Intrinsic Dimension Adaptive Learning Rates for Kernel Methods

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

vorgelegt von

Thomas Hamm

aus Sindelfingen

Hauptberichter:	Prof. Dr. Ingo Steinwart
Mitberichter:	Prof. Ph.D. (Harvard) Christian Hesse
Mitberichter:	Prof. Dr. Lorenzo Rosasco
Tag der mündlichen Prüfung:	8. Februar 2022

Institut für Stochastik und Anwendungen der Universität Stuttgart

2022

Danksagung

Meinen herzlichsten Dank möchte ich meinem Betreuer Ingo Steinwart aussprechen, sowohl für die Annahme als Doktorand, als auch für die wissenschaftliche Betreuung über die letzten Jahre, in denen diese Arbeit entstanden ist. Weiterhin möchte ich mich herzlich bei Philipp Hennig und Sebastian Trimpe für viele hilfreiche Diskussionen und Kommentare im Rahmen des *Thesis Advisory Committee* bedanken, sowie bei den Mitberichtern für ihr Mitwirken im Prüfungsausschuss. Zu guter Letzt möchte ich mich bei all meinen aktuellen und ehemaligen Kollegen am Fachbereich bedanken mit denen sowohl die Zusammenarbeit stets eine große Freude war, als auch die Zusammenkünfte abseits des wissenschaftlichen Alltags ein willkommener Ausgleich waren. Ganz besonders möchte ich mich jedoch bei Simon Fischer bedanken für das Korrekturlesen dieser Arbeit, die vielen konstruktiven Vorschlge und zahlreiche interessante und lehrreiche Diskussionen in den letzten Jahren.

Kurzfassung

Lernraten nicht-parametrischer Lernmethoden weisen in der Regel eine unvorteilhafte Abhängigkeit von der Dimension des Eingaberaums auf, ein Phänomen, das in der statistischen Lerntheorie allgemein als Fluch der Dimensionalität bekannt ist. In der Praxis geht man jedoch davon aus, dass hochdimensionale Daten in der Regel eine niedrigdimensionale intrinsische Struktur aufweisen. Daher ist es eine interessante Frage, ob Lernmethoden eine solche niedrigdimensionale Struktur in dem Sinne ausnutzen können, dass sie Lernraten erzielen, bei denen die Abhängigkeit von der Dimension des umgebenden Raums durch die intrinsische Dimension der Daten ersetzt wird. Die verbreitetste Methode um die intrinsische Dimension der Daten zu beschreiben beruht auf der Annahme, dass die datenerzeugende Verteilung auf einer niedrigdimensionalen glatten Mannigfaltigkeit getragen ist. Wir schwächen diese Annahme erheblich ab, indem wir die fraktale Dimension des Trägers der datenerzeugenden Verteilung betrachten.

Genauer leiten wir in Kapitel 2 für Support Vector Machines (SVMs), die einen Gauß-Kern verwenden, für die Kleinste-Quadrate-Regression und die binäre Klassifikation unter Verwendung der Hinge-Verlustfunktion Lernraten her, in denen die Abhängigkeit von der Dimension des umgebenden Raums durch die Box-Counting-Dimension des Trägers der datenerzeugenden Verteilung ersetzt wird. Die sich daraus ergebenden Lernraten für die Regression sind bis auf logarithmischen Faktoren minimax-optimal und die Lernraten für die Klassifikation sind bis auf logarithmischen Faktoren in einem bestimmten Bereich unserer Annahme minimax-optimal und haben ansonsten die Form der besten bekannten Raten. Wir zeigen außerdem, dass diese Raten adaptiv durch einen Trainingsvalidierungsansatz für die Auswahl der Hyperparameter erreicht werden können. Genauer gesagt zeigen wir, dass es zum Erreichen optimaler Raten auf adaptive Weise ausreicht, wenn die Anzahl der Kandidatenwerte für die Hyperparameter logarithmisch mit dem Stichprobenumfang wächst, während existierende vergleichbare Resultate erfordern, dass die Anzahl der Kandidatenwerte mindestens linear mit dem Stichprobenumfang wächst.

In Kapitel 3 beweisen wir ähnliche Ergebnisse für eine räumlich lokalisierte Variante von Gaußschen-SVMs, die eine verbreitete Methode zur Minderung des Rechenaufwands gewöhnlicher SVMs sind, welche quadratisch im Platz- und mindestens quadratisch Zeitbedarf sind. Bei diesem lokalisierten SVM-Ansatz wird eine Partition des Eingaberaums berechnet, und dann für jede Zelle dieser Partition eine SVM-Entscheidungsfunktion berechnet, wobei nur die in der jeweiligen Zelle enthaltenen Stichproben verwendet werden. Während existierende Ergebnisse zu Verfahren, die mit unserem vergleichbar sind, eine a-priori festgelegte Partition des Eingaberaums betrachten, die einige technische Annahmen erfüllt, betrachten wir eine vollständig datenabhängige Partition, die auf dem Farthest-First-Traversal-Algorithmus basiert. In diesem Kapitel hängt unser Begriff der intrinsischen Dimension von der Assouad-Dimension des Trägers der datenerzeugenden Verteilung ab. Wir beweisen erneut die gleichen minimax-optimalen Raten unter diesem etwas stärkeren Begriff der intrinsischen Dimensionalität. Wir beweisen ebenfalls erneut, dass diese Raten adaptiv durch ein Trainingsvalidierungsverfahren mit logarithmisch wachsenden Kandidatenmengen erreicht werden können. Die Ergebnisse dieses Kapitels sind die ersten, die die Adaptivität an die intrinsische Dimensionalität der Daten für eine Beschleunigungsstrategie für Kern-Methoden berücksichtigen.

In Kapitel 4 ergänzen wir schließlich unsere theoretischen Erkenntnisse aus Kapitel 2 und 3 durch experimentelle Untersuchungen. Dazu betten wir gegebene Datensätze mittels einer nicht-trivialen randomisierten Einbettung in einen höherdimensionalen Raum ein und vergleichen, wie sich der Testfehler in Abhängigkeit von der Anzahl der künstlich hinzugefügten Dimensionen verhält. Wir führen dieses Verfahren für eine Reihe von Regressions- und Klassifikationsdatensätzen für die in Kapitel 2 und 3 betrachteten Lernmethoden durch. Die Ergebnisse unserer Experimente deuten darauf hin, dass die intrinsische Dimension der Daten tatsächlich die entscheidende Größe ist, die die Generalisierungsleistung bestimmt, anstatt der Dimension des umgebenden Raums.

Summary

Learning rates for non-parametric learning methods usually exhibit a poor dependency on the dimension of the input space, a phenomenon commonly known as the curse of dimensionality in statistical learning theory. In practice however, high dimensional data usually is hypothesized to have some low dimensional intrinsic structure and therefore it is an interesting question whether learning methods can exploit such a low dimensional structure in the sense that they achieve learning rates where the dependence of the dimension of the ambient space is replaced with the intrinsic dimension of the data. The most common notion of intrinsic dimension of data relies on the assumption that the data generating distribution is supported on a low-dimensional smooth manifold. We substantially weaken this assumption by considering the fractal dimension of the support of the data generating distribution.

More precisely, in Chapter 2 we derive learning rates for support vector machines (SVMs) using a Gaussian kernel for least-squares regression and binary classification using the hinge loss, where the dependence of the dimension of the ambient space is replaced with the box-counting dimension of the support of the data generating distribution. The resulting learning rates for regression are minimax optimal up to logarithmic factors and the learning rates for classification are minimax optimal up to logarithmic factors in a certain range of our assumption and otherwise of the form of the best known rates. We further show that these rates can be achieved adaptively by a training validation approach for hyperparameter selection. More specifically, we show that in order to achieve optimal rates adaptively it is sufficient for the size of the candidate sets of values for the hyperparameters to grow logarithmically in the sample size, whereas existing similar results require the size of candidate sets to grow at least linearly in the sample size.

In Chapter 3 we prove similar results for a spatially localized version of Gaussian SVMs, which is a popular method for circumventing the computational costs of ordinary SVMs, which are quadratic in space and at least quadratic in time. In this localized SVM approach a partition of the input space is computed and then

for each cell of that partition an SVM decision function is computed using only the samples contained in that respective cell. Whereas existing results similar to ours consider an a-priori fixed partition of the input space satisfying some technical assumptions, we consider a fully data dependent partitioning based on the farthest first traversal algorithm. In this chapter our notion of intrinsic dimension depends on the Assouad dimension of the support of the data generating distribution. We again prove the same minimax optimal rates under this slightly stronger notion of intrinsic dimensionality. We also again prove that these rates can be achieved adaptively using a training validation procedure using logarithmically growing candidate sets. The results of this chapter are the first to consider adaptivity to the intrinsic dimension of the data for a speed-up strategy for kernel methods.

Finally, in Chapter 4 we complement our theoretical findings of Chapter 2 and 3 by experimental investigation. To this end, we embed a given dataset into a higher dimensional space via a non-trivial randomized embedding and compare how the test error changes depending on the number of artificially added dimensions. We perform this procedure for a number of regression and classification datasets using the learning methods considered in Chapter 2 and 3. The results of our experiments suggest that, in fact, the intrinsic dimension of the data is the critical quantity determining the generalization performance, as opposed to the dimension of the ambient space.

List of Figures

List of Tables

Notation

- We denote the set of natural numbers $\{1, 2, 3, \dots\}$ by $\mathbb{N}$ and $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$. $\mathbb{R}$ denotes the set of real numbers.

- For $p \in [1, \infty]$ let $\|\cdot\|_p$ be the usual p-norm which for $x \in \mathbb{R}^d$ is defined by

$$\|x\|_p := \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} \qquad \text{for } p < \infty$$

  and $\|x\|_\infty := \max_{i=1,\dots,d} |x_i|$. Further, we denote the space $\mathbb{R}^d$ equipped with the norm $\|\cdot\|_p$ by ℓ_p^d.

- $\|x\|$ for $x \in \mathbb{R}^d$ (without subscript) denotes the Euclidean norm, i.e. $\|x\| := \|x\|_2$. For $x \in \mathbb{R}^d$ and $r > 0$ let $B_r(x) := \{y \in \mathbb{R}^d : \|x - y\| \leq r\}$ be the closed ball (w.r.t. Euclidean distance) with center x and radius r.

- Given a normed space E, we denote the closed unit ball centered at the origin by B_E.

- For a non-empty set X let $\ell_\infty(X)$ be the space of bounded functions $f : X \to \mathbb{R}$ equipped with the norm $\|f\|_{\ell_\infty(X)} := \sup_{x \in X} |f(x)|$. If the domain X is known from the context we may also use the shorthand $\|f\|_\infty$.

