor}(\psi, h \circ Z_{k+n}) \leq \|\psi\|_{L_1(\mu)} \|h\|_{\mathcal{C}}\, d_n, \tag{4.7}$$

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

$$\operatorname{cor}(h \circ Z_k, \varphi) \leq \|h\|_{\mathcal{C}} \|\varphi\|_{L_1(\mu)}\, d_n. \tag{4.8}$$

In the rest of this section we consider examples of (time-reversed) $\mathcal{C}$-mixing processes. To begin with, let us assume that $\mathcal{Z}$ is a stationary ϕ-mixing process with rate $(d_n)_{n\geq 0}$. By [33, Inequality (1.1)] we then have

$$\operatorname{cor}(\psi, \varphi) \leq \|\psi\|_{L_1(\mu)} \|\varphi\|_{L_\infty(\mu)} d_n, \qquad n \geq 1, \tag{4.9}$$

for all $\mathcal{A}_0^k$-measurable $\psi \in L_1(\mu)$ and all $\mathcal{A}_{k+n}^\infty$-measurable $\varphi \in L_\infty(\mu)$. By taking $\|\cdot\|_{\mathcal{C}} := \|\cdot\|_\infty$ and $\varphi := h \circ Z_{k+n}$, we then see that (4.7) is satisfied, i.e. $\mathcal{Z}$ is $\mathcal{C}$-mixing with rate $(d_n)_{n\geq 0}$. Finally, by similar arguments we can deduce that time-reversed ϕ-mixing processes are also time-reversed $\mathcal{C}$-mixing with the same rate. In other words we have found

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

To deal with processes that are not α-mixing [85], Rio [84] introduced the following relaxation of ϕ-mixing coefficients

$$\begin{aligned}\tilde{\phi}(\mathcal{Z}, n) &:= \sup_{\substack{k\geq 0,\\ f\in BV_1}} \big\|\mathbb{E}\big(f(Z_{k+n})\big|\mathcal{A}_0^k\big) - \mathbb{E}f(Z_{k+n})\big\|_\infty \\ &= \sup\big\{\operatorname{cor}(\psi, h \circ Z_{k+n}) : k \geq 0, \psi \in B_{L_1(\mathcal{A}_0^k,\mu)}, h \in B_{BV(Z)}\big\}\end{aligned} \tag{4.10}$$

and an analogous time-reversed coefficient

$$\begin{aligned}\tilde{\phi}_{\mathrm{rev}}(\mathcal{Z}, n) &:= \sup_{\substack{k\geq 0,\\ f\in BV_1}} \big\|\mathbb{E}\big(f(Z_k)\big|A_{k+n}^\infty\big) - \mathbb{E}f(Z_k)\big\|_\infty \\ &= \sup\big\{\operatorname{cor}(h \circ Z_k, \varphi) : k \geq 0, \varphi \in B_{L_1(\mathcal{A}_{k+n}^\infty,\mu)}, h \in B_{BV(Z)}\big\},\end{aligned}$$

where the two identities follow from [35, Lemma 4]. In other words we have

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

Moreover, [34, p. 41] shows that some uniformly expanding maps are $\tilde{\phi}$-mixing but not α-mixing. Figure 4.1 summarizes the relations between ϕ, $\tilde{\phi}$, and $\mathcal{C}$-mixing.

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

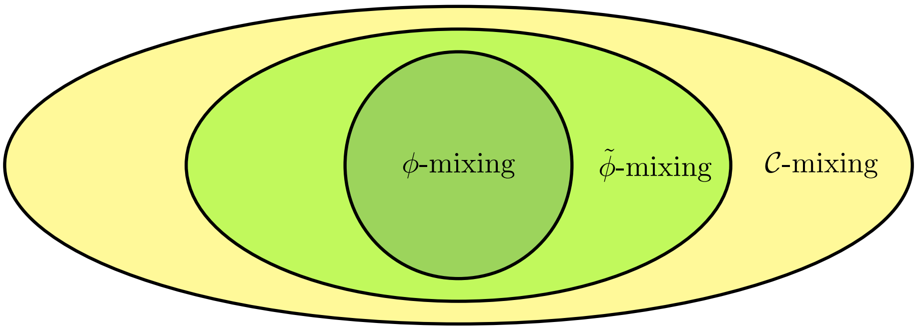


Figure 4.1: Relationship between ϕ-, $\tilde{\phi}$-, and $\mathcal{C}$-mixing processes

$\varphi = g \circ T^{k+n}$ for some suitable, measurable $g : \Omega \to \mathbb{R}$. Let us now suppose that $\|\cdot\|_{\mathcal{C}(\Omega)}$ is defined by (4.1) for some semi-norm $\|\cdot\|$. For $h \in \mathcal{C}(\Omega)$ we then find

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

The next result shows that $\mathcal{Z}$ is time-reversed $\mathcal{C}$-mixing even if we only have generic constants $C(h, g)$ in (4.8).

Theorem 4.6. *Let $(\Omega, \mathcal{A}, \mu, T)$ be a dynamical system and the stochastic process $\mathcal{Z} := (Z_n)_{n\geq 0}$ be defined by $Z_n := T^n$ for $n \geq 0$. Moreover, Let $\|\cdot\|_{\mathcal{C}}$ be defined by (4.1) for some semi-norm $\|\cdot\|$. Then, $\mathcal{Z}$ is time-reversed $\mathcal{C}$-mixing with rate $(d_n)_{n\geq 0}$ iff for all $h \in \mathcal{C}(\Omega)$ and all $g \in L_1(\mu)$ there exists a constant $C(h, g)$ such that*

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

Proof (of Theorem 4.6). ($\Rightarrow$) The proof is straightforward.
($\Leftarrow$) For $p, q \in [1, \infty]$ with $1/p + 1/q = 1$, let E_1 and E_2 be Banach spaces that are continuously embedded into $L_p(\mu)$ and $L_q(\mu)$, respectively, and let F be a Banach space that is continuously embedded into ℓ_∞. Analysis similar to that in the proof of [97, Theorem 5.1] shows that if, for all $n \geq 0$, and all $h \in E_1$, $g \in E_2$, the correlation sequence satisfies

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

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

$$\|\operatorname{cor}_{T,n}(h, g)\|_F \leq c \cdot \|h\|_{E_1} \|g\|_{E_2}, \qquad h \in E_1,\, g \in E_2. \tag{4.11}$$

In particular, (4.11) holds for $E_1 = \mathcal{C}(\Omega)$ and $E_2 = L_1(\mu)$ and the assertion is proved. □

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

For concrete examples, let us first mention that [67] presents some discrete dynamical systems that are time-reversed geometrically $\mathcal{C}$-mixing:

- Lasota-Yorke maps (piecewise expanding maps) with a finite number of intervals of monotonicity under [67, Assumption 3] with $\mathcal{C} = BV$ and $C(h,g) = \|g\|_{L_1}\|h\|_{BV}$;

- Lasota-Yorke maps with an infinite number of intervals of monotonicity under [67, Assumption 4] with $\mathcal{C} = BV$ and $C(h,g) = \|g\|_{L_1}\|h\|_{BV}$;

- Uni-modal maps under [67, Assumptions (H1)-(H3)] with $\mathcal{C} = \mathrm{Lip}$ and $C(h,g) = \|g\|_{L_1}\|h\|_{\mathrm{Lip}}$;

- Lasota-Yorke maps in higher dimension under [67, Assumption 5] with $\mathcal{C} = C_{b,\alpha}$ and $C(h,g) = \|g\|_{L_1}\|h\|_{C_{b,\alpha}}$.

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

$$|\mathrm{cor}_{T,n}(h,g)| \leq C(h,g) \cdot n^{-b}, \qquad n \geq 0, \tag{4.12}$$

for some constants $b > 0$ and $C(h,g) \geq 0$ depending on the functions h and g. Young [122] developed a powerful method for studying correlations in systems with weak chaos where correlations decay at a polynomial rate for bounded g and Hölder continuous h. Her method was applied to billiards with slow mixing rates, such as Bunimovich billiards, see [10, Theorem 3.5]. For example, modulo some logarithmic factors [66, 27] obtained (4.12) with $b = 1$ and $b = 2$ for certain forms of Bunimovich billiards and Hölder continuous h and g. Besides these results, Baladi [9] also compiles a list of "parabolic" or "intermittent" systems having a polynomial decay.

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

$$|\mathrm{cor}_{T,n}(h,g)| \leq C(h,g) \cdot \exp\bigl(-bn^{\gamma}\bigr), \qquad n \geq 0, \tag{4.13}$$

for some constants $b > 0$, $\gamma > 0$, and $C(h,g) \geq 0$ depending on h and g. Again, Baladi [9] has listed some simple examples of dynamical systems enjoying (4.13) for analytic h and g such as the angle doubling map and the Arnold's cat map. Moreover, for continuously differentiable h and g, [88, 91] proved (4.13) for two closely related classes of systems, more precisely, $C^{1+\varepsilon}$ Anosov or the Axiom-A diffeomorphisms with Gibbs invariant measures and topological Markov chains, which are also known as subshifts of finite type, see also [20]. These results were then extended by [53, 89] to expanding interval maps with smooth invariant measures for functions h and g of bounded variation. In the 1990s, similar results for Hölder continuous h and g were proved for systems with somewhat weaker chaotic behavior which is characterized by nonuniform hyperbolicity, such as quadratic interval maps, see [121], [58] and the Hénon map [12], and then extended to chaotic systems with singularities by [63] and specifically to Sinai billiards in a torus by [121, 26]. For some of these extensions, such as smooth expanding dynamics, smooth nonuniformly hyperbolic systems, and hyperbolic systems with singularities, we refer to [8] as well. Recently, for h of bounded variation and bounded g, [65] obtained (4.13) for a class of piecewise smooth one-dimensional maps with critical points and singularities. Moreover, [3] has deduced (4.13) 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.

