We present a new oracle inequality for generic regularized empirical risk minimization algorithms learning from stationary $\alpha$-mixing processes. Our main tool to derive this inequality is a rather involved version of the so-called peeling method. We then use this oracle inequality to derive learning rates for some learning methods such as empirical risk minimization (ERM), least squares support vector machines (SVMs) using given generic kernels, and 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.

1 Introduction

The goal of statistical learning is to estimate a functional relationship between an input random variable and an $\mathbb{R}$-valued output random variable from data, where the joint distribution of these two random variables is assumed to be (almost) completely unknown. During the statistical analysis of corresponding learning methods, one typically postulates that the data is generated in an i.i.d. fashion, we refer to the textbooks [6, 11, 5, 28] and the references therein. In practice, however, 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
 mixing concepts such as $\alpha$-, $\beta$-, and $\phi$-mixing, 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. As a result, these mixing concepts have already been used to analyze
 learning methods for non-i.i.d. data. For example, [18] analyzed forecasting algorithms
 by establishing a Bernstein inequality for the $\alpha$-mixing processes, while [40, 16, 19,
 20, 21] used the so-called blocking technique to study other concentration properties
 for $\beta$- and $\phi$-mixing processes. Subsequently, these results have been used to analyze
 particular learning algorithms. For example, [10] studied a classification algorithm
 based on a regularization scheme in a reproducing kernel Hilbert space and a generic
 convex loss function for $\alpha$- and $\beta$-mixing processes. For the regression problem, [31]
 established consistency of support vector machines (SVMs) learning from $\alpha$-mixing
 processes, while [39, 34, 33, 23] analyzed least squares support vector machines (LS-
 SVMs) with $\alpha$-mixing inputs, and [42] established generalization bounds for empirical
 risk minimization (ERM) when the sampling sequence satisfies an $\alpha$-mixing condition.
 More recently, [29] obtained a general oracle inequality for generic regularized learning
 algorithms from $\alpha$-mixing observations, [43] analyzed the generalized performance of
 empirical risk minimization algorithms with $\alpha$-mixing samples and [41, 9, 4] considered
 the regularized learning algorithm associated with the least-square loss and $\alpha$-mixing
 observations. For the smaller class of $\beta$-mixing processes, PAC-learning questions have
 been investigated in [37], while [15] established consistency of regularized boosting
 algorithms learning from $\beta$-mixing processes. For the even smaller class of $\phi$-mixing
 processes, we refer to [34].

The goal of this paper is to derive a new oracle inequality for generic regularized
 learning algorithms and $\alpha$-mixing observations. On the technical side, the new oracle
 inequality is achieved by a refinement of the analysis of [29]. To be more precise, the
 analysis in [29] partially ignored localization with respect to the regularization term,
 which we now address by a carefully arranged peeling approach inspired from [28]. As
 a result, the stochastic error term of our new oracle inequality is always smaller than
 that of [29]. Recall, that for i.i.d. processes localization with respect to the variance was
 pioneered by [14] and refined by [1] for general learning methods, whereas localization
 with respect to both variance and regularization is known to lead to optimal rates for
 LS-SVMs, see [32, 7]. However, these i.i.d. 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 [25]. For non-i.i.d. observations, most of these techniques are not
 available, so that only some light-weight arguments can be used.

As far as we know, the best learning rates for LS-SVMs from exponentially $\alpha$-
 mixing process are those derived in [39, 34, 33, 9]. When applied to LS-SVMs, it
 turns out that our oracle inequality leads to faster learning rates in [39] and [9]. For
 sufficiently smooth kernels, our rates are also faster than those in [33] and [34], we
refer to Examples 4.2 and 4.3 for more precisely comparison between all these learning rates. Moreover, for ERM, our results match those in the i.i.d. case, if one replaces the number of observations with the “effective number of observations”. 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 rates.

The rest of this work is organized as follows: The basic definitions of learning from \( \alpha \)-mixing processes will be given in Section 2. In Section 3 the refined oracle inequality will be formulated, some applications of the refined oracle inequality will be given in Section 4. The last section contains the proof of the refined oracle inequality.

2 Preliminaries

The goal of this section is to introduce the concepts needed throughout this work. Namely, we recall several notions of mixing and discuss properties of loss functions and risks. Moreover, we introduce the generic type of learning method we investigate in the subsequent sections.

Let \((X, X')\) be a measurable space and \(Y \subset \mathbb{R}\) be closed. Furthermore, let \((\Omega, \mathcal{A}, \mu)\) be a probability space, \(Z := (Z_i)_{i \geq 1}\) be an \(X \times Y\)-valued stochastic process on \((\Omega, \mathcal{A}, \mu)\), \(\mathcal{A}_1^i\) and \(\mathcal{A}_1^{\infty} \): be the \(\sigma\)-algebras generated by \((Z_1, \ldots, Z_i)\) and \((Z_{i+n}, Z_{i+n+1}, \ldots)\), respectively. Throughout this work, we assume that \(Z\) is stationary, that is, the \((X \times Y)^n\)-valued random variables \((Z_{i_1}, \ldots, Z_{i_n})\) and \((Z_{i_1+i}, \ldots, Z_{i_n+i})\) have the same distribution for all \(n, i, i_1, \ldots, i_n \geq 1\). We further write \(P\) for the distribution of one (and thus all) \(Z_i\), i.e., for all measurable \(A \subset X \times Y\), we have
\[
P(A) = \mu(\{\omega \in \Omega : Z_i(\omega) \in A\}).
\]

To estimate the correlation between the \(\sigma\)-algebras \(\mathcal{A}_1^i\) and \(\mathcal{A}_1^{\infty} \):, various mixing coefficients have been proposed and used in the literature:
\[
\alpha(Z, n) := \sup_{A \in \mathcal{A}_1^i, B \in \mathcal{A}_1^{\infty}} \left| \mu(A \cap B) - \mu(A)\mu(B) \right|,
\]
\[
\beta(Z, n) := \mathbb{E} \sup_{B \in \mathcal{A}_1^{\infty}} \left| \mu(B) - \mu(B|\mathcal{A}_1^i) \right|,
\]
\[
\phi(Z, n) := \sup_{A \in \mathcal{A}_1^i, B \in \mathcal{A}_1^{\infty}} \left| \mu(B) - \mu(B|A) \right|.
\]

Recall that the \(\alpha\)-coefficient was introduced by Rosenblatt [24], while the \(\beta\)-mixing coefficient, introduced by Kolmogorov, first appeared in the paper [38]. Finally, Ibragimov [12] introduced the \(\phi\)-coefficient, see also [13].

It is well-known, see e.g. [3, Vol. 1, p. 186], that these coefficients satisfy
\[
2\alpha(Z, n) \leq \beta(Z, n) \leq \phi(Z, n),
\]
and since smaller coefficients reflect weaker dependence, the \(\alpha\)-mixing coefficients are the weakest notion for describing the dependence structure.
Definition 2.1. A stochastic process $Z = (Z_i)_{i \geq 1}$ is called $\alpha$-mixing, if the $\alpha$-mixing coefficients satisfy
\[ \lim_{n \to \infty} \alpha(Z, n) = 0. \]
Similarly one can define $\beta$- and $\phi$-mixing sequences.

By (2) we know that the $\beta$- and $\phi$-mixing sequences are also $\alpha$-mixing, so we will focus here only on the $\alpha$-mixing processes. Moreover, throughout this work, we assume that the process $Z$ is geometrically $\alpha$-mixing, that is
\[ \alpha(Z, n) \leq c \exp(-bn^\gamma), \quad n \geq 1, \tag{3} \]
for some constants $b > 0$, $c \geq 0$, and $\gamma > 0$. Of course, i.i.d. processes satisfy (3) for $c = 0$ and all $b, \gamma > 0$. Moreover, several time series models such as ARMA and GARCH, which are often used to describe, e.g. financial data, satisfy (3) under natural conditions [8, Chapter 2.6.1], and the same is true for many Markov chains including some dynamical systems perturbed by dynamic noise, see e.g. [37, Chapter 3.5]. An extensive and thorough account on mixing concepts including $\beta$- and $\phi$-mixing is provided by [3].

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 $(x, y) \in X \times Y$ the value $f(x)$ is a good prediction of $y$ at $x$. The following definition will help us to define what we mean by “good”.

Definition 2.2. Let $(X, \mathcal{X})$ be a measurable space and $Y \subset \mathbb{R}$ be a closed subset. Then a function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called a loss function, or simply a loss, if it is measurable.