- Given a measure space $(X, \mathcal{F}, \mu)$ and $p \in [1, \infty]$ let $L_p(\mu)$ denote the Lebesgue space of p-integrable functions with the norm

$$\|f\|_{L_p(\mu)} := \left(\int_X |f|^p \, \mathrm{d}\mu \right)^{1/p}$$

  for $p < \infty$ and the usual modification for $p = \infty$. We use the notation $L_p(\mathbb{R}^d)$ when μ is the Lebesgue measure on $\mathbb{R}^d$. As in this work there is no ambiguity to expect, we do not distinguish between equivalence classes of functions and actual functions.

- For a random variable X let $\mathbf{E}X$ denote the expectation of X. Also, for $Y \subset \mathbb{R}^d$ equipped with some probability distribution $\mathbf{P}$, we will sometimes also write $\mathbf{E}Y$ although Y is technically not a random variable. We prefer this over the cumbersome notation $\mathbf{E}\,\mathrm{id}_Y$. In this case we may also write $\mathbf{E}_{y\sim\mathbf{P}} f(y)$ for a function f to stress the distribution with respect to which the expectation is taken.

- For real-valued functions f, g defined on some topological space (we usually consider $\mathbb{N}$ or an interval in $\mathbb{R}$) us the notation $f(x) \lesssim g(x)$ as $x \to a$ if there exists a neighborhood U of a and a constant $C > 0$ such that $f(x) \leq Cg(x)$ for all $x \in U$. Furthermore, we denote $f(x) \asymp g(x)$ as $x \to a$ if $f(x) \lesssim g(x)$ as well as $g(x) \lesssim f(x)$ as $x \to a$. The usual Landau symbol $\mathcal{O}(g)$ denotes the class of all functions f such that $f(x) \lesssim g(x)$.

Publications

The contents of Chapters 2 and 5 and parts of Section 1.3 were published in

[24] T. Hamm and I. Steinwart. Adaptive learning rates for support vector machines working on data with low intrinsic dimension. *Ann. Statist.*, 49:3153–3180, 2021

The contents of Chapters 3 and 4 will be published in

[25] T. Hamm and I. Steinwart. Intrinsic dimension adaptive partitioning for kernel methods. *SIAM J. Math. Data Sci.*, 2022 (accepted)

Contents

1 Introduction and Preliminaries

In the first section of this introductory chapter we give a brief overview, including some examples, of the main concepts of statistical learning theory, such as loss functions, the risk functional, Bayes risk and Bayes decision function, and the notion of learning rates. Associated to the latter, we illustrate the curse of dimensionality and its consequences. Subsequently in Section 1.2, we give an introduction to reproducing kernel Hilbert spaces and support vector machines as the central learning method of this thesis. Finally, in Section 1.3 we introduce our main tools for the statistical analysis of the subsequent chapters, such as variance bounds and entropy numbers, as well as some preliminary results on the latter that are used in the following chapters.

1.1 Elements of Statistical Learning Theory and the Curse of Dimensionality

The central goal in supervised learning is to learn a relationship between an input and an output variable based on a number of training examples. Formally, given an input space $X \subset \mathbb{R}^d$ and an output space $Y \subset \mathbb{R}$ as well as a training set $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ sampled from $\mathbf{P}^n$, where $\mathbf{P}$ is an unknown probability distribution on $X \times Y$, we want to find a decision function $f_D : X \to \mathbb{R}$, such that $f_D(x)$ is a good prediction for $y \in Y$ with respect to the conditional distribution of $\mathbf{P}$ on Y given x. The quality of a prediction is measured by a loss function, which is a measurable function $L : Y \times \mathbb{R} \to [0, \infty)$. Prominent examples for loss functions are the least-squares loss $L_{\mathrm{LS}}(y, t) := (y - t)^2$ for regression tasks when $Y \subset \mathbb{R}$ is some interval and the binary classification loss $L_{\mathrm{class}}(y, t) := \mathbf{1}_{(-\infty,0]}(y \operatorname{sgn} t)$ when the output space consists of two classes $Y = \{-1, 1\}$, where sgn denotes the sign function with the convention $\operatorname{sgn} 0 := 1$. The risk of a decision

function $f : X \to \mathbb{R}$ is defined as its expected loss, that is

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

Obviously, we want to find decision functions such that its risk is as small as possible. To this end, we refer to the smallest possible risk, denoted by $\mathcal{R}^*_{L,\mathbf{P}} := \inf_{f:X\to\mathbb{R}} \mathcal{R}_{L,\mathbf{P}}(f)$, as the Bayes risk and any function $f^*_{L,\mathbf{P}}$ with $\mathcal{R}_{L,\mathbf{P}}(f^*_{L,\mathbf{P}}) = \mathcal{R}^*_{L,\mathbf{P}}$ as a Bayes decision function with respect to L and $\mathbf{P}$.

Example 1.1.1. Let $L = L_{\mathrm{LS}}$ be the least-squares loss and assume that $\mathbf{E}(Y^2) < \infty$. Let $m(x) := \mathbf{E}(Y|X = x)$ be the conditional mean function, see [30, Section 8.2] for an introduction to conditional expectations. Some elementary calculations, see [23, Section 1.1], then show that for any $f \in L_2(\mathbf{P}_X)$, where $\mathbf{P}_X$ denotes the marginal distribution of $\mathbf{P}$ on X, we have

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(f) &= \int_{X\times Y} (y - f(x))^2 \, \mathrm{d}\mathbf{P}(x, y) \\ &= \int_{X\times Y} (y - m(x))^2 \, \mathrm{d}\mathbf{P}(x, y) + \int_X (m(x) - f(x))^2 \, \mathrm{d}\mathbf{P}_X(x). \qquad (1.1)\end{aligned}$$

Now, the first summand in (1.1) is independent of f and both are non-negative. From this we can conclude that

$$\mathcal{R}^*_{L,\mathbf{P}} = \int_{X\times Y} (y - m(x))^2 \, \mathrm{d}\mathbf{P}(x, y)$$

and f is a Bayes decision function if and only if $f(x) = m(x)$ for $\mathbf{P}_X$-almost all $x \in X$.

Example 1.1.2. Let $L = L_{\mathrm{class}}$ be the classification loss and let $\eta(x) := \mathbf{P}(y = 1|X = x)$ be a version of the regular conditional distribution of Y given $x \in X$, see for example [30, Section 8.3] for an introduction to regular conditional distributions. The risk of a function $f : X \to \mathbb{R}$ can be expressed by

$$\mathcal{R}_{L,\mathbf{P}}(f) = \mathbf{P}\big((x, y) \in X \times Y : \operatorname{sgn} f(x) \neq y\big)$$

and the Bayes risk is given by

$$\mathcal{R}^*_{L,\mathbf{P}} = \int_X \min\{\eta(x), 1 - \eta(x)\} \, \mathrm{d}\mathbf{P}_X(x),$$

see [55, Example 2.4], where again $\mathbf{P}_X$ denotes the marginal distribution of $\mathbf{P}$ on

X. Finally, f is a Bayes decision function if and only if $(2\eta(x)-1)\operatorname{sgn} f(x) \geq 0$ for $\mathbf{P}_X$-almost all $x \in X$.

Because of the non-convexity of the classification loss, we will also consider the hinge loss $L_{\text{hinge}} : Y \times \mathbb{R} \to [0,\infty)$ defined by $L_{\text{hinge}}(y,t) := \max\{0, 1-yt\}$ as a surrogate. Note that we call a loss function $L : Y \times \mathbb{R} \to [0,\infty)$ convex, if for all $y \in Y$ the function $L(y,\cdot)$ is convex. The use of the hinge loss as a surrogate for the classification loss is justified by Zhang's inequality [55, Theorem 2.31], which states that

$$\mathcal{R}_{L_{\text{class}},\mathbf{P}}(f) - \mathcal{R}_{L_{\text{class}},\mathbf{P}} \leq \mathcal{R}_{L_{\text{hinge}},\mathbf{P}}(f) - \mathcal{R}_{L_{\text{hinge}},\mathbf{P}} \tag{1.2}$$

for all measurable $f : X \to [-1,1]$ and any distribution $\mathbf{P}$ on $X \times Y$.

Particularly interesting in the statistical analysis of learning methods is the behavior of the *excess risk* $\mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}^*_{L,\mathbf{P}}$ as the number of samples n goes to infinity. Specifically, from a good learning method we would expect that the excess risk converges to zero. To this end, a learning method is called strongly universal consistent if $\mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}^*_{L,\mathbf{P}} \to 0$ almost surely as $n \to \infty$ for *any* distribution $\mathbf{P}$. Note that the excess risk is a random variable through its dependence on the dataset D. Therefore, different notions of universal consistency exist, for example where convergence in probability or expectation is considered instead. A series of learning methods are known to be universally consistent, see e.g. [14] for a number of examples of consistent classifying methods and [23] for consistent regression methods. However, consistency does not specify the speed of convergence of the excess risk. In fact, by the no-free-lunch theorem, a fundamental result in statistical learning theory, there exists no learning method $D \mapsto f_D$ such that the excess risk converges to zero with an a-priori specified speed for *any* distribution $\mathbf{P}$. The no-free-lunch theorem was first proved by Devroye [13] for the binary classification problem but can be generalized to a broad class of learning tasks under mild assumptions on the loss function, see for example [55, Corollary 6.8]. As a consequence, rates of convergence of the excess risk can only be proved for specified classes of distributions, defined by so-called regularity assumptions. As already noted, the excess risk is a random variable and thus there exist multiple notions of convergence. The most common type of convergence considered in the literature is convergence in expectation, that is bounds of the type

$$\mathbf{E}_{D\sim\mathbf{P}^n}\mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}^*_{L,\mathbf{P}} \leq Cn^{-r} \tag{1.3}$$

for some constant $C > 0$ independent of n and some $r \in (0, 1]$ holding for all $\mathbf{P}$ out of an often implicitly defined class of distributions. In this context, n^{-r} is referred to as a *learning rate*. In this thesis however, we will consider slightly stronger high probability bounds on $\mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}^*_{L,\mathbf{P}}$, which state that there exists a constant $C > 0$ such that the set of all samples $D \in (X \times Y)^n$ with

$$\mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau n^{-r}$$

has a probability of at least $1 - \mathrm{e}^{-\tau}$ with respect to $\mathbf{P}^n$ for all $\tau \geq 1$. Generally throughout this thesis, we will call the order of the rate of convergence of the excess risk a learning rate, regardless of their exact type of convergence. The reader should keep this minor impreciseness in terminology in mind, especially when we refer to results on learning rates in the literature.