4.2 A Bernstein-Type Inequality

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

Theorem 4.7. *Let $\mathcal{Z} := (Z_n)_{n\geq 0}$ be a Z-valued stationary geometrically (time-reversed) $\mathcal{C}$-mixing process on $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\geq 0}$ as in (4.6), $\|\cdot\|_{\mathcal{C}}$ be defined by* (4.1) *for some semi-norm $\|\cdot\|$ satisfying (4.3), and $P := \mu_{Z_0}$. Moreover, let $h \in \mathcal{C}(Z)$ with $\mathbb{E}_P h = 0$ and assume that there exist some $A > 0$, $B > 0$, and $\sigma \geq 0$ such that $\|h\| \leq A$, $\|h\|_\infty \leq B$, and $\mathbb{E}_P h^2 \leq \sigma^2$. Then, for all $\varepsilon > 0$ and all*

$$n \geq n_0 := \max\left\{\min\left\{m \geq 3 : m^2 \geq \frac{808c(3A+B)}{B} \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \geq 4\right\}, e^{\frac{3}{b}}\right\}, \tag{4.14}$$

we have

$$\mu\left(\left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^{n} h \circ Z_i \geq \varepsilon\right\}\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{8(\log n)^{\frac{2}{\gamma}}(\sigma^2 + \varepsilon B/3)}\right), \tag{4.15}$$

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

$$\mu\left(\left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^{n} h(Z_i(\omega)) \geq \sqrt{\frac{8(\log n)^{\frac{2}{\gamma}}\sigma^2\tau}{n}} + \frac{8(\log n)^{\frac{2}{\gamma}} B\tau}{3n}\right\}\right) \leq 2e^{-\tau}. \tag{4.16}$$

Note that besides the additional logarithmic factor $4(\log n)^{\frac{2}{\gamma}}$ and the constant 2 in front of the exponential, (4.15) coincides with Bernstein's classical inequality for i.i.d. processes. Moreover, notice that (4.15) is also of the general form (2.16) with n_0 as in (4.14), $C = 2$, $C_\sigma(n) = 0$, $C_\eta(n) = 8(\log n)^{\frac{2}{\gamma}}$, $\eta = 1$, $C_E(n) = 0$, and $C_B(n) = 8(\log n)^{\frac{2}{\gamma}}/3 \geq 8/3$.

4.2.1 Proof of Theorem 4.7

The following lemma, which may be of independent interest, supplies the key to the proof of Theorem 4.7.

Lemma 4.8. *Let $\mathcal{Z} := (Z_n)_{n\geq 0}$ be a Z-valued stationary (time-reversed) $\mathcal{C}$-mixing process on the probability space $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\geq 0}$, and $P := \mu_{Z_0}$. Moreover, for $f : Z \to [0, \infty)$, suppose that $f \in \mathcal{C}(Z)$ and write $f_n := f \circ Z_n$. Finally, assume that we have natural numbers k and l satisfying*

$$2l \cdot \|f\|_{\mathcal{C}} \cdot d_k \leq \|f\|_{L_1(P)}. \tag{4.17}$$

Then we have

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

Proof (of Lemma 4.8). We divide the proof into two parts.
(i) Suppose that the correlation inequality (4.7) holds. Obviously the case $f = 0$ P-a.s. is trivial. For $f \neq 0$, we define

$$D_l := \left| \mathbb{E}_\mu \prod_{j=0}^{l} f_{jk} - \prod_{j=0}^{l} \mathbb{E}_\mu f_{jk} \right| . \tag{4.18}$$

Then we have

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

Since the stochastic process $\mathcal{Z}$ is stationary, the decay of correlations (4.7) together with $\psi := \prod_{j=0}^{l-1} f_{jk}$, $h := f$, and the assumption $f \geq 0$ yields

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

Moreover, for the second term, we find

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

These estimates together imply that

$$\begin{aligned} D_l &\leq \left(D_{l-1} + \|f\|_{L_1(P)}^l \right) \|f\|_{\mathcal{C}}\, d_k + \|f\|_{L_1(P)} D_{l-1} \\ &= \left(\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}}\, d_k \right) D_{l-1} + \|f\|_{\mathcal{C}} \|f\|_{L_1(P)}^l\, d_k . \end{aligned} \tag{4.19}$$

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

$$D_l \leq \|f\|_{L_1(P)} \left(\left(\|f\|_{L_1(P)} + \|f\|_{\mathcal{C}}\, d_k \right)^l - \|f\|_{L_1(P)}^l \right) . \tag{4.20}$$

When $l = 1$, (4.20) is true because of (4.7). Now let $l \geq 1$ be given and suppose (4.20) is true for l. Then (4.19) and (4.20) imply

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

Thus, (4.20) holds for $l + 1$, and the proof of the induction step is complete. By the principle of induction, (4.20) is thus true for all $l \geq 1$.

Using the binomial formula, we obtain

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

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

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

The assumption (4.17) implies for $i = 0, \ldots, l-1$

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

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

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

This implies

$$D_l \leq \|f\|_{L_1(P)} \left(\sum_{i=0}^{l} a_i - \|f\|_{L_1(P)}^l\right)$$

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

Using the definition of D_l we thus obtain

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

(ii) Suppose that the correlation inequality (4.8) holds.
Again, the case $f = 0$ P-a.s. is trivial. For $f \neq 0$, we estimate D_l defined as in (4.18) in a slightly different way from above:

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

Since the stochastic process $\mathcal{Z}$ is stationary, the decay of correlations (4.8) together with $h := f$, $\phi := \prod_{j=1}^{l} f_{jk}$, and the assumption $f \geq 0$ yields

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

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

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

Combining the above estimates, we get

$$
D_l \leq \|f\|_{\mathcal{C}} \left(D_{l-1} + \|f\|^l_{L_1(P)}\right) d_k + \|f\|_{L_1(P)} D_{l-1}
$$

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

This estimate coincides with (4.19). The rest of the argument is the same as in *(i)*, and the assertion is proved. □

To prove Theorem 4.7, we need to introduce some notations. In the following, we write $h_i := h \circ Z_i$ and

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

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

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

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

Note that the number of the terms satisfies

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

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

$$\sum_{i=1}^k |I_i| = \sum_{i=1}^r |I_i| + \sum_{i=r+1}^k |I_i| = r(l+1) + (k-r)l = n. \tag{4.21}$$

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

$$g_i = \sum_{j \in I_i} h_j \tag{4.22}$$

such that

$$S_n = \sum_{i=1}^k g_i. \tag{4.23}$$

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

$$p_i := \frac{|I_i|}{n}. \tag{4.24}$$

It follows from (4.21) that

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

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

Lemma 4.9. *Let $\mathcal{Z} := (Z_n)_{n\geq 0}$ 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 (4.22) and (4.23), respectively. Then, for all $t > 0$, we have*

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

Proof (of Lemma 4.9). It is well-known that the exponential function is convex. Jensen's inequality together with $\sum_{i=1}^k p_i = 1$, (4.23), and (4.24) yields

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

Lemma 4.10. *Let $\mathcal{Z} := (Z_n)_{n\geq 0}$ be a Z-valued stationary (time-reversed) $\mathcal{C}$-mixing process on the probability space $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\geq 0}$, and $P := \mu_{Z_0}$. Moreover, for $h : Z \to [0, \infty)$, we write $h_n := h \circ Z_n$. Finally, let k and l be defined as above. Then, for all $t > 0$ satisfying*

$$e^{\frac{t}{|I_i|}h} \in \mathcal{C}(Z) \quad \textit{and} \quad 2l \cdot \|e^{\frac{t}{|I_i|}h}\|_{\mathcal{C}} \cdot d_k \leq \|e^{\frac{t}{|I_i|}h}\|_{L_1(P)}, \tag{4.25}$$

we have

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

Proof (of Lemma 4.10). The ith block sum g_i in (4.22) depends only on h_{i+jk} with j ranging from 0 through $|I_i| - 1$. Since $\mathcal{Z}$ is stationary, Lemma 4.8 with $f := \exp(\frac{t}{|I_i|}h)$ then yields

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

Lemma 4.11. *Let $\mathcal{Z} := (Z_n)_{n\geq 0}$ be a Z-valued stationary (time-reversed) $\mathcal{C}$-mixing process on the probability space $(\Omega, \mathcal{A}, \mu)$ with rate $(d_n)_{n\geq 0}$, and $P := \mu_{Z_0}$. Moreover, for $h : Z \to [0, \infty)$, we write $h_n := h \circ Z_n$ and suppose that $\mathbb{E}_P h = 0$, $\|h\| \leq A$, $\|h\|_\infty \leq B$, and $\mathbb{E}_P h^2 \leq \sigma^2$ for some $A > 0$, $B > 0$ and $\sigma \geq 0$. Finally, let k and l be defined as above. Then, for all $i = 1, \dots, k$, and all $t > 0$ satisfying $0 < t < 3l/B$ and (4.25), we have*

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

Proof (of Lemma 4.11). Because of $\|h\|_\infty \leq B$ and $2 \cdot 3^{j-2} \leq j!$, we obtain