Following [28, Definition 2.22], we only investigate loss functions that in some sense can be restricted to domains of the form $X \times Y \times [-M, M]$.

Definition 2.3. We say that a loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ can be clipped at $M > 0$, if, for all $(x, y, t) \in X \times Y \times \mathbb{R}$, we have
\[ L(x, y, \hat{t}) \leq L(x, y, t), \tag{4} \]
where $\hat{t}$ denotes the clipped value of $t$ at $\pm M$, that is
\[ \hat{t} := \begin{cases} -M & \text{if } t < -M, \\ t & \text{if } t \in [-M, M], \\ M & \text{if } t > M. \end{cases} \]

Throughout this work we assume that $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is a loss that can be clipped at some $M > 0$. Additionally, we assume that $L$ is both bounded in the sense of $L(x, y, t) \leq 1$ and Lipschitz continuous, that is,
\[ |L(x, y, t) - L(x, y, t')| \leq |t - t'|. \tag{5} \]
Here both inequalities are supposed to hold for all \((x, y) \in X \times Y\) and \(t, t' \in [-M, M]\). Note that the former assumption can be enforced by scaling.

To illustrate the generality of the made assumptions on \(L\), 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 \(1_{(-\infty,0]}(y \text{sign } t)\), since the latter may lead to computational infeasible approaches. Typical surrogates \(L\) belong to the class of margin-based losses, that is, \(L\) is of the form \(L(y, t) = \varphi(yt)\), where \(\varphi : \mathbb{R} \to [0, \infty)\) is a suitable, convex function. Then \(L\) can be clipped, if and only if \(\varphi\) has a global minimum, see [28, 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, [26] 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.

Bounded regression is another class of learning problems, where the assumptions made on \(L\) are often satisfied. Indeed, if \(Y := [-M, M]\) and \(L\) is a convex, distance-based loss represented by some \(\psi : \mathbb{R} \to [0, \infty)\), that is \(L(y, t) = \psi(y - t)\), then \(L\) can be clipped whenever \(\psi(0) = 0\), see again [28, Lemma 2.23]. In particular, the least squares loss

\[
L(y, t) = (y - t)^2
\]

and the \(\tau\)-pinball loss

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

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

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

**Definition 2.4.** Let \(L : X \times Y \times \mathbb{R} \to [0, \infty)\) be a loss function and \(P\) be a probability measure on \(X \times Y\). Then, for a measurable function \(f : X \to \mathbb{R}\) the \(L\)-risk is defined by

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

Moreover, the minimal \(L\)-risk

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

Informally, the goal of learning from a training set \( D \in (X \times Y)^n \) is to find a decision function \( f_D \) such that \( R_{L,P}(f_D) \) is close to the minimal risk \( R_{L,P}^* \). Our next goal is to formalize this idea. We begin with the following definition.

**Definition 2.5.** Let \( X \) be a set and \( Y \subset \mathbb{R} \) be a closed subset. A learning method \( L \) on \( X \times Y \) maps every set \( D \in (X \times Y)^n, n \geq 1 \), to a function \( f_D : X \to \mathbb{R} \).

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

\[
R_{L,P}(f_D) \to R_{L,P}^*, \quad n \to \infty .
\]

If this convergence takes place for all \( P \), then the learning method is called universally consistent. In the i.i.d. case many learning methods are known to be universally consistent, see e.g. [6] for classification methods, [11] for regression methods, and [28] for generic SVMs. For consistent methods, it is natural to ask how fast the convergence above, is. Unfortunately, in most situations uniform convergence rates are impossible, see [6, 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. In the non-i.i.d.-case, [22] showed that no uniform consistency is possible if one only assumes that the data generating process \( Z \) is stationary and ergodic. On the other hand, if some further assumptions of the dependence structure of \( Z \) are made, then consistency is possible, see e.g. [31]. The most widely made assumptions in this direction are in terms of the above mentioned mixing coefficients, but these are by no means necessary. Indeed, [27] establishes consistency for SVMs under the assumption that the data is generated by certain dynamical systems. Finally, learning rates are possible for data generated by some mixing process, if one makes additional assumptions on \( P \), see, e.g. the references given in the introduction.

Let us now describe the learning algorithms we are interested in. To this end, we assume that we have a hypothesis set \( F \) consisting of bounded measurable functions \( f : X \to \mathbb{R} \), which is pre-compact with respect to the supremum norm \( \| \cdot \|_\infty \). Since \( F \) can be infinite, we need to recall the following concept, which will enable us to approximate infinite \( F \) by finite subsets.

**Definition 2.6.** Let \((T,d)\) be a metric space and \( \varepsilon > 0 \). We call \( S \subset T \) an \( \varepsilon \)-net of \( T \) if for all \( t \in T \) there exists an \( s \in S \) with \( d(s,t) \leq \varepsilon \). Moreover, the \( \varepsilon \)-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 \( \varepsilon \).
Note that our hypothesis set $\mathcal{F}$ is assumed to be pre-compact, and hence for all $\varepsilon > 0$, the covering number $N(\mathcal{F}, \| \cdot \|_\infty, \varepsilon)$ is finite.

In order to introduce our generic learning algorithms, we write

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

for a training set of length $n$ that is distributed according to the first $n$ components of $Z$. Furthermore, we write $D_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i,Y_i)}$, where $\delta_{(X_i,Y_i)}$ denotes the (random) Dirac measure at $(X_i,Y_i)$. In other words, $D_n$ is the empirical measure associated to the data set $D_n$. Finally, the risk of a function $f : X \to \mathbb{R}$ with respect to this measure

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

is called the empirical $L$-risk.

With these preparations we can now introduce the class of learning methods we are interested in:

**Definition 2.7.** 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 $\Upsilon$ 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

$$\Upsilon(f_{D_n,\Upsilon}) + R_{L,D_n}(f_{D_n,\Upsilon}) \leq \inf_{f \in \mathcal{F}} (\Upsilon(f) + R_{L,D_n}(f)) + \delta$$

(8)

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

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

Note that on the right-hand side of (8) the unclipped loss is considered, and hence CR-ERMs do not necessarily minimize the regularized clipped empirical risk $\Upsilon(\cdot) + R_{L,D_n}(\cdot)$. Moreover, in general CR-ERMs do not minimize the regularized risk $\Upsilon(\cdot) + R_{L,D_n}(\cdot)$ either, because on the left-hand side of (8) the clipped function is considered. However, if we have a minimizer of the unclipped regularized risk, then it automatically satisfies (8). In particular, ERM decision functions satisfy (8) for the regularizer $\Upsilon := 0$ and $\delta := 0$, and SVM decision functions satisfy (8) for the regularizer $\Upsilon := \lambda \| \cdot \|^2_H$ and $\delta := 0$. In other words, ERM and SVMs are CR-ERMs.

## 3 Main Results

In this section we present our main result, a refined oracle inequality for learning from geometrically $\alpha$-mixing processes. Before we present this inequality, we need to introduce a few more notations. Let $\mathcal{F}$ be a hypothesis set in the sense of Definition 2.7. For

$$r^* := \inf_{f \in \mathcal{F}} \Upsilon(f) + R_{L,P}(f) - R_{L,P}^*$$

(9)
and $r > r^*$, we write
\[ \mathcal{F}_r := \{ f \in \mathcal{F} : \Upsilon(f) + \mathcal{R}_{L,P}(\hat{f}) - \mathcal{R}_{L,P}^* \leq r \}. \] (10)

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
\[ \ln \mathcal{N}(\mathcal{F}_r, \| \cdot \|_\infty, \varepsilon) \leq \varphi(\varepsilon)r^p \] (11)
for all $\varepsilon > 0, r > 0$ and a suitable constant $p \in (0, 1]$. Note that there are actually many hypothesis sets satisfying Assumption (11), see Section 4 for some examples.