A common observation on learning rates for non-parametric estimators is, that learning rates are significantly deteriorated when the dimension of the input space increases. We want to illustrate this in the case of regression using the least squares-loss $L = L_{\mathrm{LS}}$, where regularity assumptions on $\mathbf{P}$ are usually expressed by smoothness properties of the Bayes decision function $f^*_{L,\mathbf{P}}(x) = \mathbf{E}(Y|X = x)$. To this end, let $\mathcal{P}$ be the class of distributions $\mathbf{P}$ on $X \times Y$ with $X = [0, 1]^d$ such that the marginal distribution $\mathbf{P}_X$ has a Lebesgue density bounded away from zero and infinity and such that $f^*_{L,\mathbf{P}}$ is k-times continuously differentiable and all derivatives of order k are bounded by some constant $M > 0$. Then, a variety of learning methods were shown to achieve a learning rate of $n^{-2k/(2k+d)}$. Moreover, the classical result of Stone [60] states that this learning rate is minimax optimal, i.e. there exists no learning method $D \mapsto f_D$ that achieves a learning rate faster than $n^{-2k/(2k+d)}$ uniformly over the class $\mathcal{P}$. As a consequence, to halve the error of a regression estimator in this setting, one usually has to increase the sample size by a factor of $2^{(2k+d)/(2k)}$, which is exponential in d. This poor dependence of the generalization error on the dimension of the input space is usually termed the curse of dimensionality and is also present in binary classification, see for example the results of Audibert and Tsybakov [2].

As a consequence of the discussion above, learning from very high dimensional data should infeasible from a theoretical point of view. However, note that a central assumption in the result of Stone mentioned in the previous paragraph is that the marginal distribution $\mathbf{P}_X$ has a Lebesgue density bounded away from zero and

infinity, which means that the data points $x_1, \ldots, x_n$ are roughly uniformly spread out over the whole input space. In practice however, especially in high dimensions, real world datasets tend to be concentrated on a significantly smaller portion of the input space. Levina and Bickel [37] even went as far as saying that

> "There is a consensus in the high-dimensional data analysis community that the only reason any methods work in very high dimensions is that, in fact, the data are not truly high-dimensional."

To illustrate this reasoning consider the standard introductory textbook example for supervised learning that consists of estimating the price of a house based on certain characteristics. Now, it should be obvious that some of these characteristics, like *living space* and *number of rooms* are highly correlated in the sense that a large amount of living space implies a big number of rooms and vice versa. In other words, our dataset,in this case the characteristics of the set of houses we know the prices of, can not contain any arbitrary combination of characteristics but only very specific ones. Therefore, it seems plausible that generally, the more features we collect in a supervised learning scenario, the more likely they are to contain some highly correlated features, which in turn implies that the data is concentrated on a small portion of the input space.

It is therefore an interesting question if common learning methods are able to exploit a, yet to be formalized, assumption on the intrinsic dimensionality of the data. This question has received considerable attention in the literature, see for example [5, 11, 31, 32, 33, 35, 40, 53, 64, 65, 66]. The by far widest spread notion to express this low intrinsic dimensionality of data is to assume that the data generating distribution $\mathbf{P}_X$ is supported on a low-dimensional, smooth manifold. The obvious goal under this assumption then is to prove learning rates that coincide with the well known-ones, but where the dependence on the dimension of the ambient space is replaced with the dimension of the manifold on which $\mathbf{P}_X$ is supported, or at least, prove learning rates that only depend on the dimension of this manifold. However, there is a considerable gap between the commonly accepted hypothesis that the data is not uniformly spread over the input space and the assumption, derived from this hypothesis, that the data lies on a smooth manifold. In this thesis we prove learning rates, coinciding with some well-known minimax optimal rates, for regression and classification using Gaussian SVMs that depend on the *fractal* dimension of the support of $\mathbf{P}_X$, allowing also non-integer dimensions and considerably weakening the prevailing manifold assumption. More precisely we will

show in Chapter 2 that Gaussian SVMs achieve some well-known minimax optimal learning rates for regression and classification, where the dependence of the ambient space is replaced with the box-counting dimension of the support of the data generating distribution. Similarly, in Chapter 3 we will show that analogous results hold for Gaussian SVMs working on a data dependent partition of the input space, where the notion of intrinsic dimensionality is based on the Assouad dimension of the support of the data generating distribution, a slightly stronger notion of fractal dimension than the box-counting dimension. An interesting finding of this thesis therefore is that the differentiable structure of the data, which is actively exploited in the proofs of existing publications which work under a manifold assumption, such as [64, 65, 66], is apparently not necessary for results as described above to hold true.

1.2 Reproducing Kernel Hilbert Spaces and Support Vector Machines

In this section we introduce the central learning method this thesis is concerned with, namely support vector machines. To this end, we first need to start with a recap on reproducing kernel Hilbert spaces (RKHSs), which serve as the hypothesis space of support vector machines. We start with the definition of a kernel.

Definition 1.2.1 (Kernel)**.** Let X be a non-empty set. A function $k : X \times X \to \mathbb{R}$ is called a *kernel* if there exists a (real) Hilbert space H_0 and a map $\Phi : X \to H_0$ such that $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{H_0}$ for all $x, y \in X$. The map Φ is called a *feature map* and H_0 a *feature space* of k.

By [55, Theorem 4.16], a symmetric function $k : X \times X \to \mathbb{R}$ is a kernel if and only if it is positive definite, that is if for all $n \in \mathbb{N}$ and all choices $x_1, \ldots, x_n \in X$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) \geq 0.$$

Definition 1.2.2 (Reproducing Kernel/Reproducing Property)**.** Let X be a non-empty set and H a Hilbert space consisting of functions $f : X \to \mathbb{R}$. A function k :

$X \times X \to \mathbb{R}$ is a *reproducing kernel* if $k(x, \cdot) \in H$ for all $x \in X$ and

$$f(x) = \langle f, k(x, \cdot) \rangle \quad \text{for all } f \in H, x \in X. \tag{1.4}$$

Property (1.4) is called the *reproducing property*.

Definition 1.2.3 (RKHS)**.** Let X be a non-empty set and H a Hilbert space consisting of functions $f : X \to \mathbb{R}$. Then H is called a *reproducing kernel Hilbert space* (RKHS) if the evaluation functional $H \to \mathbb{R}$ defined by $f \mapsto f(x)$ is continuous for every $x \in X$.

Definitions 1.2.1, 1.2.2, and 1.2.3 are connected in the following way: Every reproducing kernel in the sense of Definition 1.2.2 is a kernel in the sense of Definition 1.2.1 via the *canonical feature map* $\Phi(x) := k(x, \cdot)$, see [55, Lemma 4.19]. Additionally, every RKHS H over a non-empty set X has a unique reproducing kernel [55, Theorem 4.20] given by $k(x, y) = \langle \delta_x, \delta_y \rangle_{H'}$ for $x, y \in X$ where $\delta_x, \delta_y : H \to \mathbb{R}$ denote the evaluation functionals at x, respectively y and H' denotes the dual space of H. Conversely, every kernel k has a unique RKHS H, for which it is the reproducing kernel, consisting of the functions $x \mapsto \langle \Phi(x), w \rangle_{H_0}, w \in H_0$, where $\Phi : X \to H_0$ is a feature map of k and the norm in H is given by

$$\|f\|_H = \inf \{ \|w\|_{H_0} : w \in H_0 \text{ with } f = \langle \Phi(\cdot), w \rangle \}, \tag{1.5}$$

see [55, Theorem 4.21].

Example 1.2.4. The basic idea behind the usage of kernels in machine learning is to extend linear algorithms to non-linear algorithms without much effort. To illustrate this, let $X = \mathbb{R}^2$ and define the feature map $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$ by $\Phi(x) := (x_1, x_2, x_1^2 + x_2^2)$. Now, in the left graphic of Figure 1.1 the red and blue points can not be separated by a linear decision function. However, if we map the points into the feature space H_0 via Φ they are linearly separable in H_0. This extension of the expressivity of linear algorithms by the introduction of a kernel is known as the *kernel trick*. For more details on the kernel trick and examples of "kernelizable" algorithms we refer to [52, Chapters 13-16]. To continue our example, the kernel k associated to the feature map Φ is then given by $k(x, y) = \langle x, y \rangle + \|x\|^2 \|y\|^2$ for $x, y \in \mathbb{R}^2$. The RKHS H of k consists of all functions of the form $f_w(x) = w_1 x_1 + w_2 x_2 + w_3 (x_1^2 + x_2^2)$ with $w \in \mathbb{R}^3$. One can easily check that the representation of f_w via w is unique. Equation (1.5) therefore gives us $\|f_w\|_H = \|w\|$.

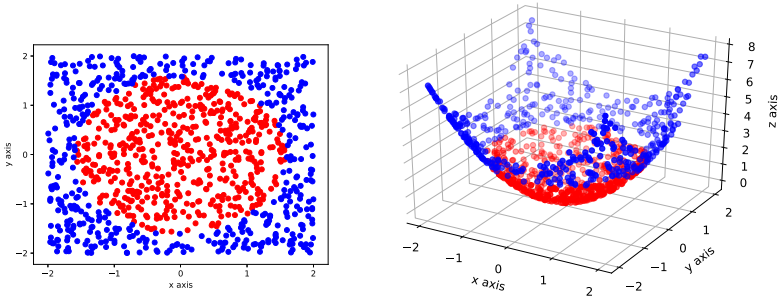

Figure 1.1: The red and blue points are not linearly separable in the original space (left), but after they are mapped into the feature space $\mathbb{R}^3$ via Φ they are linearly separable.

A *support vector machine* (SVM) is a regularized empirical risk minimizer over a reproducing kernel Hilbert space. Formally, given $D = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ as dataset and an RKHS H with kernel k on X, a support vector machine computes the minimizer of the regularized empirical risk

$$f_{D,\lambda} := \arg\min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)), \tag{1.6}$$

where $\lambda > 0$ is a regularization parameter and $L : Y \times \mathbb{R} \to [0, \infty)$ is a loss function. For some remarks on the history of SVMs and their originally geometrical interpretation as hard margin or soft margin classifiers using the classification loss or hinge loss, respectively, and their extension using the kernel trick we refer to [55, Chapter 1]. Note that, due to this geometrically motivated origin of SVMs, nomenclature for the learning method defined by (1.6) is not consistent in the literature. While some authors generally denote this generic learning method support vector machine, as we do, others use this term exclusively for the classification method (1.6) using the hinge loss $L = L_{\text{hinge}}$. Method (1.6) using the least-squares loss $L = L_{\text{LS}}$ is also commonly called kernel ridge regression.