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

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

$$\begin{aligned}
\left(\mathbb{E}_P \exp\left(t\frac{h}{|I_i|}\right)\right)^{|I_i|} &\leq \left(1 + \frac{1}{2}\left(\frac{t}{|I_i|}\right)^2 \sigma^2 \frac{1}{1 - tB/(3|I_i|)}\right)^{|I_i|} \\
&\leq \left(\exp\left(\frac{1}{2}\left(\frac{t}{|I_i|}\right)^2 \sigma^2 \frac{1}{1 - tB/(3|I_i|)}\right)\right)^{|I_i|} \\
&= \exp\left(\frac{t^2\sigma^2}{2(|I_i| - tB/3)}\right) \\
&\leq \exp\left(\frac{t^2\sigma^2}{2(l - tB/3)}\right), \qquad (4.26)
\end{aligned}$$

since the assumed $tB/(3l) < 1$ implies $tB/(3|I_i|) < 1$. Lemma 4.10 then yields

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

Proof (of Theorem 4.7). For k and l as above we define

$$t := \frac{l\varepsilon}{\sigma^2 + \varepsilon B/3}. \qquad (4.27)$$

Then we have

$$\frac{t}{|I_i|} \leq \frac{t}{l} = \frac{\varepsilon}{\sigma^2 + \varepsilon B/3} \leq \frac{\varepsilon}{\varepsilon B/3} = \frac{3}{B}. \qquad (4.28)$$

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

$$\left\|\exp\left(\frac{t}{|I_i|}h\right)\right\|_\infty \leq \exp\left(\frac{3}{B}\cdot B\right) = e^3. \tag{4.29}$$

Then, the assumption (4.3) together with the bounds (4.29) and (4.28) implies

$$\left\|\exp\left(\frac{t}{|I_i|}h\right)\right\| \leq \left\|\exp\left(\frac{t}{|I_i|}h\right)\right\|_\infty \left\|\frac{t}{|I_i|}h\right\| \leq e^3 \cdot \frac{t}{|I_i|}\|h\| \leq \frac{3e^3 A}{B}. \tag{4.30}$$

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

$$\left\|\exp\left(\frac{t}{|I_i|}h\right)\right\|_{L_1(P)} = \mathbb{E}_P \exp\left(\frac{t}{|I_i|}h\right) \geq \exp\left(\frac{3}{B}\cdot(-B)\right) = e^{-3}. \tag{4.31}$$

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

$$l = \frac{n-r}{k} \geq \frac{n}{k} - 1 \geq \frac{1}{2}\frac{n}{(\log n)^{\frac{2}{\gamma}}} - 1 \geq \frac{1}{4}\frac{n}{(\log n)^{\frac{2}{\gamma}}}, \tag{4.32}$$

since we have $n \geq 4(\log n)^{\frac{2}{\gamma}}$ for $n \geq n_0$. Now, by (4.29), (4.30), (4.31), (4.6), and (4.14) we obtain

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

i.e., the assumption (4.25) is valid.

Summarizing, the value of t defined as in (4.27) satisfies $0 < t < 3l/B$ and the assumption (4.25). In other words, all the requirements on t in Lemma 4.11 are satisfied.

Now, for this t, by using Markov's inequality, Lemma 4.9, and Lemma 4.11, we obtain for any $\varepsilon > 0$,

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

$$= 2\exp\left(-t\varepsilon + \frac{t^2\sigma^2}{2(l - tB/3)}\right). \tag{4.33}$$

Substituting the definition of t into the exponent of inequality (4.33), we get

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

hence

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

Using the estimate (4.32), we thus obtain

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

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

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

Simple transformations and estimations then yield

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

for all $n \ge n_0$ and $\tau > 0$. □

4.2.2 Comparisons

In the section, we compare Theorem 4.7 with some other concentration inequalities for non-i.i.d. processes $\mathcal{Z}$. Here, $\mathcal{Z}$ is real-valued and h is the identity map if not specified otherwise.

Example 4.12. Theorem 2.3 in [8] shows that smooth expanding systems on $[0,1]$ have exponential decay of correlations (4.7). Moreover, if, for such expanding systems, the transformation T is Lipschitz continuous and satisfies the conditions at the end of Section 4 in [35] and the ergodic measure μ satisfies [35, condition (4.8)], then [35, Theorem 2] shows that for all $\varepsilon \ge 0$ and $n \ge 1$, the left-hand side of (4.15) is bounded by

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

where C is some constant independent of n. The same result has been proved in [30, Theorem III.1] as well. Obviously, this is a Hoeffding-type bound instead of a Bernstein-type one. Hence, it is always larger than ours if the denominator of the exponent in (4.15) is smaller than C. ■

Example 4.13. For dynamical systems with exponentially decreasing $\tilde{\phi}$-coefficients, see [116, condition (3.1)], [116, Theorem 3.1] provides a Bernstein-type inequality for 1-Lipschitz functions $h : Z \to [-1/2, 1/2]$ w.r.t. some metric d on Z, in which the left-hand side of (4.15) is bounded by

$$\exp\left(-\frac{C\varepsilon^2 n}{\sigma^2 + \varepsilon \log f(n)}\right) \tag{4.34}$$

for some constant C independent of n and $f(n)$ being some function monotonically increasing in n. Note that modulo the logarithmic factor $\log f(n)$ the bound (4.34) is the same as the one for i.i.d. processes. Moreover, if $f(n)$ grows polynomially, cf. [116, Section 3.3], then (4.34) has the same asymptotic behaviour as our bound. However, geometrically $\mathcal{C}$-mixing is weaker than Condition (3.1) in [116]: Indeed, the required exponential form of Condition (3.1) in [116], i.e.

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

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

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

for some $c, \tilde{b} > 0$ and all $n \geq 1$, where $\mathcal{F}$ is the set of 1-Lipschitz functions $f : Z \to [-\frac{1}{2}, \frac{1}{2}]$ w.r.t. the metric d. In other words, processes satisfying Condition (3.1) in [116] are $\tilde{\phi}$-mixing, see (4.10), which is stronger than geometrically $\mathcal{C}$-mixing, see again Figure 4.1. Moreover, our result holds for all $\gamma > 0$, while [116] only considers the case $\gamma = 1$. ■

Example 4.14. In general, the probability bound in the inequality (3.10) and our result are not comparable, since not every α-mixing process satisfies (4.7) and conversely, not every process satisfying (4.7) is necessarily α-mixing, see Figure 4.2. 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 4.15. If the additional $\zeta > 0$ is ignored, (3.12) will have the same asymptotic behavior as our bound. In general, however, the additional ζ does influence the asymptotic behavior. For example, the oracle inequality we obtain in the next section would be slower by a factor of n^ξ, where $\xi > 0$ is arbitrary, if we used (3.12) instead. Finally, note that in general the bound (3.12) and ours are not comparable, see again Figure 4.2.

Moreover, it is easily seen that modulo the term $n^{-1}B$ in the denominator, the bound (3.16) coincides with ours for geometrically ϕ-mixing processes with $\gamma = 1$. However, our bound also holds for such processes with $\gamma \in (0, 1)$. ■

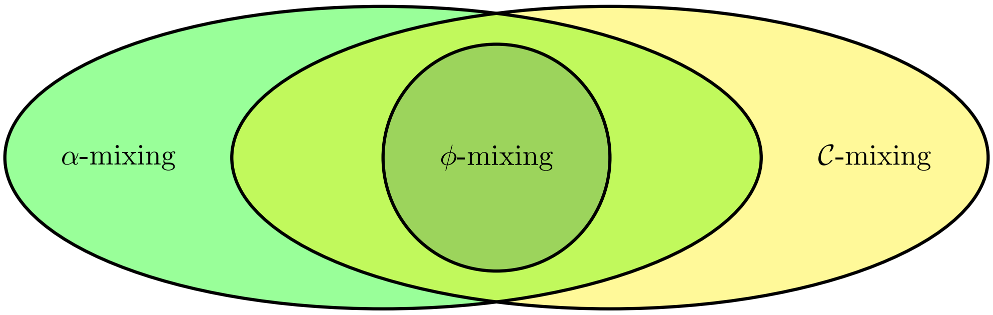


Figure 4.2: Relationship between α-, ϕ-, and $\mathcal{C}$-mixing processes

Example 4.16. Similarly as in Example 4.15, we conclude that modulo some arbitrary small number $\zeta > 0$ and the logarithmic factor $\log n$ instead of $(\log n)^2$, the bound (3.18) coincides with ours. Again, this bound and our result are not comparable, see Figure 4.2. ■

Example 4.17. For stationary, weakly dependent processes of centered and bounded random variables with $|\text{cov}(X_1, X_n)| \leq c \cdot \exp(-bn)$ for some $c, b > 0$ and all $n \geq 1$, [57, Theorem 2.1] bounds the left-hand side of (4.15) by

$$\exp\left(-\frac{\varepsilon^2 n}{C_1 + C_2 \varepsilon^{5/3} n^{2/3}}\right) \tag{4.35}$$

where C_1 is some constant depending on c and b, and C_2 is some constant depending on c, b, and B. Note that the denominator in (4.35) is at least C_1, and therefore the bound (4.35) is more of Hoeffding type. ■

4.3 Learning Rates for $\mathcal{C}$-mixing Processes

In this section, we derive learning rates for SVMS and observations generated by a geometrically $\mathcal{C}$-mixing processes. More precisely, in Subsection 4.3.1, we use the oracle inequality established in Section 2.6 to derive the learning rates for SVMs. Then, in Subsection 4.3.2, we present an oracle inequality for forecasting of time-reversed $\mathcal{C}$-mixing dynamical systems and derive the learning rates for SVMs.