Now our main result reads as follows:

**Theorem 3.1.** Let $Z := (Z_i)_{i \geq 1}$ be an $X \times Y$-valued process that satisfies (3), $\alpha := \frac{\gamma + 1}{\gamma + 1}$ with $\gamma$ as in (3), and $P$ be defined by (1). Furthermore, let $L$ be a loss that can be clipped at some $M > 0$ and that is bounded and Lipschitz continuous in the sense of Section 2. Assume that there exists a Bayes decision function $f^*_L,P$ and constants $\vartheta \in [0, 1]$ and $V \geq 1$ such that
\[ \mathbb{E}_P(L \circ \hat{f} - L \circ f^*_L,P) \leq V \cdot \mathbb{E}_P(L \circ \hat{f} - L \circ f^*_L,P)^\vartheta, \quad f \in \mathcal{F}, \] (12)
where $\mathcal{F}$ is a hypothesis set with $0 \in \mathcal{F}$. We define $r^*$ and $\mathcal{F}_r$ by (9) and (10), respectively and assume that (11) 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 $\varepsilon > 0, \delta \geq 0, \tau \geq 1, n \geq \max \{ b/8, 2^{1+5/\gamma}b^{-1/\gamma} \}$, and $r \in (0, 1]$ satisfying
\[ r \geq \max \left\{ \left( \frac{c_V(\tau + \varphi(\varepsilon/2)2p^p)}{n^\alpha}, 8c_B B_0 \tau \frac{1}{n^\alpha}, r^* \right) \right\}, \] (13)
every learning method defined by (8) satisfies with probability $\mu$ not less than $1 - 8Ce^{-\tau}$:
\[ \Upsilon(f_{D_n,Y}) + \mathcal{R}_{L,P}(\hat{f}_{D_n,Y}) - \mathcal{R}_{L,P}^* < 2\Upsilon(f_0) + 4\mathcal{R}_{L,P}(f_0) - 4\mathcal{R}_{L,P}^* + 4r + 5\varepsilon + 2\delta. \] (14)

Here the constants are
\[ C := 1 + 4e^{-2c}, \]
\[ c_B := c_\sigma/3, \]
\[ c_V := \max \{ c_\sigma V, 2^{8+2/\gamma + 1}b^{-1/\tau}(12V + 1)/3 \}, \]
where $c_\sigma := (\frac{2^2 + 2}{b})^{1/1+\gamma}$.

Before we illustrate this theorem in the next section with the help of a few examples, let us briefly discuss the variance bound (12). For example, if $Y = [-M, M]$ and $L$
is the least squares loss, then it is well-known that (12) is satisfied for \( V := 16M^2 \) and \( \vartheta = 1 \), see e.g. [28, Example 7.3]. Moreover, under some assumptions on the distribution \( P \), [30] established a variance bound of the form (12) for the so-called pinball loss used for quantile regression. In addition, for the hinge loss, (12) is satisfied for \( \vartheta := q/(q + 1) \), if Tsybakov’s noise assumption [35, Proposition 1] holds for \( q \), see [28, Theorem 8.24]. Finally, based on [2], [26] established a variance bound with \( \vartheta = 1 \) for the earlier mentioned clippable modifications of strictly convex, twice continuously differentiable margin-based loss functions.

One might wonder, why the constant \( B_0 \) is necessary in Theorem 3.1, since apparently it only adds further complexity. However, a closer look reveals that the assumed boundedness of \( L \) only guarantees \( \|L \circ \hat{f}\|_\infty \leq 1 \), 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 \( \hat{f} = f \), we believe that in general \( B_0 \) is necessary. We refer to Examples 4.2, 4.3 and 4.4 for situations, where \( B_0 \) is significantly larger than 1.

Finally, we like to point out that we do not know, whether the used Bernstein inequality for \( \alpha \)-mixing processes, see Theorem 5.1, is optimal. However, if there is, or will be, better versions of this inequality (with or without some additional assumptions on \( \mathcal{Z} \)), our oracle inequalities can be easily improved, since our techniques are modular in the sense that they only require a generic form of Bernstein’s inequality.

4 Applications

To illustrate the oracle inequality from Section 3, we will now use it to establish learning rates for some algorithms including ERM over finite sets and SVMs using either a given generic kernel or a Gaussian kernel with varying widths. In the ERM case, our results match those in the i.i.d. case, if one replaces the number of observations with the “effective number of observations” \( n^\alpha \) while, for LS-SVMs with given generic kernels, our rates are slightly worse than the recently obtained optimal rates [32] for i.i.d. observations. The latter difference is not surprising, when considering the fact that [32] 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 Bernstein’s inequality and the peeling method. However, when using Gaussian kernels, we actually recover the optimal rates for i.i.d. observations for LS-SVMs and SVMs for quantile regression.

Let us now present the first example, that is empirical risk minimization over a finite set.

Example 4.1 (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 \( \|f\|_\infty \leq M \) for all \( f \in \mathcal{F} \). Then, for accuracy \( \delta := 0 \), the learning method described by (8) is ERM, and Theorem 3.1 shows by some simple estimates that

\[
\mathcal{R}_{L,P}(f_{D_n,T}) - \mathcal{R}^*_L \leq 4 \inf_{f \in \mathcal{F}} \left( \mathcal{R}_{L,P}(f) - \mathcal{R}^*_L \right) + 4 \left( \frac{c_V(T + \ln |\mathcal{F}|)}{n^\alpha} \right)^{1/(2-\vartheta)} + 32c_B \tau \frac{\mathcal{R}^*_L}{n^\alpha}.
\]
hold with probability \( \mu \) not less than \( 1 - 8Ce^{-T} \). Note that in the i.i.d. case we have \( \alpha = 1 \). Besides constants, the oracle inequality (14) is thus an exact analogue to standard oracle inequality for ERM learning from i.i.d. processes, see e.g. [28, Theorem 7.2].

For further examples let us begin by briefly recalling SVMs, see [28] for details. To this end, let \( X \) be a measurable space, \( Y := [-1, 1] \) and \( k \) be a measurable (reproducing) kernel on \( X \) with reproducing kernel Hilbert space (RKHS) \( H \). Given a regularization parameter \( \lambda > 0 \) and a convex loss \( L \), SVMs find the unique solution

\[
 f_{D_n, \lambda} = \arg \min_{f \in H} (\lambda\|f\|_H^2 + R_{L, D_n}(f)).
\]

In particular, SVMs using the least-squares loss (6) are called least-squares SVMs (LS-SVMs), while SVMs using the \( \tau \)-pinball loss (7) are called SVMs for quantile regression.

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

\[
 A(\lambda) := \inf_{f \in H} (\lambda\|f\|_H^2 + R_{L, P}(f) - R_{L, P}^\ast), \quad \lambda > 0.
\]

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

**Example 4.2 (Generic Kernels).** Let \((X, \mathcal{X})\) be a measurable space, \( Y = [-1, 1] \), and \( Z \) 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

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

for some constants \( p \in (0, 1] \) and \( c > 0 \). Since \( F = \lambda^{-1/2}B_H \) and \( \mathcal{F}_r \subset r^{1/2}\lambda^{-1/2}B_H \), we find

\[
 \ln \mathcal{N}(\mathcal{F}_r, \| \cdot \|_\infty, \varepsilon) \leq a\lambda^{-p}\varepsilon^{-2pr^p}
\]

and thus we can define \( \varphi(\varepsilon) := a\lambda^{-p}\varepsilon^{-2pr^p} \). For the least square loss the variance bound (12) is valid with \( \vartheta = 1 \), hence the condition (13) is satisfied if

\[
 r \geq \max \left\{ \left( cV^2 + 3p \right) \frac{1}{2p} \lambda^{-\frac{p}{1-p}} n^{-\frac{\alpha}{1-\alpha} p - \frac{2p}{1-2p}} \cdot \frac{2cV \tau}{n^{\alpha}} \cdot \frac{8cB_0 \tau}{n^{\alpha}}, r^\ast \right\}.
\]

Therefore, let \( r \) be the sum of the terms on the right-hand side. In addition, assume that the approximation error function satisfies \( A(\lambda) \leq c\lambda^\beta \) for some \( c > 0 \), \( \beta \in (0, 1] \), and all \( \lambda > 0 \). Since for large \( n \) the first and next-to-last term in (18) dominate, the oracle inequality (14) becomes

\[
 \lambda\|f_{D_n, \lambda}\|_H^2 + R_{L, P}(f_{D_n, \lambda}) - R_{L, P}^\ast \leq 4\lambda\|f_{P, \lambda}\|_H^2 + 4R_{L, P}(f_{P, \lambda}) - 4R_{L, P}^\ast + 4r + 5\varepsilon \leq C \left( \lambda^\beta + \lambda^{-\frac{p}{1-2p}} n^{-\frac{\alpha}{1-\alpha} p - \frac{2p}{1-2p}} \cdot \frac{2cV \tau}{n^{\alpha}} \cdot \frac{8cB_0 \tau}{n^{\alpha}} + \lambda^\beta - n^{-\alpha} + \varepsilon \right),
\]