By the representer theorem [55, Theorem 5.5] $f_{D,\lambda}$ exists and is unique for convex loss functions L and has an expansion

$$f_{D,\lambda}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) \tag{1.7}$$

with coefficients $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, which makes the optimization problem (1.6) over a generally infinite dimensional space H computable in practice. We will also consider a modification of the learning method (1.6) intended for alleviating computational constraints of this learning method, which are quadratic in space and at least quadratic in time, and denote this method a *local SVM* (LSVM). In this modification a spatial partition of the input space X is computed as a first step and then on each cell of this partition a decision function (1.6) is computed using only the samples contained in that cell. In this thesis we consider a fully data dependent partitioning procedure based on the farthest first traversal algorithm, whereas existing results on similar partitioning based estimators consider an a-priori fixed partition satisfying some technical assumptions. We will describe the construction of our considered partition as well as how to conveniently describe this learning method mathematically by only defining a modified kernel in detail in Section 3.2.

Remark 1.2.5. To deal with questions regarding measurability of SVMs, or general learning methods, one usually considers the so-called universal completion of the Borel σ-algebra on $(X \times Y)^n$, see for example the introduction of Section 6.1 in [55]. To this end, recall that the completion of a measurable space $(\Omega, \mathcal{F})$ with respect to a measure μ on $\mathcal{F}$ is defined as the smallest σ-algebra that contains $\mathcal{F}$ and all subsets of every μ-zero set. The universal completion of $\mathcal{F}$ is then defined as the intersection of all μ-completions, where μ runs through all probability measures on $\mathcal{F}$. Using this universal completion and a separable RKHS H, all problems concerning measurability of SVMs are then answered by [55, Lemma 6.23]. In other words, we can ignore measurability conditions by implicitly assuming that we are given the universal completion of the Borel σ-algebra on $(X \times Y)^n$ and using a separable RKHS H, which we will do for the rest of this thesis.

The probably most popular kernel used for SVMs is the Gaussian kernel, which is also the kernel we will focus on in this thesis.

Definition 1.2.6. Given some non-empty set $X \subset \mathbb{R}^d$ and some $\gamma > 0$, the Gaussian RKHS, denoted by $H_\gamma(X)$, is the RKHS associated to the Gaussian kernel

$$k_\gamma(x, y) := \exp(-\gamma^{-2}\|x - y\|^2), \quad x, y \in X.$$

The parameter γ is called bandwidth.

The Gaussian kernel has been shown to yield superior performance in practical applications, making it the default kernel in the majority of SVM implementations.

Also, from a theoretical point of view the Gaussian kernel has its merits as it is well-studied in the literature. For example, in [59] an explicit description of the functions contained in the Gaussian RKHS as well as a countable orthonormal basis are derived, showing the separability of $H_\gamma(X)$. Other useful properties of the Gaussian RKHS, which will be important later, are summarized Section 2.3. The Gaussian kernel not only plays an outstanding role in machine learning applications but also in mathematical physics, for example in the derivation of the fundamental solution of the heat equation, see [17, Section 2.3].

As an SVM using a Gaussian kernel has an additional hyperparameter γ, we denote the decision function of the learning method (1.6) with $H = H_\gamma(X)$ by $f_{D,\lambda,\gamma}$, that is

$$f_{D,\lambda,\gamma} = \operatorname*{arg\,min}_{f \in H_\gamma(X)} \lambda \|f\|^2_{H_\gamma(X)} + \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)). \tag{1.8}$$

To deal with some technical issues connected to unbounded loss functions in the subsequent statistical analysis of Gaussian SVMs, we also need to introduce the clipping operation, see [55, Definition 2.22]. To this end, we say that a loss function $L : Y \times \mathbb{R} \to [0, \infty)$ can be clipped at $M > 0$, if $L(y, \widehat{t}) \leq L(y, t)$ for all $y \in Y$ and $t \in \mathbb{R}$, where

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

is the clipped value of t at M. Then, for a Gaussian SVM using a loss that can be clipped at M, we will apply the clipping operation point wise to $f_{D,\lambda,\gamma}$ and denote the resulting decision function by $\widehat{f}_{D,\lambda,\gamma}$.

Example 1.2.7. Let $L = L_{\mathrm{LS}}$ be the least-squares loss and assume that the output space $Y \subset [-M, M]$ is bounded by some $M > 0$. Then the least-squares loss can be clipped at M, since we obviously have $(y - \widehat{t})^2 \leq (y - t)^2$ for all $|y| \leq M$, where $\widehat{\cdot}$ denotes the clipped value at M.

Example 1.2.8. Let $Y = \{-1, 1\}$. The hinge loss $L = L_{\mathrm{hinge}}$ can be clipped at $M = 1$, since we have $\max\{0, 1 - y\widehat{t}\} \leq \max\{0, 1 - yt\}$ for all $y \in Y, t \in \mathbb{R}$.

As usual, the optimal choice for the hyperparameters λ and γ depends on char-

acteristics of the unknown distribution $\mathbf{P}$. Therefore, we will also consider a training validation procedure for data dependent hyperparameter selection. To this end, we split our dataset $D = ((x_1, y_1), \ldots, (x_n, y_n))$ into a training set $D_1 := ((x_1, y_1), \ldots, (x_l, y_l))$ and a validation set $D_2 := ((x_{l+1}, y_{l+1}), \ldots, (x_n, y_n))$, where $l := \lfloor n/2 \rfloor + 1$. Further, we fix finite sets Λ_n and Γ_n of candidate values for λ and γ. Then, we compute the SVM decision functions $\widehat{f}_{D_1,\lambda,\gamma}$ for all $(\lambda, \gamma) \in \Lambda_n \times \Gamma_n$ using the training set D_1 and for the final estimator we pick the parameters $\lambda_{D_2} \in \Lambda_n, \gamma_{D_2} \in \Gamma_n$ which have the best empirical error on the validation set D_2, that is

$$\sum_{i=l+1}^{n} L\left(y_i, \widehat{f}_{D_1,\lambda_{D_2},\gamma_{D_2}}(x_i)\right) = \min_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \sum_{i=l+1}^{n} L\left(y_i, \widehat{f}_{D_1,\lambda,\gamma}(x_i)\right).$$

We call the resulting learning method a *training validation support vector machine* (TV-SVM). In the subsequent chapters we will show that it is sufficient for the candidate sets Λ_n and Γ_n to grow logarithmically with the sample size n in order to achieve optimal rates adaptively, which is a significant improvement compared to previous results, which required the size of the candidate sets to grow at least linearly in n, see for example [55, Theorem 7.24].

1.3 Tools for the Statistical Analysis

A main ingredient for our statistical analysis is a so-called variance bound, which intuitively guarantees a small variance of the excess risk whenever our estimator is close to the optimum, see for example [8, Section 5.2] for a more detailed discussion of the effects of a variance bound on the statistical properties of the excess risk. In the following definition, for a loss function $L : Y \times \mathbb{R} \to [0, \infty)$ and a function $f : X \to \mathbb{R}$, we use the symbol $L \circ f$ to denote the function $(x, y) \mapsto L(y, f(x))$.

Definition 1.3.1. Let L be a loss that can be clipped at $M > 0$ and let $\mathcal{F}$ be some function class of measurable functions $f : X \to \mathbb{R}$. Assume there exists a Bayes decision function $f^*_{L,\mathbf{P}} : X \to [-M, M]$. We say, that the supremum bound is satisfied, if there exists a constant $B > 0$, such that $L(y, t) \leq B$ for all $(y, t) \in Y \times [-M, M]$. We further say, that the variance bound is satisfied, if there exist $\vartheta \in [0, 1]$ and $V \geq B^{2-\vartheta}$, such that

$$\mathbf{E}(L \circ \widehat{f} - L \circ f^*_{L,\mathbf{P}})^2 \leq V \cdot \left(\mathbf{E}\, L \circ \widehat{f} - L \circ f^*_{L,\mathbf{P}}\right)^{\vartheta} \quad \text{for all } f \in \mathcal{F}.$$

Example 1.3.2. Let $L = L_{\mathrm{LS}}$ be the least-squares loss and assume that the output space is bounded $Y \subset [-M, M]$ by some $M > 0$. Then the variance bound is satisfied for the best possible exponent $\vartheta = 1$ and $V = 16M^2$. The supremum bound is satisfied for $B = 4M^2$. For details on the derivation we refer to [55, Example 7.3].

As the example above shows, a significant advantage of the least-squares loss is that the variance bound is satisfied for the best exponent for all distributions with bounded output space. For the hinge loss, this problem is quite more intricate. Non-trivial variance bounds can only be established under some assumptions on the distribution $\mathbf{P}$, which we will summarize in the following. To this end, recall the definition of the conditional class probability $\eta : X \to [0, 1]$ in Example 1.1.2.

Assumption 1.3.3. There exist constants $C_* > 0$ and $q \in [0, \infty]$ such that

$$\mathbf{P}_X\big(\{x \in X : |2\eta(x) - 1| < t\}\big) \leq (C_* t)^q$$

for all $t \geq 0$, where we use the convention $t^\infty = 0$ for $t \in (0, 1)$.

Assumption 1.3.3 is in the literature widely known as Tsybakov noise condition. It was first introduced in [38] and since then has become a standard regularity assumption in non-parametric binary classification. Intuitively, Assumption 1.3.3 restricts the mass of points $x \in X$ such that $\eta(x)$ is close to $1/2$ and evidently it is hard to predict the label of $x \in X$ with high probability whenever $\eta(x) \approx 1/2$. Assumption 1.3.3 can be used to establish a non-trivial variance bound, which we record in the following example.

Example 1.3.4. Assume $\mathbf{P}$ satisfies Assumption 1.3.3 for the constant $C_* > 0$ and exponent $q \in [0, \infty]$. Then, the hinge loss satisfies the variance bound for $\vartheta = q/(q + 1)$ and $V = 6C_*^{q/(q+1)}$, see [55, Theorem 8.24]. The supremum bound is obviously satisfied for $B = 2$.

Finally, we have to introduce one last regularity condition we need to impose on our loss functions.