4.3.1 Learning Rates for SVMs

Since the Bernstein-type inequality (4.15) is of the general form (2.16) with n_0 that depends on the semi-norm bounds A and B, $C = 2$, $C_\sigma(n) = 0$, $C_\eta(n) = 8(\log n)^{\frac{2}{\gamma}}$, $\eta = 1$, $C_E(n) = 0$, and $C_B(n) = 8(\log n)^{\frac{2}{\gamma}}/3 \geq 8/3$, we immediately conclude that the oracle inequality (2.23) holds also for geometrically (time-reversed) $\mathcal{C}$-mixing processes with

$$C_V(n) := 512(12V + 1)(\log n)^{\frac{2}{\gamma}}/3, \tag{4.36}$$

$$C_\Sigma(n) := 16(\log n)^{\frac{2}{\gamma}}, \tag{4.37}$$

and some n_0^* associated with the semi-norm to be determined in the following.

In order to formulate the threshold number n_0^* for the oracle inequality, we need to make following assumptions on the semi-norm bounds.

Assumption 4.18. *Let $\|\cdot\|$ be a semi-norm satisfying (4.3) and $\mathcal{F}_r$ be defined as in (2.19), assume that we have a monotonic decreasing sequence $(A_r)_{r\in(0,1]}$ such that*

$$\|L\circ \widehat{f}\| \leq A_r \quad \text{for all } f\in\mathcal{F}_r \text{ and } r\in(0,1]\,. \tag{4.38}$$

Then it is easily to conclude that $\|L\circ\widehat{f}\|\leq A_1$ for all $f\in\mathcal{F}_r$ and $r\in(0,1]$.

Moreover, let $f_0\in\mathcal{F}$ be a fixed function, and $A_0, A^*\geq 0$ be constants such that $\|L\circ f_0\|\leq A_0$, $\|L\circ\widehat{f}_0\|\leq A_0$, and $\|L\circ f^*_{L,P}\|\leq A^*$.

Now, recall that in the proof of Theorem 2.23, the Bernstein-type inequality has been used three times, namely, for $\mathbb{E}_{D_n}(h_{f_0}-h_{\widehat{f}_0})-\mathbb{E}_P(h_{f_0}-h_{\widehat{f}_0})$ and $\mathbb{E}_{D_n}h_{\widehat{f}_0}-\mathbb{E}_P h_{\widehat{f}_0}$ in (2.30), and $\mathbb{E}_P h_{\widehat{f}_{D_n,\Upsilon}}-\mathbb{E}_{D_n}h_{\widehat{f}_{D_n,\Upsilon}}$ in (2.29).

(i) $h:=(h_{f_0}-h_{\widehat{f}_0})-\mathbb{E}_P(h_{f_0}-h_{\widehat{f}_0})$.
Obviously, we have $L\circ f_0-L\circ\widehat{f}_0\geq 0$ which implies $h_{f_0}-h_{\widehat{f}_0}=L\circ f_0-L\circ\widehat{f}_0\in[0,B_0]$. Moreover, we find

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

Thus, n_0^* should at least be the maximum of $e^{\frac{3}{b}}$ and

$$n_0^{(1)}:=\min\left\{m\geq 3: m^2\geq\frac{808c(6A_0+B_0)}{B_0} \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}}\geq 4\right\}.$$

(ii) $h:=h_{\widehat{f}_0}-\mathbb{E}_P h_{\widehat{f}_0}$.
We first observe that the assumed $L(x,y,t)\leq 1$ for all $(x,y)\in X\times Y$ and $t,t'\in[-M,M]$ implies $\|h_{\widehat{f}_0}\|_\infty\leq 1$, and hence we have $\|h_{\widehat{f}_0}-\mathbb{E}_P h_{\widehat{f}_0}\|_\infty\leq 2$. Furthermore, we have

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

Thus, n_0^* should at least be the maximum of $e^{\frac{3}{b}}$ and

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

(iii) $h:=h_{\widehat{f}}-\mathbb{E}_{D_n}h_{\widehat{f}}$, $f\in\mathcal{F}_r$.
For $f\in\mathcal{F}_r$, we have $\|\mathbb{E}_P h_{\widehat{f}}-h_{\widehat{f}}\|_\infty\leq 2$. Furthermore, for $f\in\mathcal{F}_r$ and $k\geq 0$ with $2^k r\leq 1$, by the assumption (4.38) we find

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

Thus, n_0^* should at least be the maximum of $e^{\frac{3}{b}}$ and

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

To make the above conditions satisfied, we can set

$$\begin{aligned} n_0^* &:= \max\left\{ n_0^{(1)}, n_0^{(2)}, n_0^{(3)}, e^{\frac{3}{b}} \right\} \\ &:= \max\left\{ \min\left\{ m \geq 3 : m^2 \geq K \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \geq 4 \right\}, e^{\frac{3}{b}} \right\} \end{aligned} \tag{4.39}$$

with $K = 1212c(4A_0 + A^* + A_1 + 1)$.

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

Lemma 4.19. *Let (X, d) be a metric space, $Y \subset [-M, M]$ with $M > 0$. Moreover, let $f : X \to \mathbb{R}$ be a bounded, Lipschitz continuous function. Then the following statements hold true:*

(i) For the least square loss L, see (2.8), we have

$$|L \circ f|_1 \leq 2\sqrt{2}\,(M + \|f\|_\infty)\,(1 + |f|_1). \tag{4.40}$$

(ii) For the τ-pinball loss L, see (2.9), we have

$$|L \circ f|_1 \leq \sqrt{2}(1 + |f|_1). \tag{4.41}$$

Proof (of Lemma 4.20). *(i)* For the least square loss (2.8), by using $a + b \leq (2(a^2 + b^2))^{1/2}$, we obtain

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

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

(ii) Let L be the the τ-pinball loss (2.9) and define

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} |L(x,y,f(x)) - L(x',y',f(x'))| &\leq |(y-y') - (f(x) - f(x'))| \\ &\leq |y - y'| + |f(x) - f(x')|. \end{aligned}$$

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

If $\mathcal{C} = C^1$ and the involved functions $f \in \mathcal{F}$ are also continuous differentiable, we have a similar result as (4.40) for least square loss,

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

However, an estimate like (4.41) can not be achieved for general $f \in C^1$, since the τ-pinball loss is not differentiable at the point 0.

Finally, we give the required bounds on $\|L \circ f\|$ for $\mathcal{C} = BV$ and the involved functions $f \in \mathcal{F}$ are also of bounded variation. For the sake of brevity, we omit the proof.

Lemma 4.20. *Let (X,d) be a metric space, $Y \subset [-M,M]$ with $M > 0$. Moreover, let $f : X \to \mathbb{R}$ be a bounded, Lipschitz continuous function. Then the following statements hold true:*

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

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

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

$$\|L \circ f\|_{BV} \leq 2(M + \|f\|_{BV}).$$

In this rest of section, we derive the learning rates for LS-SVMs and SVMs for quantile regression using Gaussian RBF kernels from geometrically (time-reversed) $\mathcal{C}$-mixing processes.

Example 4.21 (Least Square Regression with Gaussian Kernels). For $M > 0$, let $Y := [-M,M]$ and P be a distribution on $\mathbb{R}^d \times Y$ such that $X := \mathrm{supp}P_X \subset B_{\ell_2^d}$ is a bounded domain with $\mu(\partial X) = 0$, where $B_{\ell_2^d}$ denotes the closed unit ball of d-dimensional Euclidean space ℓ_2^d. Furthermore, let P_X be absolutely continuous w.r.t. the Lebesgue measure μ on X with associated density $g : \mathbb{R}^d \to \mathbb{R}$ such that $g \in L_q(X)$

for some $q \geq 1$. Moreover, let $f^*_{L,P} : \mathbb{R}^d \to \mathbb{R}$ be a Bayes decision function such that $f^*_{L,P} \in L_2(\mathbb{R}^d) \cap \mathrm{Lip}(\mathbb{R}^d)$ as well as $f^*_{L,P} \in B^t_{2s,\infty}$ for some $t \geq 1$ and $s \geq 1$ with $\frac{1}{q} + \frac{1}{s} = 1$.

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

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

see e.g. [43, Section 2]. Moreover, [97, Lemma 5.5] shows every function f in H_σ is Lipschitz continuous with

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

and this implies

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

Moreover, there exists a constant $C^* < \infty$ such that $|f^*_{L,P}|_1 \leq C^*$, since we have assumed that $f^*_{L,P} \in \mathrm{Lip}(\mathbb{R}^d)$. Then, Lemma 4.20 (i) yields

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

for all $\sigma, \lambda \in (0, 1]$ with $\lambda\sigma^2 \geq n^{-2}$, where C is a constant independent of n, λ, and σ. Consequently, from (4.39) we obtain

$$n^*_0 := \max\left\{2C, \min\left\{m \geq 3 : \frac{m}{(\log m)^{\frac{2}{\gamma}}} \geq 4\right\}, e^{\frac{3}{b}}\right\}.$$

Then, similar arguments as in Example 3.17 yield that, for all $\xi > 0$, all $n \geq n^*_0$, the LS-SVM using Gaussian RKHS H_σ and

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

learns with rate

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

since the requirement $\lambda_n\sigma^2_n \geq n^{-2}$ is automatically satisfied by the assumed $t \geq 1$. ■

Again, modulo the arbitrarily small $\xi > 0$, these learning rates are optimal, see e.g. [103, Theorem 13] or [49, Theorem 3.2]. Moreover, adaptivity can be discussed along the lines in Section 3.4.

The following example discusses learning rates for SVMs for quantile regression. For more information on such SVMs we refer to [43, Section 4].