\( \ldots \)
where \( f_{P, \lambda} \) is the function at which the infimum in (15) is attained and \( C \) is a constant independent of \( n, \lambda, \tau, \) and \( \varepsilon \). Now optimizing over \( \varepsilon \), we then see by [28, Lemma A.1.7] that the LS-SVM using \( \lambda_n := n^{-\alpha \rho / \beta} \) learns with rate \( n^{-\alpha \rho} \), where

\[
\rho := \min \left\{ \beta, \frac{\beta}{\beta + p \beta + p} \right\}.
\]  

(19)

Let us now compare our rate \( n^{-\alpha \rho} \), where \( \rho \) is defined in (19), with the rates previously established for LS-SVMs in the literature. For example, [29] proved a rate of the form

\[
n^{-\alpha \min(\beta, \beta + 2p \beta + p)}
\]

under exactly the same assumptions. Since \( \beta > 0 \) and \( p > 0 \), our new rate is always better than that of [29]. In addition, [9] generalized the rates of [29] to regularization terms of the form \( \lambda\| \cdot \|_q^H \) 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 \( q = 2 \); [39] established the rate

\[
n^{-\frac{\beta}{2(\beta + 1)}}
\]

which is always slower than ours, too. Finally, in the case \( p = 1 \), [33] established the rate \( n^{-\frac{\beta}{2(\beta + 1)}} \), which was subsequently improved to \( n^{-\frac{\beta}{2(\beta + 1)}} \) in [34]. 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 [33, 34] is restricted to LS-SVMs, while our results hold for rather generic learning algorithms.

**Example 4.3 (Smooth Kernels).** Let \( X \subset \mathbb{R}^d \) be a compact subset, \( Y = [-1, 1] \), and \( Z \) 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. [32] or [28, Theorem 6.26], that

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

(20)

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 \( \infty \). Then there exists a constant \( C_p > 0 \) such that

\[
\| f \|_\infty \leq C_p \| f \|_H^p \| f \|_{L^2(P_X)}^{1-p}, \quad f \in H
\]

for the same \( p \), see [17] and [32, Corollary 3]. Consequently, we can bound \( B_0 \leq \lambda^{(\beta-1)p} \) as in [32]. Moreover, the assumption on the approximation error function is satisfied for \( \beta := s/m \), whenever \( f_L^m \in W^s(X) \) and \( s \in (0, m] \). Therefore, the resulting learning rate is

\[
n^{-\frac{2s \alpha}{2s + 2s + m}}.
\]

11
Note that in the i.i.d. case, where \( \alpha = 1 \), this rate is worse than the optimal rate \( n^{-2/3+\sigma} \), 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 \( k \).

Let us finally compare this rate with the rate from [34]. Here it turns out that their rate is worse than ours, if \( m \geq \frac{1}{p_0^2}(2s + 3d + \sqrt{4s^2 + 108sd + 9d^2}) \). Note that by the constraint \( s \leq m \), the latter is always satisfied for \( m \geq d \). \( \blacksquare \)

In the next example for LS-SVMs we will consider the Gaussian RBF kernels \( k_\gamma \) on \( X \), which are defined by

\[
F_\gamma(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\gamma^2}\right), \quad x, x' \in X,
\]

for some width \( \gamma \in (0, 1] \). We write \( H_\gamma \) for the RKHS of \( k_\gamma \) and \( B_{H_\gamma} \) for its closed unit ball. Similarly, \( B_{\mathcal{B}_2^d} \) denotes the closed unit ball of \( d \)-dimensional Euclidean space \( \ell_2^d \). Finally, for \( t \in \mathbb{R} \), \( \lfloor t \rfloor \) is the largest integer \( n \) satisfying \( n \leq t \), and similarly, \( \lceil t \rceil \) is the smallest integer \( n \) satisfying \( n \geq t \).

**Example 4.4 (Gaussian Kernels)**. Let \( Y := [-M, M] \) for \( M > 0 \), and \( P \) be a distribution on \( \mathbb{R}^d \times Y \) such that \( X := \text{supp}P_X \subset B_{\mathcal{B}_2^d} \) is a bounded domain with \( \mu(\partial X) = 0 \). Furthermore, let \( P_X \) be absolutely continuous w.r.t. the Lebesgue measure \( \mu \) 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 \( f_{L,p}^* \in B_{2q,\infty}^d \) for \( t \geq 1 \) and \( s \geq 1 \) with \( 1 + 1/s = 1.0 \). Here, \( B_{2q,\infty}^d \) denotes the usual Besov space with the smoothness parameter \( t \). Furthermore, the entropy numbers for Gaussian kernels [28, Theorem 7.34] and the equivalence of covering and entropy numbers [28, Lemma 6.21] yield a constant \( a_{p,\eta} > 0 \) such that

\[
\ln \mathcal{N}(B_{H_\gamma}, \| \cdot \|_\infty, \varepsilon) \leq a_{p,\eta} \gamma^{-(1+p)(1+\eta)d\varepsilon^{-2p}}, \quad \varepsilon > 0,
\]

for all \( \eta > 0 \) and \( p \in (0, 1) \). Similarly as above, since \( F = \lambda^{-1/2} B_{H_\gamma} \) and \( \mathcal{F}_r \subset r^{1/2} \lambda^{-1/2} B_{H_\gamma} \), we have

\[
\ln \mathcal{N}((\mathcal{F}_r, \| \cdot \|_\infty, \varepsilon) \leq a_{p,\eta} \gamma^{-(1+p)(1+\eta)d\lambda^{-p}\varepsilon^{-2p}},
\]

and thus we can define

\[
\varphi(\varepsilon) := a_{p,\eta} \gamma^{-(1+p)(1+\eta)d\lambda^{-p}\varepsilon^{-2p}}.
\]

Since the variance bound (12) is valid with \( \vartheta = 1 \) for the least-square loss, the condition (13) is satisfied if

\[
r \geq \max\left\{(CV_2^{1+3p}a_{p,\eta})^{1-r/p} \gamma^{-(1+\eta)d \lambda^{-p} n^{-1/p} - \alpha} \varepsilon^{-1/p}, \frac{2cv_2^2}{n^\alpha}, \frac{8c_B B_0 \tau}{n^\alpha}, r^*\right\}.
\]
Moreover, [7, Section 2] shows that there exists a constant $c > 0$ such that for all $\lambda > 0$ and all $\gamma \in (0, 1]$, there is an $f_0 \in H_\gamma$ with $\|f_0\|_\infty \leq c$ and
\[
\lambda \|f_0\|^2_{H_\gamma} + R_{L, P}(f_0) - R^*_L, P \leq c\lambda \gamma^{-d} + c\gamma^{2t}.
\]
Since for large $n$ the first term in (24) dominates, the oracle inequality (14) becomes
\[
\lambda \|f_{D_n, \lambda}\|^2_{H_\gamma} + R_{L, P}(f_{D_n, \lambda}) - R^*_L, P \leq 4\lambda \|f_0\|^2_{H_\gamma} + 4R_{L, P}(f_0) - 4R^*_L, P + 4\varepsilon \leq C(\lambda \gamma^{-d} + \gamma^{2t} + \gamma^{-(1+n)d} \lambda^{-\frac{d}{n+2}} - \frac{\varepsilon}{n+2} - \frac{\varepsilon}{n+2} + \varepsilon),
\]
where $C$ is a constant independent of $n$, $\lambda$, $\gamma$, $\tau$, and $\varepsilon$. Again, optimizing over $\varepsilon$ together with some standard techniques, see [28, Lemma A.1.7], we then see that for all $\xi > 0$ we can find $\eta, p \in (0, 1)$ sufficiently close to 0 such that the LS-SVM using Gaussian RKHS $H_\gamma$ and $\lambda_n = n^{-\alpha}$ and $\gamma_n = n^{-\frac{d}{n+2}}$, (25)
learns with rate
\[
n^{-\frac{d}{n+2} + \eta + \xi}.
\]
In the i.i.d. case we have $\alpha = 1$, and hence the learning rate (26) becomes
\[
n^{-\frac{d}{n+2} + \xi}.
\]
Recall that modulo the arbitrarily small $\xi > 0$ these learning rates are essentially optimal, see e.g. [32, Theorem 13] or [11, Theorem 3.2].
To achieve these rates, however, we need to set $\lambda_n$ and $\gamma_n$ as in (25), which in turn requires us to know $t$ and $\alpha$. 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. [28, Chapters 6.5, 7.4, 8.2], to achieve the same rates adaptively, i.e. without knowing $t$ and $\alpha$. To this end, let $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be sequences of finite subsets $\Lambda_n, \Gamma_n \subset (0, 1]$ such that $\Lambda_n$ is an $\epsilon_n$-net of $(0, 1]$ and $\Gamma_n$ is an $\delta_n$-net of $(0, 1]$ with $\epsilon_n \leq n^{-1}$ and $\delta_n \leq n^{-\frac{d}{n+2}}$. Furthermore, assume that the cardinalities $|\Lambda_n|$ and $|\Gamma_n|$ grow polynomially in $n$. For a data set $D := (x_1, y_1, \ldots, x_n, y_n)$, we define
\[
D_1 := ((x_1, y_1), \ldots, (x_m, y_m))
\]
\[
D_2 := ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n))
\]
where $m := \left\lfloor \frac{n}{2} \right\rfloor + 1$ and $n \geq 4$. We will use $D_1$ as a training set by computing the SVM decision functions
\[
f_{D_1, \lambda, \gamma} := \arg \min_{f \in H_\gamma} \lambda \|f\|^2_{H_\gamma} + R_{L, D_1}(f), \quad (\lambda, \gamma) \in \Lambda_n \times \Gamma_n
\]
and use $D_2$ to determine $(\lambda, \gamma)$ by choosing a $(\lambda_{D_2}, \gamma_{D_2}) \in \Lambda_n \times \Gamma_n$ such that
\[
\mathcal{R}_{L,D_2} \left( \int_{D_2} \lambda_{D_2}, \gamma_{D_2} \right) = \min_{(\lambda, \gamma) \in \Lambda_n \times \Gamma_n} \mathcal{R}_{L,D_2} \left( \int_{D_2} \lambda, \gamma \right) .
\]