Definition 1.3.5. A loss function $L : Y \times \mathbb{R} \to [0, \infty)$ is called locally Lipschitz continuous if for every $a > 0$ the functions $L(y, \cdot)|_{[-a,a]}, y \in Y$ are uniformly Lipschitz

continuous, that is

$$|L|_{a,1} := \sup_{\substack{s,t\in[-a,a],s\neq t\\ y\in Y}} \frac{|L(y,t)-L(y,s)|}{|t-s|} < \infty.$$

Example 1.3.6. Let $L = L_{\mathrm{LS}}$ be the least-squares loss and assume that the output space is bounded $Y \subset [-M, M]$ by some $M > 0$. A simple application of the mean value theorem shows that

$$|L|_{M,1} \leq \sup_{\substack{s,t\in[-M,M],s\neq t\\ y\in Y}} \frac{\mathrm{d}}{\mathrm{d}t}(t-y)^2 = \sup_{\substack{s,t\in[-M,M],s\neq t\\ y\in Y}} 2(t-y) \leq 4M,$$

and we conclude that the least-squares loss is locally Lipschitz continuous.

Example 1.3.7. The hinge loss is locally Lipschitz continuous with $|L_{\mathrm{hinge}}|_{1,1} = 1$.

For the analysis of the statistical error of our considered learning methods we need to introduce a number of covering quantities which, roughly speaking, quantify how well a set can be approximated by a finite number of points.

Definition 1.3.8. Given a normed space E and a subset $A \subset E$ we say that the points $x_1, \ldots, x_m \in E$ are an ε-net of A, if

$$A \subset \bigcup_{j=1}^{m} (x_j + \varepsilon B_E).$$

Given an $\varepsilon > 0$ the covering number $\mathcal{N}_E(A, \varepsilon)$ of A is defined as the minimum cardinality of an ε-net of A. We may also write $\mathcal{N}(A, \varepsilon) := \mathcal{N}_E(A, \varepsilon)$, if the ambient space E is known from the context. Finally, given a second normed space F and a bounded, linear operator $T : E \to F$, the covering numbers of T are defined by $\mathcal{N}(T, \varepsilon) := \mathcal{N}_F(TB_E, \varepsilon)$.

Note that a set A is precompact if and only if $\mathcal{N}(A, \varepsilon) < \infty$ for all $\varepsilon > 0$. Instead of fixing an $\varepsilon > 0$ and minimizing the number of ε-balls necessary to cover a set, we can also fix the number of balls and minimize the (common) radius of the balls. This is the principle of the related concept of (dyadic) entropy numbers, which we will only introduce for operators.

Definition 1.3.9. Given normed spaces E, F and a bounded, linear operator $T : E \to F$, for $i \in \mathbb{N}$ the i-th dyadic entropy number of T is defined as

$$e_i(T) := \inf \left\{ \varepsilon > 0 : \exists x_1, \dots, x_{2^{i-1}} \in F \text{ such that } TB_E \subset \bigcup_{j=1}^{2^{i-1}} (x_j + \varepsilon B_F) \right\}.$$

One can easily check that a bounded, liner operator T is compact if and only if $e_i(T)$ converges to 0 as $i \to \infty$. In this regard, entropy numbers provide a quantitative notion of compactness in the sense that an operator $T : E \to F$ is *more compact* than an operator $S : E \to F$ if $e_i(T)$ converges to 0 faster than $e_i(S)$. Also, the entropy numbers of an operator T are related to approximation properties of T by finite rank operators. Note that in the definition of $e_i(T)$ ε-nets of cardinality 2^{i-1} are considered. The reason for this is that for non-trivial compact operators between *infinite* dimensional spaces E and F, which are commonly considered, the non-dyadic entropy numbers usually converge to 0 slower than any polynomial and there is no loss of information in switching to dyadic entropy numbers, see [9, Section 1.3]. A basic, yet useful property of entropy numbers, which we will use throughout this thesis, is that entropy numbers are dominated by decompositions, that is, if we can decompose $T : E \to F$ into $T = RS$ with bounded, linear operators $S : E \to \tilde{F}$, $R : \tilde{F} \to F$, and an intermediate normed space $\tilde{F}$, then we have $e_i(T) \leq \|R\| e_i(S)$ as well as $e_i(T) \leq e_i(R)\|S\|$ for all $i \in \mathbb{N}$, where $\|R\|$, $\|S\|$ denotes the operator norm, see [9, p. 21]. As one would expect, there is a close connection between entropy and covering numbers.

Lemma 1.3.10. *Let E, F be normed spaces and let $T : E \to F$ be a bounded, linear operator.*

(i) *If there exist constants $a > 0$ and $q > 0$ such that $e_i(T) \leq a\, i^{-1/q}$ for all $i \in \mathbb{N}$, then we have*

$$\log \mathcal{N}(T, \varepsilon) \leq \log 4 \left(\frac{a}{\varepsilon}\right)^q$$

for all $\varepsilon > 0$.

(ii) *If there exist constants $a > 0$ and $q > 0$ such that $\log \mathcal{N}(T, \varepsilon) \leq (a/\varepsilon)^q$ for all $\varepsilon > 0$, then we have*

$$e_i(T) \leq 3^{\frac{1}{q}} a\, i^{-\frac{1}{q}}$$

for all $i \in \mathbb{N}$.

The first assertion is the statement of [55, Lemma 6.21], the second assertion is the content of [55, Exercise 6.8].

In the statistical analysis of Chapters 2 and 3 we have to bound the expectation of the entropy numbers of the embedding $\mathrm{id} : H \to L_2(\mathbf{D})$, where H is an RKHS on X and $\mathbf{D} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical distribution associated to a sample $D = (x_1, \ldots, x_n)$ drawn from $\mathbf{P}_X^n$. More specifically, we will be interested in bounds of the form

$$\mathbf{E}_{D\sim\mathbf{P}_X^n} e_i(\mathrm{id} : H \to L_2(\mathbf{D})) \leq a\, i^{-\frac{1}{2p}} \quad \text{for all } i \in \mathbb{N} \tag{1.9}$$

for some constants $a > 0$ and $p \in (0, 1)$. Such bounds in turn can be used to bound Rademacher averages of H using a standard symmetrization procedure and Dudley's chaining, see [55, Section 7.3]. In the following, we collect some preliminary results on RKHSs and their covering numbers that will help us to derive appropriate bounds of the form (1.9) in Chapter 2 and 3. The subsequent results appeared in [24].

Lemma 1.3.11. *Let k be a kernel on X with RKHS H and let $\psi : Y \to X$ be a map. Then $k_\psi(\cdot,\cdot) := k(\psi(\cdot), \psi(\cdot))$ is a kernel on Y with RKHS $H_\psi = \{f \circ \psi : f \in H\}$ and the map $V : H \to H_\psi$ defined by $f \mapsto f \circ \psi$ is a metric surjection. The norm in H_ψ can be computed by*

$$\|g\|_{H_\psi} = \inf\{\|f\|_H : f \text{ with } g = f \circ \psi\}.$$

If ψ is bijective, then V is an isometric isomorphism.

Proof. Let $\Phi : X \to H, x \mapsto k(x,\cdot)$ be the canonical feature map of k and define $\Phi_\psi : Y \to H, y \mapsto \Phi(\psi(y))$. Then by construction we have $\langle \Phi_\psi(y), \Phi_\psi(y')\rangle_H = k_\psi(y, y')$ for all $y, y' \in Y$, that is, Φ_ψ is a feature map of k_ψ. The first two assertions now follow from (1.5). For the third assertion additionally apply this result on ψ^{-1}. □

Corollary 1.3.12. *Let k be a kernel on $X \subset \mathbb{R}^d$, H its RKHS, and $Y \subset X$. Then $H|_Y := \{f|_Y : f \in H\}$ is the RKHS of $k|_{Y\times Y}$ and the restriction $H \to H|_Y$ is a metric surjection.*

Proof. This follows from Lemma 1.3.11 with $\psi : Y \to X$ being the inclusion. □

A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called radial if there exists a function $\kappa : [0,\infty) \to \mathbb{R}$ such that $k(x, y) = \kappa(\|x - y\|)$ for all $x, y \in \mathbb{R}^d$. Radial kernels are a special case of translation invariant kernels, which by Bochner's theorem [47, Theorem IX.9]

are characterized as the inverse Fourier transform of a finite Borel measure on $\mathbb{R}^d$. Similarly, by Schoenberg's theorem [51] $\kappa : [0,\infty) \to \mathbb{R}$ defines a radial kernel on $\mathbb{R}^d$ via $k(\cdot,\cdot) = \kappa(\|\cdot - \cdot\|)$ for every $d \in \mathbb{N}$, if and only if there exists a finite Borel measure μ on $[0,\infty)$ such that

$$\kappa(t) = \int_0^\infty \mathrm{e}^{-xt^2} \,\mathrm{d}\mu(x)$$

for all $t \in [0,\infty)$, i.e. radial kernels are mixtures of Gaussians. The following corollary shows, that RKHSs of radial kernels are in some sense translational and rotational invariant.

Corollary 1.3.13. *Let k be a radial kernel on $\mathbb{R}^d$ and for $X \subset \mathbb{R}^d$ denote the restriction of k onto $X \times X$ by k_X and its RKHS by $H(X)$. Fix an $a \in \mathbb{R}^d$ and an orthogonal matrix $U \in \mathbb{R}^{d\times d}$. Then the operator $T : H(X) \to H(a + UX)$ defined by $Tf(x) = f(U^{-1}(x-a))$ is well-defined and an isometric isomorphism.*

Proof. This follows from Lemma 1.3.11 with the map $\psi : a + UX \to X$ defined by $x \mapsto U^{-1}(x-a)$, since $k_X(\psi(\cdot),\psi(\cdot)) = k_{a+UX}(\cdot,\cdot)$. □

Lemma 1.3.14. *Let k be a kernel on X, H be its RKHS, and $X_1,\dots,X_N \subset X$ pairwise disjoint subsets with $X_1 \cup \ldots \cup X_N = X$. Then for all $\varepsilon > 0$ we have*

$$\mathcal{N}_{\ell_\infty(X)}(B_H,\varepsilon) \le \prod_{k=1}^N \mathcal{N}_{\ell_\infty(X_k)}(B_{H|_{X_k}},\varepsilon).$$

Proof. As the general statement easily follows inductively, we will only prove the case $N = 2$. To this end, let $f_1,\dots,f_n \in \ell_\infty(X_1)$ be a minimal ε-net of $B_{H|_{X_1}}$ and let $g_1,\dots,g_m \in \ell_\infty(X_2)$ be a minimal ε-net of $B_{H|_{X_2}}$. Let $f \in B_H$. Then by Corollary 1.3.12 we have $f|_{X_l} \in H|_{X_l}$ with $\|f|_{X_l}\|_{H|_{X_l}} \le \|f\|_H \le 1$ for $l = 1,2$. Hence, there exist i,j with $\|f|_{X_1} - f_i\|_{\ell_\infty(X_1)} \le \varepsilon$ and $\|f|_{X_2} - g_j\|_{\ell_\infty(X_2)} \le \varepsilon$. If we denote the zero-extensions of f_i, g_j to X by $\widehat{f}_i, \widehat{g}_j$, we see that $\{\widehat{f}_i + \widehat{g}_j : i = 1,\dots,n,\ j = 1,\dots,m\}$ is an ε-net of B_H with cardinality $n \cdot m$. □

Corollary 1.3.15. *Let $X \subset \mathbb{R}^d$. For all $\gamma > 0$ and $\varepsilon > 0$ we have*

$$\begin{aligned}&\log \mathcal{N}(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X),\varepsilon)\\ \le &\mathcal{N}(X,\gamma) \cdot \log \mathcal{N}\left(\mathrm{id} : H_1(B) \to \ell_\infty(B),\varepsilon\right),\end{aligned}$$

where $\mathcal{N}(X,\gamma)$ denotes the covering numbers with respect to an arbitrary norm on $\mathbb{R}^d$ and B denotes the closed unit ball in $\mathbb{R}^d$ with respect to that norm.