Example 4.22 (Quantile Regression with Gaussian Kernels). Let $X \subset \mathbb{R}^d$, $Y := [-1, 1]$, P be a distribution on $X \times Y$ such that $\operatorname{supp} P_X \subset B_{\ell_2^d}$ and P_X is absolutely continuous with respect to the Lebesgue measure μ. Assume that the corresponding conditional density $h(\,\cdot\,, x) := \frac{dP(\,\cdot\,|x)}{d\mu|_Y}$ is uniformly bounded, that is, $h(y, x) \leq b$ for Lebesgue-almost all $y \in Y$. Then, for $p = \infty$, P has a τ-quantile of upper p-average type $q = 2$ with $\varphi(x) := b$, see [43, Definition 4.4]. Furthermore, if we assume that, for P_X-almost all $x \in X$, the density $h(\,\cdot\,, x)$ is bounded away from 0, i.e., $h(y, x) \geq \hat{b}$ for some $0 < \hat{b} \leq b$ for Lebesgue-almost all $y \in Y$, then, for $p = \infty$, P also has a τ-quantile of lower p-average type $q = 2$ with $\kappa(x) := 2\hat{b}$, see [43, Definition 4.2]. Then for the τ-pinball loss L_τ, [100, Theorem 2.8] yields a variance bound of the form

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

for all $f : X \to \mathbb{R}$, where $V \geq 2$ is a suitable constant. Moreover, let P_X be absolutely continuous w.r.t. the Lebesgue measure on X with associated density $g \in L_u(X)$ for some $u \geq 1$ and for $\tau \in (0, 1)$, let $f^*_{\tau,P} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $f^*_{\tau,P} \in B^t_{2s,\infty}$ for some $t \geq 1$ and $s \geq 1$ such that $\frac{1}{s} + \frac{1}{u} = 1$. Similar arguments to Theorem 4.21 shows then that the essentially optimal learning rate (4.50) can be achieved as well. Note that this rate is for the excess L_τ-risk, but since [100, Theorem 2.7] shows

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

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. ■

4.3.2 Forecasting of Dynamical Systems

In this section, we proceed with the study of the forecasting problem of dynamical systems considered in [97]. First, let us recall some basic notations and assumptions. Let Ω be a compact subset of $\mathbb{R}^d$, $(\Omega, \mathcal{A}, \mu, T)$ be a dynamical system, and $S_0 \in \Omega$ be a random variable describing the true but unknown state at time 0. Moreover, for $E > 0$, assume that all observations of the stochastic process described by the sequence $\mathcal{T} := (T^n)_{n\geq 0}$ are additively corrupted by some i.i.d., $[-E, E]^d$-valued noise process $\mathcal{E} = (\varepsilon_n)_{n\geq 0}$ defined on the probability space $(\Theta, \mathcal{C}, \nu)$ which is (stochastically) independent of $\mathcal{T}$. It follows that all possible observations of the system at time $n \geq 0$ are of the form

$$X_n = T^n(S_0) + \varepsilon_n. \tag{4.42}$$

In other words, the process that generates the noisy observations (4.42) is $(T^n(S_0)+\varepsilon_n)_{n\geq 0}$. In particular, a sequence of observations $(X_0, \ldots, X_n)$ generated by this process is of the form (4.42) for a conjoint initial state S_0.

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

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

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

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

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

$$\boldsymbol{f}_{\boldsymbol{D}_n} := \left(f_{\boldsymbol{D}_n^{(1)}}, \dots, f_{\boldsymbol{D}_n^{(d)}} \right), \tag{4.43}$$

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

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

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

In other words, we build the forecaster $\boldsymbol{f}_{\boldsymbol{D}_n}$ by training separately d different decision functions on the training sets $\boldsymbol{D}_n^{(1)}, \dots, \boldsymbol{D}_n^{(d)}$. These problems can be considered as the (supervised) statistical learning problems formulated in Chapter 2 with the help of the following Notations.

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

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

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

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

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

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

$$\boldsymbol{L}(\boldsymbol{r}) = \psi(r_1) + \cdots + \psi(r_d), \qquad \boldsymbol{r} = (r_1, \dots, r_d) \in \mathbb{R}^d. \tag{4.44}$$

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

$$\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f}) := \iint \boldsymbol{L}\left(T(x) + \varepsilon_1 - \boldsymbol{f}(x + \varepsilon_0)\right) \nu(d\varepsilon)\, \mu(dx), \tag{4.45}$$

where $\varepsilon = (\varepsilon_i)_{i\geq 0}$ and $\boldsymbol{P} := \nu \otimes \mu$. Naturally, the smaller the risk, the better the forecaster is. Hence, we ideally would like to have a forecaster $\boldsymbol{f}^*_{\boldsymbol{L},\boldsymbol{P}} : \mathbb{R}^d \to \mathbb{R}^d$ that attains the minimal $\boldsymbol{L}$-risk

$$\mathcal{R}^*_{\boldsymbol{L},\boldsymbol{P}} := \inf \left\{ \mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f}) | \boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d \text{ measurable} \right\}. \tag{4.46}$$

The assumption (4.44) then implies $\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f}) = \sum_{j=1}^d \mathcal{R}_{L,P}(f_{\boldsymbol{D}_n{}^{(j)}})$ and

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

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

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

$$\boldsymbol{\Upsilon}(\boldsymbol{f}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}}) + \mathcal{R}_{\boldsymbol{L},\boldsymbol{D}_n}(\widehat{\boldsymbol{f}}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}}) < \inf_{\boldsymbol{f}\in\mathcal{F}^d} (\boldsymbol{\Upsilon}(\boldsymbol{f}) + \mathcal{R}_{\boldsymbol{L},\boldsymbol{D}_n}(\boldsymbol{f})) + d\delta \tag{4.47}$$

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

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

Theorem 4.23. *Let $\Omega \subset \mathbb{R}^d$ be compact and $(\Omega, \mathcal{A}, \mu, T)$ be a dynamical system. Suppose that the stationary stochastic process $\mathcal{T} := (T^n)_{n\geq 0}$ is geometrically time-reversed $\mathcal{C}$-mixing and $\mathcal{E} = (\varepsilon_n)_{n\geq 0}$ is some i.i.d. noise process defined on $(\Theta, \mathcal{C}, \nu)$ which is independent of $\mathcal{T}$. Furthermore, let $\boldsymbol{L} : \mathbb{R}^d \to [0, \infty)$ be a clippable and separable loss function with the corresponding loss function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ satisfying the properties described as in Theorem 2.23. Finally, let $\boldsymbol{\Upsilon} : \mathcal{F}^d \to [0, \infty)$ be a separable regularizer. Then, for all fixed $\boldsymbol{f}_0 = (f_0, \ldots, f_0)$, $\varepsilon > 0$, $\delta \geq 0$, $\tau \geq 1$, $n \geq n_0$ as in (4.39), and $r \in (0, 1]$ satisfying (2.22), every learning method defined by (4.47) satisfies with probability $\mu \otimes \nu$ not less than $1 - 16e^{-\tau}$:*

$$\begin{aligned}\boldsymbol{\Upsilon}(\boldsymbol{f}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}}) &+ \mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\widehat{\boldsymbol{f}}_{\boldsymbol{D}_n,\boldsymbol{\Upsilon}}) - \mathcal{R}^*_{\boldsymbol{L},\boldsymbol{P}} \\ &< 2\boldsymbol{\Upsilon}(\boldsymbol{f}_0) + 4\mathcal{R}_{\boldsymbol{L},\boldsymbol{P}}(\boldsymbol{f}_0) - 4\mathcal{R}^*_{\boldsymbol{L},\boldsymbol{P}} + 9dr + 5d\varepsilon + 2d\delta.\end{aligned} \tag{4.48}$$

Here the constants $C_V(n)$ and $C_\Sigma(n)$ are defined by (4.36) and (4.37), respectively.

Proof (of Theorem 4.23). From the discussion in the beginning of Section 4.3.1 we know that

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

holds with probability $\mu \otimes \nu$ not less than $1 - 16e^{-\tau}$. Using (4.44) and the definition (4.43) we then easily obtain the assertion. □

Again, this general oracle inequality can be applied to SVMs.

Example 4.24 (Least Square Regression with Gaussian Kernels). For $M > 0$, let $\Omega := [-M, M]^d$ and P be a distribution on $[-M, M]^{d+1}$ such that $\mathrm{supp} P_X \subset B_{\ell_2^d}$ is a bounded domain with $\mu(\partial X) = 0$, where $B_{\ell_2^d}$ denotes the closed unit ball of d-dimensional Euclidean space ℓ_2^d. Furthermore, let P_X be absolutely continuous w.r.t. the Lebesgue measure μ on X with associated density $g : \mathbb{R}^d \to \mathbb{R}$ such that $g \in L_q(X)$ for some $q \geq 1$. Moreover, let $f^*_{L,P} : \mathbb{R}^d \to \mathbb{R}$ be a Bayes decision function such that $f^*_{L,P} \in L_2(\mathbb{R}^d) \cap \mathrm{Lip}(\mathbb{R}^d)$ 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 $\xi > 0$, all $n \geq n_0$ with

$$n_0 := \max\left\{2C, \min\left\{m \geq 3 : \frac{m}{(\log m)^{\frac{2}{\gamma}}} \geq 4\right\}, e^{\frac{3}{b}}\right\}.$$

where C is a constant independent of n, λ, and σ, the LS-SVM using Gaussian RKHS H_σ and

$$\lambda_n = n^{-1} \quad \text{and} \quad \sigma_n = n^{-\frac{1}{2t+d}} , \tag{4.49}$$

learns with rate

$$n^{-\frac{2t}{2t+d}+\xi} . \tag{4.50}$$

Again, modulo the arbitrarily small $\xi > 0$, these learning rates are (4.50) optimal. Moreover, adaptivity can be discussed along the lines in Section 4.3.1. Finally, we present the learning rates for SVMs for quantile regression.