Then, analogous to the proof of Theorem 3.3 in [7] we can show that for all $\xi > 0$, the TV-SVM producing the decision functions $f_{D_2, \lambda_{D_2}, \gamma_{D_2}}$ with the above learning rates (26).

In the last example we will briefly discuss learning rates for SVMs for quantile regression. For more information on such SVMs we refer to [7, Section 4].

**Example 4.5 (Quantile Regression with Gaussian Kernels).** Let $Y := [-1,1]$ and $P$ be a distribution on $\mathbb{R}^d \times Y$ such that $X := \text{supp} P_{X} \subset B_{d}^{\infty}$ be a domain. Furthermore, we assume that, for $P_{X}$-almost all $x \in X$, the conditional density $h(\cdot, x)$ of $P(\cdot|x)$ is bounded from 0 and $\infty$, see also [7, Example 4.5]. Moreover, let $P_{X}$ be absolutely continuous w.r.t. the Lebesgue measure on $X$ with associated density $g \in L_{u}(X)$ for some $u \geq 1$. For $\tau \in (0,1)$, let $f^{*}_{\tau,P} : \mathbb{R}^{d} \rightarrow \mathbb{R}$ be a conditional $\tau$-quantile function that satisfies $f^{*}_{\tau,P} \in L_{2}(\mathbb{R}^{d}) \cap L_{\infty}(\mathbb{R}^{d})$. In addition, we assume that $f^{*}_{\tau,P} \in B_{d,\infty}^{\infty}$ for some $t \geq 1$ and $s \geq 1$ such that $\frac{1}{t} + \frac{1}{s} = 1$.

Then [30, Theorem 2.8] yields a variance bound of the form
\[
\mathbb{E}_{P}(L_{\tau} \circ \hat{f} - L_{\tau} \circ f^{*}_{\tau,P})^{2} \leq V \cdot \mathbb{E}_{P}(L_{\tau} \circ \hat{f} - L_{\tau} \circ f^{*}_{\tau,P}),
\]
for all $f : X \rightarrow \mathbb{R}$, where $V$ is a suitable constant and $L_{\tau}$ is the $\tau$-pinball loss. For the covering numbers of the Gaussian kernels (21) we define $\varphi(\varepsilon)$ as in (23). Consequently, Condition (13) is satisfied if
\[
r \geq \max \left\{ \left( c_{V} \frac{1 + 3p}{\alpha} a_{p,q} \right)^{\frac{1}{\gamma - 1 + \eta d}} \gamma^{-1 + \eta d} \lambda^{-\frac{p}{2\gamma - 1}} n^{-\frac{\alpha}{4\gamma - 2}} \varepsilon^{-\frac{2p}{4\gamma - 2}}, \frac{2c_{V} \tau}{\eta n}, \frac{8c_{B} \beta_{0} \tau}{n^{\alpha}}, p^{*} \right\} .
\]
Moreover, [7, Section 4] shows that there exists a constant $c > 0$ such that for all $\lambda > 0$ and all $\gamma \in (0,1]$, there is an $f_{0} \in H_{\gamma}$ with $\| f_{0} \|_{\infty} \leq c$ and
\[
\lambda \| f_{0} \|_{H_{\gamma}}^{2} + \mathcal{R}_{L_{\tau},P}(f_{0}) - \mathcal{R}_{L_{\tau},P}^{*} \leq c \lambda \gamma^{-d} + c \gamma^{2t} .
\]
For large $n$ the first term in (27) thus dominates, and hence the oracle inequality (14) becomes
\[
\lambda \| f_{D_{n},\lambda} \|_{H_{\gamma}}^{2} + \mathcal{R}_{L_{\tau},P}(f_{D_{n},\lambda}) - \mathcal{R}_{L_{\tau},P}^{*} \leq 4 \lambda \| f_{0} \|_{H_{\gamma}}^{2} + 4 \mathcal{R}_{L_{\tau},P}(f_{0}) - 4 \mathcal{R}_{L_{\tau},P}^{*} + 4r + 5 \varepsilon
\leq C( \lambda \gamma^{-d} + 2 \gamma + \gamma^{-1 + \eta d} \lambda^{-\frac{p}{2\gamma - 1}} n^{-\frac{\alpha}{4\gamma - 2}} \varepsilon^{-\frac{2p}{4\gamma - 2}}(\gamma + r) + \varepsilon),
\]
where $C$ is a constant independent of $n$, $\lambda$, $\gamma$, $\tau$, and $\varepsilon$. Again, for every $\xi > 0$ we can then find $\eta, p \in (0,1)$ sufficiently close to 0 such that the SVM for quantile regression with
\[
\lambda_{n} = n^{-\alpha} \quad \text{and} \quad \gamma_{n} = n^{-\frac{2\alpha}{2\gamma - 2}},
\]

14
learns with rate
\[ n^{-\frac{2\alpha}{\gamma+1}}. \]  
(29)

Note that this rate is for the excess \( L_\tau \)-risk, but since \([30, \text{Theorem 2.7}]\) shows
\[ \| \hat{f} - f^*_\tau, P \|_{L_2(P_X)}^2 \leq c(R_{L_\tau, P}(\hat{f}) - R^*_L, P) \]
for some constant \( c > 0 \) and all \( f : X \rightarrow \mathbb{R} \), we actually obtain the same rates for
\[ \| \hat{U}_f - f^*_\tau, P \|_{L_2(P_X)}^2. \] Last but not least, optimality and adaptivity can be discussed along
the lines of LS-SVMs.

5 Proofs

The first key result we need for the proof of Theorem 3.1 is the following Bernstein
type inequality for geometrically \( \alpha \)-mixing processes, which was established in \([18, \text{Theorem 4.3}]\):

**Theorem 5.1.** Let \( Z := (Z_t)_{t \geq 1} \) be an \( X \times Y \)-valued stochastic process that satisfies
(3) and \( P \) be defined by (1). Furthermore, let \( h : X \times Y \rightarrow \mathbb{R} \) be a bounded measurable
function for which there exist constants \( B > 0 \) and \( \sigma \geq 0 \) such that
\[ \mathbb{E} P h = 0, \quad \mathbb{E} P h^2 \leq \sigma^2, \quad \| h \|_{\infty} \leq B. \] For \( n \geq 1 \), we define
\[ n^{(\gamma)} := \left[ n \left( \frac{8n}{b} \right)^{\frac{\gamma}{\gamma+1}} \right]^{\frac{1}{\gamma+1}}. \]

Then, for all \( n \geq 1 \) and all \( \epsilon > 0 \), we have
\[ \mu \left( \left\{ \omega \in \Omega : \frac{1}{n} \sum_{i=1}^{n} h(Z_i(\omega)) \geq \epsilon \right\} \right) \leq (1 + 4e^{-2c}) \exp \left( \frac{-3\epsilon^2 n^{(\gamma)}}{6\sigma^2 + 2\epsilon B} \right). \]  
(30)