Proof. Let $x_1, \dots, x_n \in \mathbb{R}^d$ be a minimal γ-net of X with respect to a given norm. We partition X into $X_1, \dots, X_n$, where X_j consists of the points $x \in X$ that are closest to x_j with respect to the given norm. Here we break ties, for example, in favor of a smaller index j. Combining Lemma 1.3.14, Corollary 1.3.12 and Corollary 1.3.13 we get

$$\begin{aligned}\log \mathcal{N}_{\ell_\infty(X)}\left(B_{H_\gamma(X)}, \varepsilon\right) &\leq \sum_{j=1}^{n} \log \mathcal{N}_{\ell_\infty(X_j)}\left(B_{H_\gamma(X_j)}, \varepsilon\right) \\ &\leq \sum_{j=1}^{n} \log \mathcal{N}_{\ell_\infty(x_j+\gamma B)}\left(B_{H_\gamma(x_j+\gamma B)}, \varepsilon\right) \\ &= n \log \mathcal{N}_{\ell_\infty(\gamma B)}\left(B_{H_\gamma(\gamma B)}, \varepsilon\right),\end{aligned}$$

where B denotes the closed unit ball in $\mathbb{R}^d$ with respect to the given norm. The result now follows from [55, Proposition 4.37], which states that the scaling operator $\tau_\gamma : H_\gamma(\gamma B) \to H_1(B)$ defined by $\tau_\gamma f(x) = f(\gamma x)$ is an isometric isomorphism. □

In essence, Corollary 1.3.15 reduces the problem of bounding the ℓ_∞-covering numbers of $H_\gamma(X)$ to bounding the ℓ_∞-covering numbers of $H_1(B)$. The latter are well understood, [34, Theorem 3] showed that[1]

$$\log \mathcal{N}(\mathrm{id} : H_1(B) \to \ell_\infty(B), \varepsilon) \asymp \frac{\left(\log \frac{1}{\varepsilon}\right)^{d+1}}{\left(\log \log \frac{1}{\varepsilon}\right)^{d}} \quad \text{as} \quad \varepsilon \to 0. \tag{1.10}$$

for any bounded $B \subset \mathbb{R}^d$ with non-empty interior. However, we will rely on the slightly suboptimal bound

$$\log \mathcal{N}(\mathrm{id} : H_1(B) \to \ell_\infty(B), \varepsilon) \lesssim \log^{d+1} \frac{1}{\varepsilon} \quad \text{as} \quad \varepsilon \to 0, \tag{1.11}$$

since it is very hard to make use of the extra double logarithmic factor in (1.10).

Theorem 1.3.16. *There exists a universal constant K_d only depending on d, such that*

$$e_i(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X)) \leq K_d^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \mathcal{N}(X, \gamma)^{\frac{1}{2p}} i^{-\frac{1}{2p}}$$

holds for all $i \in \mathbb{N}$, $p \in (0,1)$ and $\gamma > 0$.

[1] The result is actually only stated for $B = [0,1]^d$, but the generalization is straightforward.

Proof. For $f \in B_{H_1(B)}$ by [55, Lemma 4.23] we have $\|f\|_\infty \leq 1$ and consequently we find $\mathcal{N}_{\ell_\infty(B)}(B_{H_1(B)}, \varepsilon) = 1$ for all $\varepsilon \geq 1$. Furthermore, by [34, Theorem 3] there exists a constant $K \geq 1$ such that

$$\log \mathcal{N}_{\ell_\infty(B)}\left(B_{H_1(B)}, \varepsilon\right) \leq K \log^{d+1}\frac{2}{\varepsilon}$$

for all $\varepsilon \in (0,1]$. Some elementary calculations show that

$$\sup_{\varepsilon \in (0,2)} \varepsilon^q \log^{d+1}\frac{2}{\varepsilon} = 2^q \left(\frac{d+1}{eq}\right)^{d+1},$$

which combined with Corollary 1.3.15 and the estimate above gives us

$$\begin{aligned} \log \mathcal{N}(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X), \varepsilon) &\leq K\, \mathcal{N}(X,\gamma) \log^{d+1}\frac{2}{\varepsilon} \\ &\leq 4K\, \mathcal{N}(X,\gamma) \left(\frac{d+1}{eq}\right)^{d+1} \varepsilon^{-q} \end{aligned} \tag{1.12}$$

for $\varepsilon > 0$ and $q \in (0,2)$. As a final step we convert the latter bound on the covering numbers of $\mathrm{id} : H_\gamma(X) \to \ell_\infty(X)$ into a bound on the entropy numbers. To this end, we fix an $i \geq 2$ and define $\varepsilon > 0$ by $\exp(a/\varepsilon^q) = 2^{i-1}$, where

$$a := 4K\, \mathcal{N}(X,\gamma) \left(\frac{d+1}{eq}\right)^{d+1}.$$

By (1.12) this implies

$$e_i(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X)) \leq \left(\frac{(i-1)\log 2}{a}\right)^{-\frac{1}{q}} \leq \left(\frac{2a}{\log 2}\right)^{\frac{1}{q}} i^{-\frac{1}{q}}$$

for all $i \geq 2$. Since $e_1(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X)) \leq 1$ we get

$$e_i(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X)) \leq \left(\frac{\mathcal{N}(X,\gamma) 8K}{\log 2} \left(\frac{d+1}{eq}\right)^{d+1}\right)^{\frac{1}{q}} i^{-\frac{1}{q}}$$

for all $i \in \mathbb{N}$. Now substitute $2p = q$ and absorb all irrelevant constants into a constant K_d. □

The following theorem, which is the content of [55, Theorem 7.23], states an oracle inequality for SVMs using a kernel satisfying (1.9) for the given distribution $\mathbf{P}$.

Theorem 1.3.17. *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a locally Lipschitz continuous loss that can be clipped at $M > 0$ and satisfies the supremum bound for a $B > 0$. Moreover, let H be a separable RKHS over X and $\mathbf{P}$ be a distribution over $X \times Y$ such that a variance bound is satisfied for constants $\vartheta \in [0,1], V \geq B^{2-\vartheta}$, and all $f \in H$. Assume that for fixed $n \geq 1$ there exist constants $p \in (0,1)$ and $a \geq B$ such that (1.9) is satisfied. Finally, fix an $f_0 \in H$ and a constant $B_0 \geq B$ such that $\|L \circ f_0\|_\infty \leq B_0$. Then, for all $\tau \geq 0$ and $\lambda > 0$, the SVM using H and L satisfies*

$$\lambda \|f_{D,\lambda}\|_H^2 + \mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda}) - \mathcal{R}^*_{L,\mathbf{P}} \leq 9(\lambda \|f_0\|_H^2 + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}}) \\ K\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}$$

with probability not less than $1 - 3\mathrm{e}^{-\tau}$, where $K \geq 1$ is a constant only depending on p, M, B, ϑ, and V.

For our later applications of this theorem we need to take a closer look at the constant K and especially how it depends on the given parameters. First of all, the constant K is given by (see [55, proof of Theorem 7.23])

$$\begin{aligned} K &= \max\left\{2B, 3\left(30 \cdot 2^p C_1(p) |L|_{M,1}^p V^{\frac{1-p}{2}}\right)^{\frac{2}{2-p-\vartheta+\vartheta p}}, 90 \cdot 120^p C_2^{1+p}(p) |L|_{M,1}^p B^{1-p}\right\} \\ &\leq \max\left\{B, \left(|L|_{M,1}^p V^{\frac{1-p}{2}}\right)^{\frac{2}{2-p-\vartheta+\vartheta p}}, |L|_{M,1}^p B^{1-p}, 1\right\} \\ &\quad \cdot \max\left\{2, 3\left(30 \cdot 2^p C_1(p)\right)^{\frac{2}{2-p-\vartheta+\vartheta p}}, 90 \cdot 120^p C_2^{1+p}(p)\right\}. \end{aligned}$$

The constants $C_1(p)$ and $C_2(p)$ are given by

$$C_1(p) := \frac{2\sqrt{\log 256} C_p^p}{(\sqrt{2}-1)(1-p)2^{\frac{p}{2}}}, \quad C_2(p) = \left(\frac{8\sqrt{\log 16} C_p^p}{(\sqrt{2}-1)(1-p)4^p}\right)^{\frac{2}{1+p}},$$

where

$$C_p = \frac{\sqrt{2}-1}{\sqrt{2}-2^{\frac{2p-1}{2p}}} \cdot \frac{1-p}{p},$$

which can be tracked in [55, proof of Theorem 7.16]. It was shown in [20, Proof of Theorem 7] that $C_1(p)$ and $C_2^{1+p}(p)$ are uniformly bounded in $p \in (0, \frac{1}{2}]$. More

precisely, we have

$$\sup_{p\in(0,\frac{1}{2}]} C_1(p) \le 46\mathrm{e} \quad \text{and} \quad \sup_{p\in(0,\frac{1}{2}]} C_2^{1+p}(p) \le 1035\mathrm{e}^2,$$

which implies for all $p \in (0, 1/2]$ and $\vartheta \in [0, 1]$ that

$$K \le \tilde{K} \max\left\{B, \left(|L|_{M,1}^p V^{\frac{1-p}{2}}\right)^{\frac{2}{2-p-\vartheta+\vartheta p}}, |L|_{M,1}^p B^{1-p}, 1\right\} \tag{1.13}$$

with a constant $\tilde{K}$ independent of p and ϑ.