Example 4.25 (Quantile Regression with Gaussian Kernels). Let $\Omega = [-1, 1]^d$, P be a distribution on $X \times Y$ such that $\mathrm{supp}\, P_X \subset B_{\ell_2^d}$ and P_X is absolutely continuous with respect to the Lebesgue measure μ. Assume that the corresponding conditional density $h(\,\cdot\,, x) := \frac{dP(\,\cdot\,|x)}{d\mu|_Y}$ is uniformly bounded, that is, $h(y, x) \leq b$ for Lebesgue-almost all $y \in Y$. Then, for $p = \infty$, P has a τ-quantile of upper p-average type $q = 2$ with $\varphi(x) := b$, see [43, Definition 4.4]. Furthermore, if we assume that, for P_X-almost all $x \in X$, the density $h(\,\cdot\,, x)$ is bounded away from 0, i.e., $h(y, x) \geq \hat{b}$ for some $0 < \hat{b} \leq b$ for Lebesgue-almost all $y \in Y$, then, for $p = \infty$, P also has a τ-quantile of lower p-average type $q = 2$ with $\kappa(x) := 2\hat{b}$, see [43, Definition 4.2]. Then for the τ-pinball loss L_τ, [100, Theorem 2.8] yields a variance bound of the form

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

for all $f : X \to \mathbb{R}$, where $V \geq 2$ is a suitable constant. Moreover, let P_X be absolutely continuous w.r.t. the Lebesgue measure on X with associated density $g \in L_u(X)$ for some $u \geq 1$ and for $\tau \in (0, 1)$, let $f^*_{\tau,P} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $f^*_{\tau,P} \in B^t_{2s,\infty}$ for some $t \geq 1$ and $s \geq 1$ such that $\frac{1}{s} + \frac{1}{u} = 1$. Similar arguments to Theorem 4.21 shows then that the essentially optimal learning rate (4.50) can be achieved as well. Note that this rate is for the excess L_τ-risk, but since [100, Theorem 2.7] shows

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

for some constant $c > 0$ and all $f : X \to \mathbb{R}$, we actually obtain the same rates for $\|\widehat{f} - f^*_{\tau,P}\|^2_{L_2(P_X)}$. Again, optimality and adaptivity can be discussed along the lines of LS-SVMs.

4.4 Experiments

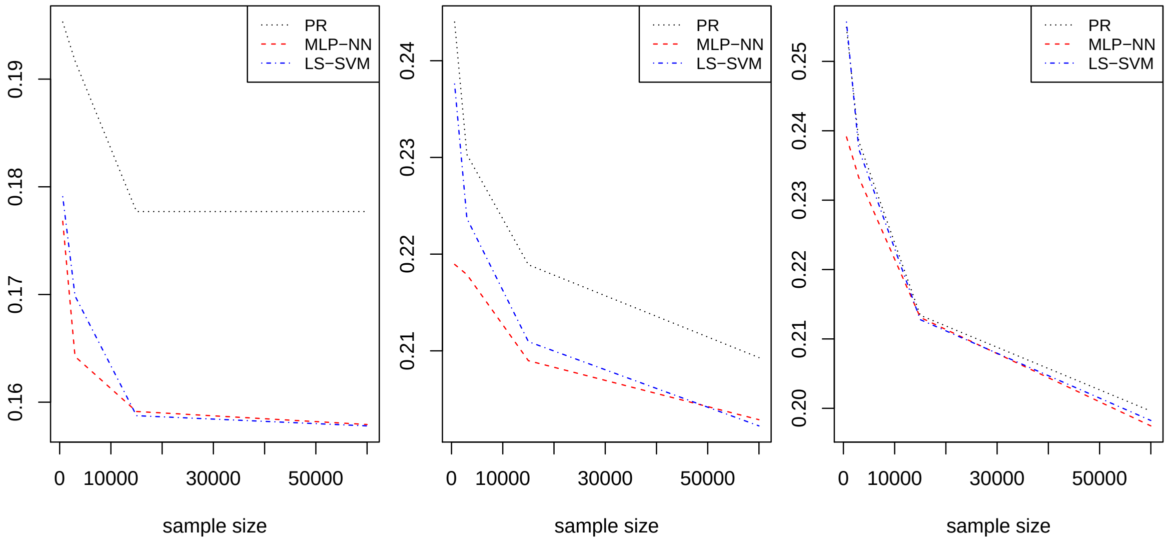
In this section, we study the learning performances of LS-SVMs on data sets generated from some dynamical systems including Logistic map, Hénon map and Lorenz System and compare the square root of mean square errors (SR-MSE) with other algorithms such as the polynomial regression (PR) and the multilayer perceptron neural network (MLP-NN). For PR, we will use polynomials of degree 2. Moreover, for MLP-NN, we always use Bayesian regularization, 10 neurons in the hidden layer, 70% as training data, 15% as validation data and 15% as testing data. Finally, we employ a new SVM library provided by Steinwart [96] to train hyper-parameter λ and σ based on a geometrically spaced 10 by 10 grid by using 5-fold cross validation.

4.4.1 Logistic Map

For $r \in [0, 4]$, the logistic map is defined as follows:

$$x_n = r \cdot x_{n-1} \cdot (1 - x_{n-1}).$$

The selection for the parameter $r = 4$ has been widely studied and in this case the dynamical system has an exponential decay of correlations of the form (4.6), see [37] and [120]. With a start value from the uniform distribution on $[0, 1]$ we generated data samples D from this system with $r = 4$. Then, by adding $\mathcal{N}(0, \sigma_i^2)$ noises to D, $\sigma_i = 0.1 \cdot i$, $i = 1, 2, 3$, we obtain three data sets D_1, D_2, D_3, respectively, with different noises.



(a) low noise **(b)** medium noise **(c)** high noise

Figure 4.3: The SR-MSE of PR, MLP-NN, and LS-SVM for the data generated from the logistic map with parameter $r = 4$ on the training set size 600, 3000, 15000, and 60000. Subfigures (a), (b), and (c) show the results for data with low noise $\mathcal{N}(0, 0.01)$, medium noise $\mathcal{N}(0, 0.04)$, and high noise $\mathcal{N}(0, 0.09)$, respectively.

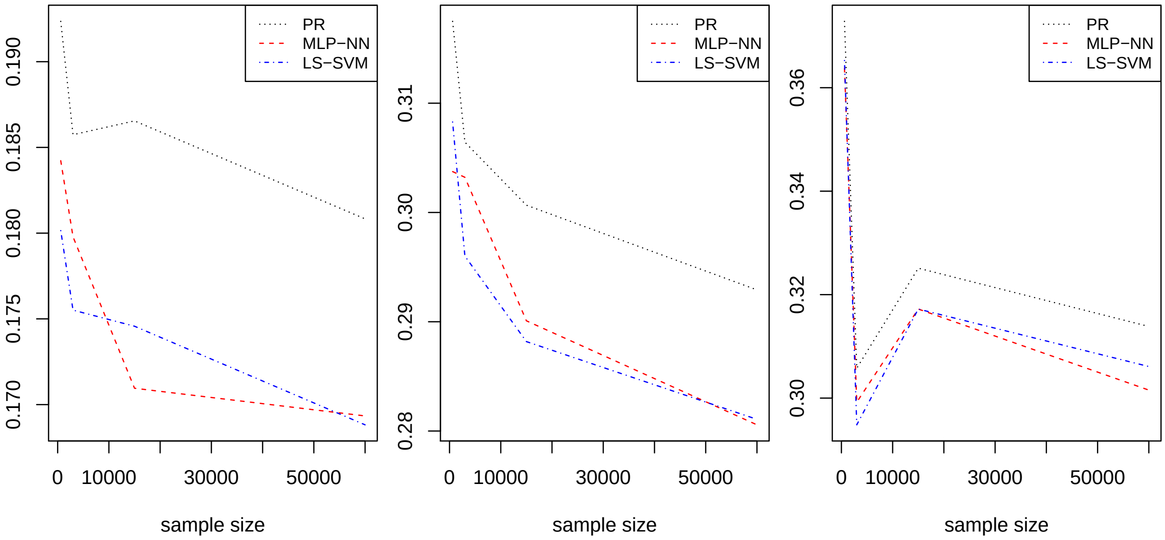
Note that for data sets with low and medium noises, MLP-NN and LS-SVM are significantly better than PR. In these cases, LS-SVM is only worse than MLP-NN for small data sets. Moreover, LS-SVM has almost the same behaviour in all three cases.

4.4.2 Hénon Map

For $a, b \in \mathbb{R}$, the Hénon map is defined as follows:

$$x_{n+1} = 1 - a \cdot x_n^2 + y_n$$
$$y_{n+1} = b \cdot x_n$$

The map depends on two parameters, a and b, [12] has proved that the Hénon map has good convergence properties, if $a < 2$ and b small. In particular, the classical Hénon map with parameters $a = 1.4$ and $b = 0.3$ has an exponential decay of correlations. With start values from the uniform distribution on $[-0.5, 0.5]$, we generated data samples from the classical Hénon map.



(a) low noise **(b)** medium noise **(c)** high noise

Figure 4.4: The SR-MSE of PR, MLP-NN, and LS-SVM for the data generated from the Hénon map with parameters $a = 1.4$ and $b = 0.3$ on the training set size 600, 3000, 15000, and 60000. Subfigures (a), (b), and (c) show the results for data with low noise $\mathcal{N}(0, 0.01)$, medium noise $\mathcal{N}(0, 0.04)$, and high noise $\mathcal{N}(0, 0.09)$, respectively.