Before we prove Theorem 3.1, we need to slightly modify (30). To this end, we first
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
\[ n^{(\gamma)} \geq 2 \left( \frac{2b + 5}{b^{1/\gamma}} \right)^{n^{\alpha}}, \]
where \( \alpha := \frac{\gamma}{\gamma+1} \). For \( C := 1 + 4e^{-2c} \), \( c_\sigma := (\frac{b^{2+5/\gamma}}{b})^{1/1+\gamma} \) and \( c_B := c_\sigma/3 \), we thus obtain
\[ \mu \left( \left\{ \omega \in \Omega : \frac{1}{n} \sum_{i=1}^{n} h(Z_i(\omega)) \geq \epsilon \right\} \right) \leq C e^{-\tau}, \quad n \geq n_0, \]
where \( \tau := \frac{c^2\sigma^2}{c_\sigma^2 + 3c_B}. \) Simple transformations and estimations then yield
\[ \mu \left( \left\{ \omega \in \Omega : \frac{1}{n} \sum_{i=1}^{n} h(Z_i(\omega)) \geq \sqrt{\frac{\tau c_\sigma^2}{n^{\alpha}} + \frac{c_B^2}{n^{\alpha}}} \right\} \right) \leq C e^{-\tau} \]  
(31)
for all $n \geq \max\{b/8, 2^{2+5/\gamma}b^{-1/\gamma}\}$ and $r > 0$.

For our proof we also need the so-called peeling method (see for example [36, Chapter 5.3]). 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

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

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

$$\mathcal{G}_k = \{g \in \mathcal{G} : m_k - 1 \leq \Gamma(g) < m_k\}, \quad k = 1, \ldots, K + 2.$$

Now we can formulate the peeling as following:

**Theorem 5.2.** Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $Z = (Z_g)_{g \in \mathcal{G}}$ be a stochastic process indexed by $\mathcal{G}$, we have for all $\epsilon \geq 0$ and $r > 0$,

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

**Proof.** With the peeling (32) we obtain

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

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

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

\[\square\]

In addition, we will need the following simple and well-known lemma (see e.g. [28, Lemma 7.1]):

**Lemma 5.3.** 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 $(qa)^{2/q}(qb)^{2/q'} \leq (a+b)^2$ and $ab \leq a^q/b^{q'}$.

Since the proof of Theorem 3.1 is rather complicated, we first describe its main steps briefly: First we decompose the regularized excess risk into an approximation error term and two stochastic error terms. The approximation error and the first stochastic error term can be estimated by standard techniques. Similarly, the first step in the
estimation of the second error term is a rather standard quotient approach, see e.g. [28, 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 $F$ into disjoint “spheres”. For each sphere, we then use localized covering numbers and Bernstein’s inequality from Theorem 5.1, 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 [29].

**Proof of Theorem 3.1. Main Decomposition.** For $f : X \rightarrow \mathbb{R}$ we define $h_f := L \circ f - L \circ f_L^\star$. By the definition of $f_D, \tau$, we then have

$$\Upsilon(f_D, \tau) + \mathbb{E}_{D_n} h_f^\tau \leq \Upsilon(f_0) + \mathbb{E}_{D_n} h_{f_0} + \delta,$$

and consequently we obtain

$$\Upsilon(f_D, \tau) + R_{L,P}(\hat{f}_D, \tau) - R_{L,P} = \Upsilon(f_0) + \mathbb{E}_{D_n} h_f - \mathbb{E}_{D_n} h_{f_D, \tau} + \mathbb{E} h_{\hat{f}_D, \tau} + \delta$$

$$= (\Upsilon(f_0) + \mathbb{E}_{h_{f_0}} + (\mathbb{E}_{D_n} h_f - \mathbb{E}_{h_{f_0}}) + (\mathbb{E} h_{\hat{f}_D, \tau} - \mathbb{E}_{D_n} h_{\hat{f}_D, \tau}) + \delta). \quad (34)$$

**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

$$\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_{P} h_{f_0} = (\mathbb{E}_{D_n} (h_{f_0} - h^\tau_{f_0}) - \mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0})) + (\mathbb{E}_{D_n} h^\tau_{f_0} - \mathbb{E}_{P} h^\tau_{f_0}). \quad (35)$$

Now $L \circ f_0 - L \circ f_0 \geq 0$ implies $h_{f_0} - h^\tau_{f_0} = L \circ f_0 - L \circ f_0 \in [0, B_0]$, and hence we obtain

$$\mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0}) - \mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0}) \leq \mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0}) \leq B_0 \mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0}).$$

Inequality (31) applied to $h := (h_{f_0} - h^\tau_{f_0}) - \mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0})$ thus shows that

$$\mathbb{E}_{D_n} (h_{f_0} - h^\tau_{f_0}) - \mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0}) \leq \sqrt{\tau c_\sigma B_0 \mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0}) h^\tau_{f_0}} + \frac{c_\sigma B_0 \tau}{n^{\alpha}} + \frac{\sqrt{\tau c_\sigma B_0 \mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0}) n^{\alpha}}}{n^{\alpha}}$$

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

$$\sqrt{n^{\alpha} \tau c_\sigma B_0 \mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0})} \leq \mathbb{E}_{P} (h_{f_0} - h^\tau_{f_0}) + n^{-\alpha} c_\sigma B_0 \tau / 4,$$
and consequently we have with probability \( \mu \) not less than \( 1 - Ce^{-\tau} \) that
\[
\mathbb{E}_{Dn}(h_{\hat{f}_o} - h_{\tilde{f}_o}) - \mathbb{E}_P(h_{\hat{f}_o} - h_{\tilde{f}_o}) \leq \mathbb{E}_P(h_{\hat{f}_o} - h_{\tilde{f}_o}) + \frac{7eB_0\tau}{4n^\alpha}.
\]  
(36)

In order to bound the remaining term in (35), that is \( \mathbb{E}_{Dn}h_{\tilde{f}_o} - \mathbb{E}_P h_{\tilde{f}_o} \), we first observe that (5) implies \( \|h_{\tilde{f}_o}\|_\infty \leq 1 \), and hence we have \( \|h_{\tilde{f}_o} - \mathbb{E}_P h_{\tilde{f}_o}\|_\infty \leq 2 \). Moreover, (12) yields
\[
\mathbb{E}_P(h_{\tilde{f}_o} - \mathbb{E}_P h_{\tilde{f}_o})^2 \leq \mathbb{E}_P h_{\tilde{f}_o}^2 \leq V(\mathbb{E}_P h_{\tilde{f}_o})^2.
\]

In addition, if \( \vartheta \in (0, 1] \), the second inequality in Lemma 5.3 implies for \( q := \frac{2}{1-\vartheta} \), 
\( q' := \frac{2}{\vartheta} \), \( a := (n^{-\alpha}c_\sigma 2^{-\vartheta} \vartheta^\vartheta V^\vartheta)^{1/2} \), and \( b := (2^{\vartheta+1})^{\vartheta/2} \), that
\[
\sqrt{\frac{c_\sigma V^\tau(\mathbb{E}_P h_{\tilde{f}_o})^\vartheta}{n^\alpha}} \leq 2 \left( \frac{c_\sigma 2^{-\vartheta} \vartheta^\vartheta V^\vartheta}{n^\alpha} \right)^{1/\vartheta} + \mathbb{E}_P h_{\tilde{f}_o} \leq \left( \frac{c_\sigma V^\tau}{n^\alpha} \right)^{1/\vartheta} + \mathbb{E}_P h_{\tilde{f}_o}.
\]

Since \( \mathbb{E}_P h_{\tilde{f}_o} \geq 0 \), this inequality also holds for \( \vartheta = 0 \), and hence (31) shows that we have
\[
\mathbb{E}_{Dn}h_{\tilde{f}_o} - \mathbb{E}_P h_{\tilde{f}_o} < \mathbb{E}_P h_{\tilde{f}_o} + \left( \frac{c_\sigma V^\tau}{n^\alpha} \right)^{1/\vartheta} + \frac{2cB\tau}{n^\alpha} + \frac{7eB_0\tau}{4n^\alpha},
\]
(37)

with probability \( \mu \) not less than \( 1 - Ce^{-\tau} \). By combining this estimate with (36) and (35), we now obtain that with probability \( \mu \) not less than \( 1 - 2Ce^{-\tau} \) we have
\[
\mathbb{E}_{Dn}h_{\tilde{f}_o} - \mathbb{E}_P h_{\tilde{f}_o} < \mathbb{E}_P h_{\tilde{f}_o} + \left( \frac{c_\sigma V^\tau}{n^\alpha} \right)^{1/\vartheta} + \frac{2cB\tau}{n^\alpha} + \frac{7eB_0\tau}{4n^\alpha},
\]
(38)

since \( 1 \leq B_0, \) i.e., we have established a bound on the second term in (34).