2 Learning Rates for SVMs

In this chapter we present our results on learning rates for SVMs using a (global) Gaussian kernel. Our notion of intrinsic dimensionality is based on the box-counting dimension, which we introduce in Section 2.1 including some illustrative examples. Afterwards, in Section 2.2, we present an oracle inequality for Gaussian SVMs for general loss functions under the assumptions introduced in the previous section. Finally, Sections 2.3 and 2.4 contain our results on least-squares regression and binary classification, respectively. The contents of this chapter were published in [24].

2.1 Intrinsic Dimension Assumption

The main concept in this chapter to describe the intrinsic dimension of the data is based on the upper box-counting dimension of the support of the data generating distribution $\mathbf{P}_X$. To this end, recall that the support of a Borel measure μ, denoted by $\operatorname{supp}\mu$, is defined as the complement of the largest open μ-zero set. Also, recall the definition of covering numbers in Definition 1.3.8.

Assumption 2.1.1. There exist constants $C_{\mathrm{box}} > 0$ and $\varrho > 0$ such that for all $\varepsilon \in (0,1)$ we have

$$\mathcal{N}_{\ell_\infty^d}(\operatorname{supp}\mathbf{P}_X, \varepsilon) \leq C_{\mathrm{box}}\varepsilon^{-\varrho}.$$

The infimum over all ϱ, such that Assumption 2.1.1 is fulfilled for ϱ and some finite constant C_{box} coincides with the so-called upper box-counting dimension of $\operatorname{supp}\mathbf{P}_X$, which is defined as

$$\limsup_{\varepsilon\to 0} \frac{\log \mathcal{N}_{\ell_\infty^d}(\operatorname{supp}\mathbf{P}_X, \varepsilon)}{\log \frac{1}{\varepsilon}}, \tag{2.1}$$

cf. [18, Section 3.1]. Analogously, the lower box-counting dimension is defined by substituting $\limsup$ with $\liminf$ in (2.1) and in case those values coincide, this

common limit is the box-counting dimension of $\operatorname{supp} \mathbf{P}_X$. Also note, that we can consider in Assumption 2.1.1 the covering numbers with respect to the ℓ_p^d-norm for any $p \in [1, \infty)$, as a change of the norm will only influence the constant C_{box}, but not ϱ. The following examples give some intuition on Assumption 2.1.1 and, in ascending order, demonstrate the generality of this assumption.

Example 2.1.2. Let $X = [-1, 1]^d$. By a simple argument we see, that for $\varepsilon = 1/m$, where $m \in \mathbb{N}$, we have $\mathcal{N}_{\ell_\infty^d}(X, \varepsilon) = \varepsilon^{-d}$. Moreover, by [9, Proposition 1.3.1] there exist constants $c, C > 0$, such that

$$c\varepsilon^{-d} \leq \mathcal{N}_{\ell_\infty^d}(X, \varepsilon) \leq C\varepsilon^{-d}$$

for all $\varepsilon \in (0, 1)$. That is, Assumption 2.1.1 is fulfilled exactly for $\varrho = d$. More generally we have for any bounded $X \subset \mathbb{R}^d$ with non-empty interior that Assumption 2.1.1 is fulfilled exactly for $\varrho = d$.

Example 2.1.3. Let $X \subset \mathbb{R}^d$ be a bounded d'-dimensional differentiable manifold. Then 2.1.1 is fulfilled for $\varrho = d'$. This follows from Example 2.1.2 and the fact, that the box-counting dimension is invariant under bi-Lipschitzian maps, cf. [18, Section 3.2]. Our assumption 2.1.1 therefore includes the manifold assumption commonly used in the literature.

Example 2.1.4. The *attractor* of a dynamical system is, loosely speaking, a set in the phase space of the dynamical system to which it tends to converge to, based on the initial conditions [42]. It is not unusual for such attractors of dynamical systems describing physical systems to exhibit a fractal structure, whose complexity is, amongst others, measured by their box-counting dimension [19]. A famous example is given by the Lorenz attractor associated to the dynamical system

$$\begin{aligned} x' &= -\sigma x + \sigma y \\ y' &= -xy + rx - y \\ z' &= xy - bz \end{aligned}$$

for real parameters σ, r, b, and was originally used to describe atmospherical convection, see [50] as well as for other examples. The Lorenz attractor is estimated to have a box-counting dimension of approximately 1.98 for certain values of σ, r, b, see [39]. This shows, that our assumptions allowing for non-integer dimensions is not

only a mathematical quirk, but is also relevant for real-world datasets. Suppose, for example, the feature vectors x_i of the dataset D are generated by observing the state of such a dynamical systems at independent, random time steps. To give a concrete example, following the discussion above, it is reasonable to assume that a dataset containing meteorological data has some low-dimensional intrinsic fractal structure, as in [12] where the authors propose an approach for estimating air pollution based on meteorological and pollution data from distant sensor stations. For further examples of fractal structures in mathematical models in physics, chemistry, and finance we refer to [18, Chapter 18].

2.2 A General Oracle Inequality

The following proposition relates Assumption 2.1.1 to the averaged entropy numbers (1.9) and is the central result for bounding the statistical error of Gaussian SVMs under Assumption 2.1.1.

Proposition 2.2.1. *Let $\mathbf{P}$ satisfy Assumption 2.1.1. Then there exists a constant K_d only depending on d, such that the bound*

$$\mathbf{E}_{D\sim\mathbf{P}_X^n} e_i(\mathrm{id} : H_\gamma(X) \to L_2(\mathbf{D})) \leq (C_{\mathrm{box}} K_d)^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \gamma^{-\frac{\varrho}{2p}} i^{-\frac{1}{2p}}$$

holds for all $i \in \mathbb{N}$, $p \in (0,1)$ and $\gamma \in (0,1)$.

Proof. Consider the decomposition of $\mathrm{id} : H_\gamma(X) \to L_2(\mathbf{D})$ for a sample $D \in (\operatorname{supp} \mathbf{P}_X)^n$ described by the commutative diagram

$$\begin{array}{ccc} H_\gamma(X) & \xrightarrow{\quad\mathrm{id}\quad} & L_2(\mathbf{D}) \\ \Big\downarrow{\scriptstyle \mathrm{res}} & & \Big\uparrow{\scriptstyle \mathrm{id}} \\ H_\gamma(\operatorname{supp} \mathbf{P}_X) & \xrightarrow[\quad\mathrm{id}\quad]{} & \ell_\infty(\operatorname{supp} \mathbf{P}_X) \end{array}$$

where $\mathrm{res} : H_\gamma(X) \to H_\gamma(\operatorname{supp} \mathbf{P}_X)$ is the restriction operator. We have

$$\|\mathrm{res} : H_\gamma(X) \to H_\gamma(\operatorname{supp}, \mathbf{P}_X)\| \leq 1$$

by Corollary 1.3.12 and trivially also $\|\mathrm{id} : \ell_\infty(\operatorname{supp} \mathbf{P}_X) \to L_2(\mathbf{D})\| \leq 1$. Theorem 1.3.16 then implies

$$e_i(\mathrm{id} : H_\gamma(X) \to L_2(\mathbf{D})) \leq K_d^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \mathcal{N}_{\ell_\infty^d}(\operatorname{supp} \mathbf{P}_X, \gamma)^{\frac{1}{2p}} \, i^{-\frac{1}{2p}}$$

for all $i \in \mathbb{N}$, $\gamma > 0$ and $p \in (0,1)$ for $\mathbf{P}_X^n$-almost all $D \in X^n$. Now combine the above bound with Assumption 2.1.1 and take the expectation w.r.t. $D \sim \mathbf{P}_X^n$. □

The following theorem states an oracle inequality for Gaussian SVMs using general loss functions under Assumption 2.1.1 which is the basis for our results in Sections 2.3 and 2.4.

Theorem 2.2.2. *Assume L is a locally Lipschitz continuous loss that can be clipped at $M > 0$ and that the supremum and variance bounds are satisfied for constants $B > 0, \vartheta \in [0,1]$, and $V \geq B^{2-\vartheta}$. Furthermore, assume $\mathbf{P}_X$ satisfies Assumption 2.1.1 for $C_{\mathrm{box}} > 0$ and $\varrho > 0$ and fix an $f_0 \in H_\gamma(X)$ and a $B_0 \geq B$ with $\|L \circ f_0\|_\infty \leq B_0$. Then there exists a constant K such that for all $n \in \mathbb{N}, \gamma \in (0,1), \lambda > 0, p \in (0, 1/2]$ and $\tau > 0$ we have*

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq{}& 9(\lambda \|f_0\|^2_{H_\gamma(X)} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}}) \\ &+ C_{\mathbf{P}} K \left(\frac{p^{-d-1} \gamma^{-\varrho}}{\lambda^p n} \right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3 \left(\frac{72 V \tau}{n} \right)^{\frac{1}{2-\vartheta}} + \frac{15 B_0 \tau}{n} \end{aligned} \tag{2.2}$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$, where K is independent of $\mathbf{P}$ and

$$C_{\mathbf{P}} = \max\left\{ B, \left(|L|^p_{M,1} V^{\frac{1-p}{2}} \right)^{\frac{2}{2-p-\vartheta+\vartheta p}}, |L|^p_{M,1} B^{1-p}, 1 \right\} \cdot \max\left\{ C_{\mathrm{box}}, B^{2p} \right\}^{\frac{1}{2-p-\vartheta+\vartheta p}}.$$

Proof. By Theorem 1.3.17 in combination with the entropy estimate from Proposition 2.2.1 we have

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq{}& 9(\lambda \|f_0\|^2_{H_\gamma(X)} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}}) \\ &+ K \left(\frac{a^{2p}}{\lambda^p n} \right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3 \left(\frac{72 V \tau}{n} \right)^{\frac{1}{2-\vartheta}} + \frac{15 B_0 \tau}{n} \end{aligned}$$

with probability not less than $1 - 3\mathrm{e}^{-\tau}$, where

$$a = \max\left\{ (C_{\mathrm{box}} K_d)^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \gamma^{-\frac{\varrho}{2p}}, B \right\} \leq p^{-\frac{d+1}{2p}} \gamma^{-\frac{\varrho}{2p}} \max\left\{ (C_{\mathrm{box}})^{\frac{1}{2p}}, B \right\} K_d^{\frac{1}{2p}}$$

with the constant K_d from Proposition 2.2.1 and K is given by (see Equation (1.13))

$$K = \tilde{K} \max \left\{ B, \left(|L|_{M,1}^p V^{\frac{1-p}{2}} \right)^{\frac{2}{2-p-\vartheta+\vartheta p}}, |L|_{M,1}^p B^{1-p}, 1 \right\}$$

with a constant $\tilde{K}$ independent of p and ϑ. The proof is now completed by combining all constants depending on p or ϑ into the constant $C_{\mathbf{P}}$. □

2.3 Least-Squares Regression

In this section we derive learning rates for SVMs using the least-squares loss function $L = L_{\mathrm{LS}}$ under suitable smoothness assumptions on $f^*_{L,\mathbf{P}}$. First of all, recall that for the least-squares loss, the Bayes decision function is given by the conditional mean function $f^*_{L,\mathbf{P}}(x) = \mathbf{E}(Y|X = x)$, see Example 1.1.1. We begin by introducing some tools for our notion of smoothness.