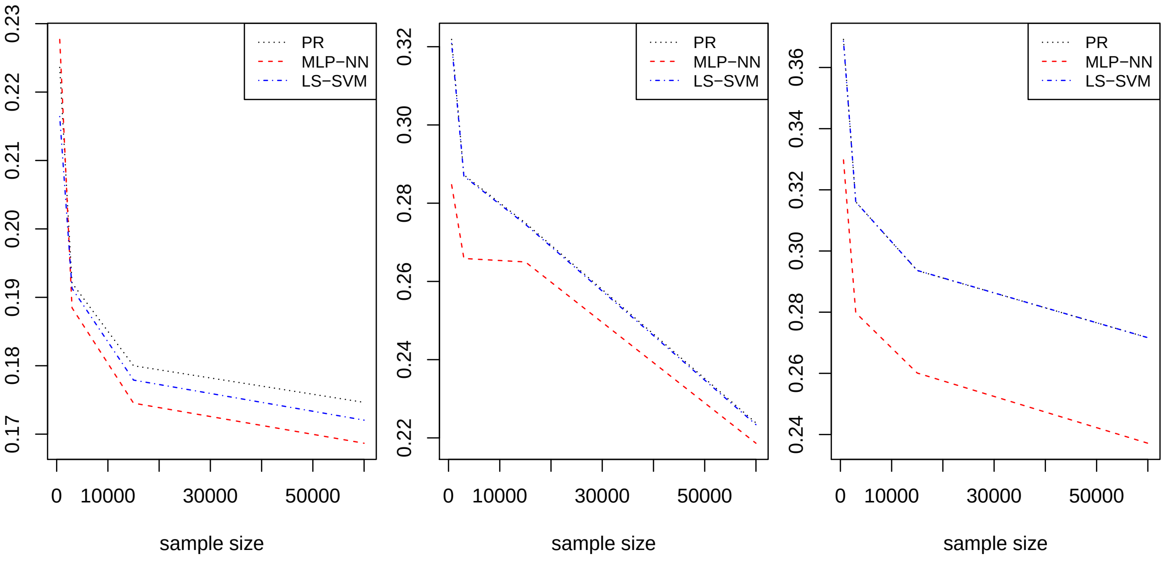
Again, for data sets with low and medium noises, MLP-NN and LS-SVM performs significantly better than PR. However, LS-SVM has a lower error than MLP-NN for small data sets and in case of low noises MLP-NN is better than LS-SVM only for medium size of data sets. Notice that there is a jump in the picture of high noises, SR-MSE becomes larger when the size of data sets increases from 3000 to 15000.

4.4.3 Lorenz System

The Lorenz system is a system of ordinary differential equations defined as:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

Recently, a first result on robust exponential decay of correlations was proved in [5] for a nonempty open subset of geometric Lorenz attractors. Unfortunately, this open set does not contain the classical Lorenz attractor with parameters $\sigma = 10$, $\rho = 28$, $\beta = 8/3$. However, [4] has shown that all C^∞ geometric Lorenz attractors including classical Lorenz attractor have superpolynomial decay of correlations in the sense of [41], that is, we have a polynomial decay of correlations of the form (4.12) with b larger than any integer. With a start value of $(x, y, z) = (-13, -14, 47)$ we have generated data samples by numerical integration using a fourth order Runge-Kutta method from the classical Lorenz attractor with parameters $\sigma = 10$, $\rho = 28$, $\beta = 8/3$.



(a) low noise **(b)** medium noise **(c)** high noise

Figure 4.5: The SR-MSE of PR, MLP-NN, and LS-SVM for the data generated from the classical Lorenz system with parameters $\sigma = 10$, $\rho = 28$, $\beta = 8/3$ on the training set size 600, 3000, 15000, and 60000. Subfigures (a), (b), and (c) show the results for data with low noise $\mathcal{N}(0, 0.01)$, medium noise $\mathcal{N}(0, 0.04)$, and high noise $\mathcal{N}(0, 0.09)$, respectively.

Roughly speaking, in all these cases, MLP-NN performs better than LS-SVM and PR. It is surprising that these latter two have almost the same behaviour.

4.4.4 Conclusions

Let us now briefly summarize the above results. Generally speaking, increasing sample sizes lead to decreasing errors except for the case of the classical Hénon map with high noises. Moreover, notice that for the logistic map and the classical Hénon map with low or medium noises, MLP-NN and LS-SVM perform significantly better than PR. In these cases, LS-SVMs has the lowest error for large data sets. However, we see that MLP-NN performs at best for the classical Lorenz systems, particularly for these with high noises. Hence, there is no algorithm which always has the best error. The performances of an algorithm depend not only on the underlying system but also on the noise strength.

5. Conclusion and Outlook

In this thesis, we established a new oracle inequality for generic regularized empirical risk minimization algorithms and used them to derive the learning rates for α- and $\mathcal{C}$-mixing processes and some learning methods such as ERM and SVMs.

In Section 2, we first presented some elementary notions of statistical learning theory. Then, based on a generic form of Bernstein's inequality for stationary stochastic processes, we derived an oracle inequality for a generic class of learning algorithms including ERM and SVMs. Chapter 3 was then dedicated to investigate geometrically α-mixing processes $\mathcal{Z}$ for which the α-mixing coefficients satisfy

$$\alpha(\mathcal{Z}, n) \leq c \exp\bigl(-bn^\gamma\bigr), \qquad n \geq 1,$$

for some constants $b > 0$, $c \geq 0$, and $\gamma > 0$. When our oracle inequality applied to ERM, it turns out that our oracle inequality coincides with the one for ERM learning from i.i.d. processes up to some constants and the effective number of observations n_{eff} as in (2.26). Furthermore, we obtained learning rates for LS-SVMs using given generic kernels of the form

$$n_{\text{eff}}^{-\min\bigl\{\beta, \frac{\beta}{\beta+p\beta+p}\bigr\}}, \tag{5.1}$$

which are slightly worse than the recently obtained optimal rates [103] for i.i.d. observations because of the factor $p\beta$ in the denominator. This difference is not surprising, when considering the fact that [103] used heavy machinery from empirical process theory such as Talagrand's inequality and localized Rademacher averages, while our results only use a light-weight argument based on a generic Bernstein inequality and the peeling method. However, when using sufficiently smooth kernels like Gaussian kernels for LS-SVMs and SVMs for quantile regression, we actually obtain the rate

$$n_{\text{eff}}^{-\frac{2t}{2t+d}+\xi}. \tag{5.2}$$

Modulo the arbitrarily small $\xi > 0$, (5.2) is optimal for geometrically α-mixing processes satisfying (3.5) up to the factor $\frac{\gamma}{\gamma+1}$ in the exponent and optimal for geometrically α-mixing processes satisfying (3.5) with $\gamma \geq 1$, geometrically α-mixing Markov chains, and geometrically ϕ-mixing processes.

In Chapter 4, we established a Bernstein-type inequality for geometrically $\mathcal{C}$-mixing processes with rate of decay

$$d_n = c \exp\bigl(-bn^\gamma\bigr), \qquad n \geq 1,$$

for some constants $b > 0$, $c \geq 0$, and $\gamma > 0$. It also turns out to be a Bernstein inequality of general form in Chapter 3. Hence, we obtain the same oracle inequality as in Chapter 3 with $n_{\mathrm{eff}} = n/(\log n)^{\frac{2}{\gamma}}$ for $n \geq n_0$ with n_0 being a number associated with the semi-norm. Applying this oracle inequality to SVMs using the Gaussian kernels for both least squares and quantile regression, it turns out that the resulting learning rates match, up to some arbitrarily small extra term in the exponent, the optimal rates for i.i.d. processes.

Finally, we list some open questions about the refinement of the analysis of learning from non-i.i.d. observations:

1. *Oracle inequalities for non-stationary processes.* In this thesis, we always assumed that the processes are stationary. However, there exist dynamical systems that are not stationary, see e.g. [47], but asymptotically mean stationary, see e.g. [46] and [102, Definition 2.2].

2. *Optimal learning rates for LS-SVMs with generic kernels.* The peeling approach enabled us to achieve the suboptimal rate (5.1) with an additional factor βp in the denominator. Recall that for i.i.d. processes, the optimal learning rates are obtained with the help of some heavy machinery from empirical process theory such as Talagrand's inequality, doubly localized Rademacher averages, and some recent estimates on expectations of random covering numbers. However, so far, for non-i.i.d. observations, these techniques are not available in the literature.

3. *Optimal learning rates for SVMs using Gaussian kernels.* So far, for Gaussian kernels, we only achieved the essentially optimal learning rates (5.2). Comparing to the optimal learning rates for i.i.d. processes, there is an additional term $\xi > 0$ in the exponent.

4. *Bernstein-type inequality for geometrically α-mixing processes.* As we have seen, the Bernstein-type inequality established by [73] has the effective number of observations $n^{\frac{\gamma}{\gamma+1}}$, which was improved to $n/(\log n)^2$ by [71] for the case $\gamma \geq 1$. However, a refinement for $\gamma \in (0, 1)$ is still a challenging problem.

Bibliography

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

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

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

[4] V. Araujo, I. Melbourne, and P. Varandas. Rapid mixing for the lorenz attractor and statistical limit laws for their time-1 maps. *arXiv preprint arXiv:1311.5017*, 2013.

[5] V. Araújo and P. Varandas. Robust exponential decay of correlations for singular-flows. *Commun. Math. Phys.*, 311(1):215–246, 2012.