**Estimating the Second Stochastic Term.** For the third term in (34) let us first consider the case \( n^\alpha < 3cB(\tau + \varphi(\varepsilon/2)2^\vartheta \tau^\vartheta) \). Combining (38) with (34) and using \( 1 \leq B_0, 1 \leq V, 3cB \leq c_\sigma, c_\sigma V \leq cV, 2 \leq 4^{1/(2-\vartheta)}, \) and \( \mathbb{E}_P h_{f_{Dn,\tau}} - \mathbb{E}_P h_{f_{\tilde{D}n,\tau}} \leq 2 \), then we find
\[
\Upsilon(f_{Dn,\tau}) + \mathcal{R}_{L.P}(\hat{f}_{Dn,\tau}) - \mathcal{R}_{L.P}
\leq \Upsilon(f_0) + 2\mathbb{E}_P h_{\tilde{f}_o} + \left( \frac{c_\sigma V^\tau}{n^\alpha} \right)^{1/\vartheta} + \frac{2cB\tau}{n^\alpha} + \frac{7eB_0\tau}{4n^\alpha}
+ (\mathbb{E}_P h_{\tilde{f}_{Dn,\tau}} - \mathbb{E}_P h_{\tilde{f}_{\tilde{D}n,\tau}}) + \delta
\leq \Upsilon(f_0) + 2\mathbb{E}_P h_{\tilde{f}_o} + \left( \frac{c_\sigma (\tau + \varphi(\varepsilon/2)2^\vartheta \tau^\vartheta)}{n^\alpha} \right)^{1/\vartheta} + \frac{4cB_0\tau}{n^\alpha}
+ 2 \left( \frac{c_\sigma (\tau + \varphi(\varepsilon/2)2^\vartheta \tau^\vartheta)}{n^\alpha} \right)^{1/\vartheta} + \delta
\]
\leq 2T(f_0) + 4\mathbb{E}_P h_{f_0} + 3 \left( \frac{c_V (\tau + \varphi (\varepsilon/2)^{2p} \rho^p)}{n^\alpha} \right)^{1/2} + \frac{8c_B B_0 \tau}{n^\alpha} + 2\delta

with probability \( \mu \) not less than \( 1 - 2Ce^{-\tau} \). It thus remains to consider the case \( n^\alpha \geq 3c_B (\tau + \varphi (\varepsilon/2)^{2p} \rho^p) \).

**Introduction of the Quotients.** To establish a non-trivial bound on the term \( \mathbb{E}_P h_{\tilde{f}_0} \tilde{U} f - \mathbb{E}_{D_n} h_{\tilde{f}_0} \tilde{U} f \) in (34), we define functions

\[
g_{f,r} := \mathbb{E}_P h_{\tilde{f}_0} \tilde{U} f - \mathbb{E}_P h_{\tilde{f}_0} \tilde{U} f, \quad f \in \mathcal{F}, \quad r > r^*.
\]

For \( f \in \mathcal{F} \), we have \( \|\mathbb{E}_P h_{\tilde{f}_0} - h_{\tilde{f}_0}\|_{\infty} \leq 2 \). Moreover, for \( f \in \mathcal{F}_r \), the variance bound (12) implies

\[
\mathbb{E}_P (h_{\tilde{f}_0} - \mathbb{E}_P h_{\tilde{f}_0})^2 \leq \mathbb{E}_P h_{\tilde{f}_0}^2 \leq V_r \leq V_r^g.
\]

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

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

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

\[
\tilde{\mathcal{C}}_{\varepsilon,r,k} := \bigcup_{l=0}^{k} \mathcal{C}_{\varepsilon,r,l}, \quad 0 \leq k \leq K + 1,
\]

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

\[
\mathcal{N}(A, \| \cdot \|_{\infty}, \varepsilon) \leq \mathcal{N}(B, \| \cdot \|_{\infty}, \varepsilon/2).
\]

By using (42) for \( \mathcal{F}_{2^k r} \setminus \mathcal{F}_{2^{k-1} r} \subset \mathcal{F}_{2^k r} \), we can estimate the cardinality of \( \tilde{\mathcal{C}}_{\varepsilon,r,k} \) by

\[
|\tilde{\mathcal{C}}_{\varepsilon,r,k}| = \left| \bigcup_{l=0}^{k} \mathcal{C}_{\varepsilon,r,l} \right| \leq \sum_{l=0}^{k} |\mathcal{C}_{\varepsilon,r,l}| = \sum_{l=0}^{k} \mathcal{N}(\mathcal{F}_{2^l r} \setminus \mathcal{F}_{2^{l-1} r}, \| \cdot \|_{\infty}, \varepsilon) \leq \mathcal{N}(B, \| \cdot \|_{\infty}, \varepsilon).}
\]
\[ \leq \sum_{l=0}^{k} \mathcal{N}(\mathcal{F}_{2^l r}, \| \cdot \|_\infty, \varepsilon/2) \]
\[ \leq \sum_{l=0}^{k} \exp \left( \varphi(\varepsilon/2)(2^l r)^p \right) \]
\[ \leq (k + 1) \exp \left( \varphi(\varepsilon/2)2^{kp} r^p \right), \quad 0 \leq k \leq K + 1. \] (43)

Peeling by Theorem 5.2 with \( Z_f := \mathbb{E}_{D_n}(Ep_h f - h_f), \Gamma(f) := \Upsilon(f) + Ep_h f \) 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 \( \varepsilon = \frac{1}{4} > 0 \) imply

\[ \mu \left( \sup_{f \in \tilde{C}_{\varepsilon, r}} \mathbb{E}_{D_n}(Ep_h f - h_f) > \frac{1}{4} \right) \]
\[ \leq \sum_{k=1}^{K+2} \mu \left( \sup_{f \in \tilde{C}_{\varepsilon, r}} \mathbb{E}_{D_n}(Ep_h f - h_f) > \frac{1}{4} (2^{k-1} r + r) \right) \]
\[ \leq \mu \left( \sup_{f \in \tilde{C}_{\varepsilon, r}} \mathbb{E}_{D_n}(Ep_h f - h_f) > \frac{1}{4} (r^* + r) \right) \]
\[ + \sum_{k=1}^{K+1} \mu \left( \sup_{f \in \tilde{C}_{\varepsilon, r}} \mathbb{E}_{D_n}(Ep_h f - h_f) > \frac{1}{4} (2^{k-1} r + r) \right) \]
\[ \leq \mu \left( \sup_{f \in \tilde{C}_{\varepsilon, r}} \mathbb{E}_{D_n}(Ep_h f - h_f) > \frac{1}{4} r \right) \]
\[ + \sum_{k=1}^{K+1} \mu \left( \sup_{f \in \tilde{C}_{\varepsilon, r}} \mathbb{E}_{D_n}(Ep_h f - h_f) > \frac{1}{4} 2^{k-1} r \right) \]
\[ \leq 2 \sum_{k=1}^{K+1} \mu \left( \sup_{f \in \tilde{C}_{\varepsilon, r}} \mathbb{E}_{D_n}(Ep_h f - h_f) > 2^{k-3} r \right) . \]

**Estimating the Error Probabilities on the “Spheres”.** Our next goal is to estimate all the error probabilities by using (30), (39) and the union bound. From \( \vartheta \in [0, 1] \) and the estimations of the covering numbers (43) follows that

\[ \mu \left( \sup_{f \in \tilde{C}_{\varepsilon, r}} \mathbb{E}_{D_n}(Ep_h f - h_f) > 2^{k-3} r \right) \]
\[ \leq C|\tilde{C}_{\varepsilon, r}| \exp \left( -\frac{3(2^{k-3} r)^2 \gamma}{6V(2^k r)^\vartheta + 4(2^{k-3} r)} \right) \]

20
\[
\leq C \cdot (k + 1) \exp \left( \varphi(\varepsilon/2)2k^p r^p \right) \cdot \exp \left( -\frac{3(2^{k-1}r)^2 n^\gamma}{192V (2^{k-1}r)^\vartheta + 16(2^{k-1}r)} \right).
\]

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

\[
p_k(r) := (k + 1) \exp \left( \varphi(\varepsilon/2)2k^p r^p \right) \cdot \exp \left( -\frac{3(2^{k-1}r)^2 n^\gamma}{192V (2^{k-1}r)^\vartheta + 16(2^{k-1}r)} \right).
\]