[6] K. B. Athreya and S. G. Pantula. Mixing properties of Harris chains and autoregressive processes. *J. Appl. Probab.*, 23(4):880–892, 1986.

[7] K. B. Athreya and S. G. Pantula. A note on strong mixing of ARMA processes. *Statist. Probab. Lett.*, 4(4):187–190, 1986.

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

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

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

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

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

[13] S. Bernstein. *The theory of Probabilities.* Gastehizdat Publishing House, Moscow, 1946.

[14] P. Billingsley. *Convergence of probability measures.* John Wiley & Sons, Inc., New York-London-Sydney, 1968.

[15] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36(2):489–531, 2008.

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

[17] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

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

[19] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

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

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

[22] R. C. Bradley. *Introduction to strong mixing conditions. Vol. 2.* Kendrick Press, Heber City, UT, 2007.

[23] R. C. Bradley. *Introduction to strong mixing conditions. Vol. 3.* Kendrick Press, Heber City, UT, 2007.

[24] Richard C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2:107–144, 2005. Update of, and a supplement to, the 1986 original.

[25] D. Chen, Q. Wu, Y. Ying, and D. Zhou. Support vector machine soft margin classifiers: error analysis. *J. Mach. Learn. Res.*, 5:1143–1175, 2003/04.

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

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

[28] A. Christmann and I. Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

[29] X. Chu and H. Sun. Regularized least square regression with unbounded and dependent sampling. In *Abstract and Applied Analysis*, volume 2013. Hindawi Publishing Corporation, 2013.

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

[31] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[32] F. Cucker and D. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007. With a foreword by Stephen Smale.

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

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

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

[36] Manfred Denker. The central limit theorem for dynamical systems. In *Dynamical systems and ergodic theory (Warsaw, 1986)*, volume 23 of *Banach Center Publ.*, pages 33–62. PWN, Warsaw, 1989.

[37] Robert L Devaney. *A first course in chaotic dynamical systems.* Westview Press, 1992.

[38] L. Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):154–157, 1982.

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

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

[41] Dmitry Dolgopyat. Prevalence of rapid mixing in hyperbolic flows. *Ergod. Th. Dynam. Syst.*, 18(05):1097–1114, 1998.

[42] P. Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics.* Springer-Verlag, New York, 1994. Properties and examples.

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

[44] J. Fan and Q. Yao. *Nonlinear Time Series.* Springer, New York, 2003.

[45] Y. Feng. Least-squares regularized regression with dependent samples and q-penalty. *Appl. Anal.*, 91(5):979–991, 2012.

[46] R. M. Gray. *Probability, random processes, and ergodic properties.* Springer, Dordrecht, second edition, 2009.

[47] R. M. Gray and J. C. Kieffer. Asymptotically mean stationary measures. *Ann. Probab.*, 8(5):962–973, 1980.

[48] Z. Guo and L. Shi. Classification with non-iid sampling. *Math. Comput. Modelling*, 54(5):1347–1364, 2011.

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

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

[51] H. Hang and I. Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *Tech. Rep. 2015-006, Fakultät für Mathematik und Physik, Universität Stuttgart*, 2015.

[52] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

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

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

[55] I. A. Ibragimov and Yu. V. Linnik. *Independent and stationary sequences of random variables.* Wolters-Noordhoff Publishing, Groningen, 1971. With a supplementary chapter by I. A. Ibragimov and V. V. Petrov, Translation from the Russian edited by J. F. C. Kingman.

[56] I. A. Ibragimov and Y. A. Rozanov. *Gaussian random processes*, volume 9 of *Applications of Mathematics.* Springer-Verlag, New York-Berlin, 1978. Translated from the Russian by A. B. Aries.

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

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

[59] H. Kesten and G. L. O'Brien. Examples of mixing sequences. *Duke Math. J.*, 43(2):405–415, 1976.

[60] A. N. Kolmogorov. On certain asymptotic characteristics of completely bounded metric spaces. *Dokl. Akad. Nauk SSSR (N.S.)*, 108:385–388, 1956.

[61] A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17:277–364, 1961.

[62] V. Koltchinskii and O. Beznosova. Exponential convergence rates in classification. In *Learning theory*, volume 3559 of *Lecture Notes in Comput. Sci.*, pages 295–307. Springer, Berlin, 2005.

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

[64] A. Lozano, S. Kulkarni, and R. Schapire. Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, pages 819–826. MIT Press, Cambridge, MA, 2006.

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

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

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

[68] C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[69] R. Meir. Nonparametric time series prediction through adaptive model selection. *Mach. Learn.*, 39:5–34, 2000.

[70] S. Mendelson and J. Neeman. Regularization in kernel learning. *Ann. Statist.*, 38(1):526–565, 2010.

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

[72] S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Cambridge University Press, Cambridge, second edition, 2009. With a prologue by Peter W. Glynn.

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

[74] M. Mohri and A. Rostamizadeh. Stability bounds for non-i.i.d. processes. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 1025–1032. MIT Press, Cambridge, MA, 2008.

[75] M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21, pages 1097–1104. MIT Press, Cambridge, MA, 2009.

[76] M. Mohri and A. Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *J. Mach. Learn. Res.*, 4:1–26, 2009.

[77] A.B. Nobel. Limits to classification and regression estimation from ergodic processes. *Ann. Statist*, 27:262–273, 1999.

[78] E. Nummelin and P. Tuominen. Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory. *Stochastic Process. Appl.*, 12(2):187–202, 1982.

[79] E. Nummelin and R. L. Tweedie. Geometric ergodicity and R-positivity for general Markov chains. *Ann. Probability*, 6(3):404–420, 1978.

[80] S. Orey. Recurrent Markov chains. *Pacific J. Math.*, 9:805–827, 1959.

[81] S. Orey. *Lecture notes on limit theorems for Markov chain transition probabilities.* Van Nostrand Reinhold Co., London-New York-Toronto, Ont., 1971. Van Nostrand Reinhold Mathematical Studies, No. 34.

[82] Z. Pan and Q. Xiao. Least-square regularized regression with non-iid sampling. *J. Statist. Plann. Inference*, 139(10):3579–3587, 2009.

[83] Magda Peligrad. A note on two measures of dependence and mixing sequences. *Adv. in Appl. Probab.*, 15(2):461–464, 1983.

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

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

[86] M. Rosenblatt. *Markov processes. Structure and asymptotic behavior.* Springer-Verlag, New York-Heidelberg, 1971. Die Grundlehren der mathematischen Wissenschaften, Band 184.

[87] Y. A. Rozanov. *Stationary random processes.* Translated from the Russian by A. Feinstein. Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam, 1967.

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

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

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

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

[92] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2(1):67–93, 2002.

[93] I. Steinwart. Support vector machines are universally consistent. *J. Complexity*, 18(3):768–791, 2002.

[94] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Inform. Theory*, 51(1):128–142, 2005.

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

[96] I. Steinwart. A new SVM library: usage and implementational details. *Institute for Stochastics and Applications, University of Stuttgart*, 2015.

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

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

[99] I. Steinwart and A. Christmann. Fast learning from non-i.i.d. observations. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 1768–1776. MIT Press, Cambridge, MA, 2009.

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

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

[102] I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. *J. Multivariate Anal.*, 100:175–194, 2009.

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

[104] I. Steinwart and C. Scovel. Fast rates for support vector machines. In *Learning theory*, volume 3559 of *Lecture Notes in Comput. Sci.*, pages 279–294. Springer, Berlin, 2005.

[105] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35(2):575–607, 2007.

[106] H. Sun and Q.Wu. A note on application of integral operator in learning theory. *Appl. Comput. Harmon. Anal.*, 26(3):416–421, 2009.

[107] H. Sun and Q.Wu. Regularized least square regression with dependent samples. *Adv. Comput. Math.*, 32:175–189, 2010.

[108] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[109] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.

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

[111] S. A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 2000.

[112] V. N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, New York, 1995.

[113] M. Vidyasagar. *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems.* Springer, London, 2^{nd} edition, 2003.

[114] V. A. Volkonskii and Yu. A. Rozanov. Some limit theorems for random functions. I. *Theor. Probability Appl.*, 4:178–197, 1959.

[115] V. A. Volkonskii and Yu. A. Rozanov. Some limit theorems for random functions. II. *Teor. Verojatnost. i Primenen.*, 6:202–215, 1961.

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

[117] C. S. Withers. Central limit theorems for dependent variables. I. *Z. Wahrsch. Verw. Gebiete*, 57(4):509–534, 1981.

[118] C. S. Withers. Conditions for linear processes to be strong-mixing. *Z. Wahrsch. Verw. Gebiete*, 57(4):477–480, 1981.

[119] Y.-L. Xu and D.-R. Chen. Learning rates of regularized regression for exponentially strongly mixing sequence. *J. Statist. Plann. Inference*, 138:2180–2189, 2008.

[120] L.-S. Young. Decay of correlations for certain quadratic maps. *Comm. Math. Phys.*, 146(1):123–138, 1992.

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

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

[123] B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Ann. Probab.*, 22:94–116, 1994.

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

[125] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.

[126] Y. Zhang, F. Cao, and C. Yan. Learning rates of least-square regularized regression with strongly mixing observation. *Int. J. Mach. Learn. & Cyber.*, 3(4):277–283, 2012.

[127] B. Zou and L. Li. The performance bounds of learning machines based on exponentially strongly mixing sequences. *Comput. Math. Appl.*, 53:1050–1058, 2007.

[128] B. Zou, L. Li, and Z. Xu. The generalization performance of ERM algorithm with strongly mixing observations. *Mach. Learn.*, 75:275–295, 2009.