Then we have

\[
q_k(r) := \frac{p_{k+1}(r)}{p_k(r)} \leq \frac{k + 2}{k + 1} \exp \left( \varphi(\varepsilon/2)2^{k+1}r^p - \varphi(\varepsilon/2)(2^k r)^p \right) \cdot \exp \left( -\frac{3 \cdot 2^2(2^{k-1}r)^2 n^\gamma}{192V \cdot 2(2^{k-1}r)^\vartheta + 16 \cdot 2(2^{k-1}r)^\vartheta} + \frac{3(2^{k-1}r)^2 n^\gamma}{192V (2^{k-1}r)^\vartheta + 16(2^{k-1}r)} \right) \leq 2 \exp \left( \varphi(\varepsilon/2)2^{k+1}r^p \right) \cdot \exp \left( -\frac{3(2^{k-1}r)^2 n^\gamma}{192V (2^{k-1}r)^\vartheta + 16(2^{k-1}r)} \right),
\]

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

\[
q_k(r) \leq 2 \exp \left( \varphi(\varepsilon/2)2^{k+1}r^p \right) \cdot \exp \left( -\frac{3(2^{k-1}r)^2 n^\gamma}{192V (2^{k-1}r)^\vartheta + 16(2^{k-1}r)} \right) \leq 2 \exp \left( \frac{2^{k-1}p \cdot 4r^p \varphi(\varepsilon/2) - 2^{k-1}(2^{-\theta}) \cdot 3}{16(12V + 1)} 2^{-\frac{2\gamma+12}{n^\alpha}} b^{\frac{1}{n^\alpha}} r^{2-\vartheta} n^\alpha \right).
\]

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

\[
2^{(k-1)p} \leq 2^{(k-1)(2^{-\vartheta})}.
\]

The first assumption in (13), namely,

\[
r \geq \left( \frac{cV(\tau + \varphi(\varepsilon/2)2^{p} r^{p})}{n^\alpha} \right)^{\frac{1}{2-\vartheta}}
\]

with \( cV := \max \{ c, V, 2^{\frac{\vartheta+12}{n^\alpha}} b^{\frac{1}{n^\alpha}} (12V + 1)/3 \} \) implies that

\[
r \geq \left( 2^{\frac{\vartheta+12}{n^\alpha}} b^{\frac{1}{n^\alpha}} \frac{12V + 1}{3} \frac{\varphi(\varepsilon/2)2^{p} r^{p}}{n^\alpha} \right)^{\frac{1}{2-\vartheta}}
\]

or equivalently that

\[
4r^p \varphi(\varepsilon/2) \leq \frac{1}{2} \frac{3}{16(12V + 1)} 2^{-\frac{2\gamma+12}{n^\alpha}} b^{\frac{1}{n^\alpha}} r^{2-\vartheta} n^\alpha,
\]

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

\[
q_k(r) \leq 2 \exp \left( \frac{1}{2} \frac{3}{16(12V + 1)} 2^{-\frac{2\gamma+12}{n^\alpha}} b^{\frac{1}{n^\alpha}} r^{2-\vartheta} n^\alpha \right).
\]
Moreover, since \( \tau \geq 1 \), the first assumption in (13) implies also

\[
r \geq \left( 2^{\frac{2\alpha+1}{3}b^{-1} + \frac{12V}{3} + 1/n^\alpha} \right) \frac{r}{n^\alpha}
\]
or equivalently that

\[
\frac{1}{2} \cdot 3 \cdot 2^{\frac{2\alpha+1}{3}b^{-1} + \frac{12V}{3} + 1/n^\alpha} \cdot r^{2-\vartheta} n^\alpha \geq 4,
\]
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

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

\[
\leq 2C \cdot p_1(r) \cdot \sum_{k=0}^{K} (2e^{-4})^k
\]

\[
\leq C \cdot \frac{2}{1-2e^{-4}} \cdot p_1(r)
\]

\[
\leq 3C \cdot p_1(r)
\]

\[
= 6C \exp (\varphi(\varepsilon/2)2^{p}r^p) \cdot \exp \left( -\frac{3r^2n(\gamma)}{192V r^\alpha + 16r} \right)
\]

\[
\leq 6C \exp (\varphi(\varepsilon/2)2^{p}r^p) \cdot \exp \left( -\frac{3r^2n(\gamma)}{192V r^\alpha + 16r} \right)
\]

\[
\leq 6C \exp (\varphi(\varepsilon/2)2^{p}r^p) \cdot \exp \left( -3 \cdot 2^{\frac{2\alpha+1}{3}b^{-1} + \frac{12V}{3} + 1/n^\alpha} \right).
\]

Then once again the first assumption in (13) gives

\[
r \geq \left( \frac{16(12V + 1)}{3 \cdot 2^{\frac{2\alpha+1}{3}b^{-1} + \frac{12V}{3} + 1/n^\alpha}} \right) \frac{r}{n^\alpha}
\]

and a simple transformation thus yields

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

Consequently we see that with probability \( \mu \) not less than \( 1 - 6Ce^{-\tau} \) we have

\[
\mathbb{E}_p h_f - \mathbb{E}_{D_n} h_f \leq \frac{1}{4} \left( T(f) + \mathbb{E}_p h_f + r \right)
\]

(44)

for all \( f \in C_{c,1} \). Since \( r \in (0,1] \), we have \( f_{D_n,r} \in F_1 \), i.e. either \( f_{D_n,r} \in F_r \), or there exists an integer \( k \leq K + 1 \) such that \( f_{D_n,r} \in F_{2^kr} \cup F_{2k-1,r} \). Thus there exists an
\(f_{D_n} \in \mathcal{C}_{\varepsilon,r,k} \subset F_{2^k-1, r} \) with \(\|f_{D_n}\|_{\infty} \leq \varepsilon\). By the assumed Lipschitz continuity of the clipped \(L\) the latter implies

\[
|h_{f_{D_n}}(x,y) - h_{f_{D_n}, Y}(x, y)| \leq |\tilde{f}_{D_n}(x) - \tilde{f}_{D_n, Y}(x)| \leq |f_{D_n}(x) - f_{D_n, Y}(x)| \leq \varepsilon
\]

(45) for all \((x, y) \in X \times Y\). For \(f_{D_n, Y}, f_{D_n} \in F_r\) we obviously have

\[
\Upsilon(f_{D_n}) + \mathbb{E} h_{f_{D_n}} \leq r
\]

and for the other cases \(f_{D_n, Y}, f_{D_n} \in F_{2^k-1, r}\) we obtain

\[
\Upsilon(f_{D_n}) + \mathbb{E} h_{f_{D_n}} \leq 2^k r = 2 \cdot 2^{k-1} r \leq 2 \left( \Upsilon(f_{D_n, Y}) + \mathbb{E} h_{f_{D_n, Y}} \right),
\]

consequently, we always have

\[
\Upsilon(f_{D_n}) + \mathbb{E} h_{f_{D_n}} \leq 2 \left( \Upsilon(f_{D_n, Y}) + \mathbb{E} h_{f_{D_n, Y}} \right) + r.
\]

(46)

Combining (45) with (44) and (46), we obtain

\[
\mathbb{E} h_{f_{D_n, Y}} - \mathbb{E} h_{f_{D_n, Y}} \leq \frac{1}{2} \left( \Upsilon(f_{D_n, Y}) + \mathbb{E} h_{f_{D_n, Y}} + \varepsilon + r \right) + 2 \varepsilon
\]

with probability \(\mu\) not less than \(1 - 6 Ce^{-\tau}\). By combining this estimate with (34) and (38), we then obtain that

\[
\Upsilon(f_{D_n, Y}) + \mathbb{E} h_{f_{D_n, Y}} \leq \Upsilon(f_0) + 2 \mathbb{E} h_{f_0} + \left( \frac{c_\sigma V_{\tau}}{n^\alpha} \right)^{\frac{\tau}{2-\tau}} + \frac{2 c_0 v \tau}{n^\alpha} + \frac{7 c_0 B_0 \tau}{4 n^\alpha} + \delta
\]

\[
+ \frac{5}{2} + \frac{1}{2} + \frac{r}{2}
\]

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

\[
\Upsilon(f_{D_n, Y}) + \mathbb{E} h_{f_{D_n, Y}} \leq 2 \Upsilon(f_0) + 4 \mathbb{E} h_{f_0} + 4 r + 5 \varepsilon + 2 \delta,
\]

i.e. we have shown the assertion. \(\square\)
References