For a function $f : \mathbb{R}^d \to \mathbb{R}$ and $h \in \mathbb{R}^d$ the difference operator Δ_h is defined by $\Delta_h f(x) := f(x+h) - f(x)$. The s-fold application of Δ_h has the explicit expansion

$$\Delta_h^s f(x) = \sum_{j=0}^{s} \binom{s}{j} (-1)^{s-j} f(x + jh). \tag{2.3}$$

Given a measure μ on $X \subset \mathbb{R}^d$ we further define the s-th modulus of smoothness by

$$\omega_{s,L_2(\mu)}(f,t) := \sup_{\|h\| \le t} \|\Delta_h^s f\|_{L_2(\mu)}. \tag{2.4}$$

Finally, given an $\alpha > 0$ we set $s := \lfloor \alpha \rfloor + 1$ and define the semi-norm

$$|f|_{B^\alpha_{2,\infty}(\mu)} := \sup_{t>0} t^{-\alpha} \omega_{s,L_2(\mu)}(f,t). \tag{2.5}$$

Remark 2.3.1. The so-called Besov spaces $B^\alpha_{2,\infty}(\mathbb{R}^d)$, commonly used in approximation theory, can be defined by means of real interpolation of Sobolev spaces, see for example [3, Chapter 3] for an introduction to the real interpolation method. More precisely for $k_0 \in \mathbb{N}_0$ and $k_1 \in \mathbb{N}_0$ with $k_0 \neq k_1$ and $\theta \in (0,1)$ such that $\alpha = k_0(1-\theta) + k_1\theta$

$$B^\alpha_{2,\infty}(\mathbb{R}^d) := \left(W^{k_0,2}(\mathbb{R}^d), W^{k_1,2}(\mathbb{R}^d) \right)_{\theta,\infty},$$

cf. [61, Section 1.6.4]. Here, $W^{k,p}(\mathbb{R}^d)$ denotes the Sobolev space of functions on $\mathbb{R}^d$ that have p-integrable weak derivatives up to order k with $p \in [1,\infty]$ and $k \in \mathbb{N}_0$. Interpolation theory allows one to *fill the gaps* of the discrete range of smoothness of Sobolev spaces, just like the Hölder spaces fill the gaps between the spaces of k-times differentiable functions. In this sense, one can think of Besov spaces as Sobolev spaces with a continuous range of smoothness. For $\alpha > d/2$ we can define an equivalent norm on $B^\alpha_{2,\infty}(\mathbb{R}^d)$ by

$$\|\cdot\|_{L_2(\mathbb{R}^d)} + |\cdot|_{B^\alpha_{2,\infty}(\mathbb{R}^d)},$$

where $|\cdot|_{B^\alpha_{2,\infty}(\mathbb{R}^d)}$ is defined with respect to the Lebesgue measure on $\mathbb{R}^d$ in (2.4) and (2.5), cf. [61, Section 2.6.1], explaining our notation. Furthermore, we see that $|f|_{B^\alpha_{2,\infty}(\mathbf{P}_X)} < \infty$, whenever $f \in B^\alpha_{2,\infty}(\mathbb{R}^d)$ and $\mathbf{P}_X$ has a bounded Lebesgue-density.

The next proposition gives a sufficient condition for f to have a finite $B^\alpha_{2,\infty}(\mu)$-norm also for singular measures μ, which $\mathbf{P}_X$ necessarily is in our main focus where Assumption 2.1.1 is satisfied for $\varrho < d$. To formulate it, we first introduce some additional definitions and recall multi-index notation.

If f is k-times continuously differentiable we denote for a multi-index $\nu = (\nu_1,\ldots,\nu_d) \in \mathbb{N}_0^d$ with $|\nu| := \nu_1 + \ldots + \nu_d = k$ the higher-order partial derivative by

$$\partial^\nu f(x) = \frac{\partial^{|\nu|} f}{\partial x_1^{\nu_1} \ldots \partial x_d^{\nu_d}}(x).$$

Further, for a multi-index $\nu \in \mathbb{N}_0^d$ and an $x \in \mathbb{R}^d$ we write $x^\nu := x_1^{\nu_1} \cdot \ldots \cdot x_d^{\nu_d}$ as well as $\nu! := \nu_1! \cdot \ldots \cdot \nu_d!$

Definition 2.3.2. For $k \in \mathbb{N}_0$ and $\beta \in [0,1]$ let $C^{k,\beta}(\mathbb{R}^d)$ be the set of k-times continuously differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ with

$$|f|_{C^{k,\beta}(\mathbb{R}^d)} := \max_{|\nu|=k} \sup_{\substack{x,y\in\mathbb{R}^d \\ x\neq y}} \frac{|\partial^\nu f(x) - \partial^\nu f(y)|}{\|x-y\|^\beta} < \infty.$$

As in our setting the Bayes decision function $f^*_{L,\mathbf{P}}$ is in general essentially only defined on a set with empty interior, we briefly want to discuss the issue of imposing differentiability properties on such a function. To this end, assume that for a function $f : S \to \mathbb{R}$ there exists a collection of functions $f_\nu : S \to \mathbb{R}$, $\nu \in \mathbb{N}_0^d, |\nu| \leq k$,

where $f_0 = f$ such that

$$f_\nu(x) = \sum_{|\nu+\iota|\leq k} \frac{f_{\nu+\iota}(y)}{\iota!}(x-y)^\iota + R_\nu(x,y) \tag{2.6}$$

with $|f_\nu(x)| \leq C$ and the residuals R_ν satisfy

$$R_\nu(x,y) \leq C\|x-y\|^{k+\beta-|\nu|} \tag{2.7}$$

for some $0 < \beta \leq 1$ and all $x, y \in S$ and $|\nu| \leq k$. The obvious motivation for conditions (2.6) and (2.7) is that if $f \in C^{k,\beta}(\mathbb{R}^d)$ and S has non-empty interior, then (2.6) and (2.7) are satisfied for the partial derivatives $f_\nu = \partial^\nu f$. By Whitney's extension theorem [54, Chapter VI, Theorem 4], for a closed set S with *empty* interior any function $f : S \to \mathbb{R}$ satisfying (2.6) and (2.7) has an extension to a function $f_0 \in C^{k,\beta}(\mathbb{R}^d)$. As a consequence, we can consider $f^*_{L,\mathbf{P}}$ as a globally extended function without imposing any further restrictions. Moreover, this extension can always be chosen compactly supported. In the subsequent results, besides $f^*_{L,\mathbf{P}} \in C^{k,\beta}(\mathbb{R}^d)$, for some technical reasons we will also require that $f^*_{L,\mathbf{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$, which therefore also poses no further constraint.

Proposition 2.3.3. *Let μ be a finite measure on $\mathbb{R}^d$ and for some $k \in \mathbb{N}_0$ and $\beta \in [0,1)$ let $f \in C^{k,\beta}(\mathbb{R}^d)$ and set $\alpha := k + \beta$. Then we have $|f|_{B^\alpha_{2,\infty}(\mu)} \leq d^{k/2}\sqrt{\mu(\mathbb{R}^d)}|f|_{C^{k,\beta}(\mathbb{R}^d)}$.*

The proof of Proposition 2.3.3 requires an auxiliary lemma on a mean value theorem for higher order differences.

Lemma 2.3.4. *Let $f : (a,b) \to \mathbb{R}$ be s-times continuously differentiable. Furthermore, fix $x \in (a,b)$, $h > 0$, and $k \in \{1,\ldots,s\}$ with $(x, x+sh) \subset (a,b)$. Then there exists a $\xi \in (x, x+kh)$ such that $h^{-s}\Delta^s_h f(x) = h^{k-s}\Delta^{s-k}_h f^{(k)}(\xi)$, where Δ^s_h is the difference operator defined by (2.3).*

Proof. Because of $\Delta^s_h = \Delta_h\Delta^{s-1}_h$ we can apply the mean value theorem to the function $h^{-1}\Delta^{s-1}_h f$, which gives us $h^{-s}\Delta^s_h f(x) = h^{1-s}\Delta^{s-1}_h f'(\xi)$ for some $\xi \in (x, x+h)$. Note that in the last step we used $\frac{\mathrm{d}}{\mathrm{d}x}\Delta_h f(x) = \Delta_h f'(x)$. That is, we have proven the assertion for $k = 1$. Now we can iterate this argument by applying the mean value theorem to $h^{-1}\Delta^{s-2}_h f'$ and so on. □

Proof of Proposition 2.3.3. Fix an $x \in X$ and an $h \in \mathbb{R}^d\backslash\{0\}$ and define the univariate function $F(t) := f(x+th)$. For $s := \lfloor\alpha\rfloor + 1$ we then have $\Delta^s_h f(x) =$

$\Delta_1^s F(0) = \Delta_1 F^{(k)}(\xi)$ by Lemma 2.3.4 for some $\xi \in (0,k)$. The k-th derivative of F is given by

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k} F(t) = \sum_{|\nu|=k} \frac{k!}{\nu!} h^\nu \, \partial^\nu f(x+th).$$

This leads us to the estimate

$$\begin{aligned}
|\Delta_1 F^{(k)}(\xi)| &= |F^{(k)}(\xi+1) - F^{(k)}(\xi)| \\
&= \left| \sum_{|\nu|=k} \frac{k!}{\nu!} h^\nu \big(\partial^\nu f(x+(\xi+1)h) - \partial^\nu f(x+\xi h) \big) \right| \\
&\leq |f|_{C^{k,\beta}(\mathbb{R}^d)} \|h\|^\beta \left| \sum_{|\nu|=s-1} \frac{(s-1)!}{\nu!} \prod_{j=1}^d |h_j|^{\nu_j} \right| \\
&= |f|_{C^{k,\beta}(\mathbb{R}^d)} \|h\|^\beta \|h\|_{\ell_1^d}^k \\
&\leq d^{\frac{k}{2}} |f|_{C^{k,\beta}(\mathbb{R}^d)} \|h\|^\alpha
\end{aligned}$$

using the Definition 2.3.2 of the Hölder semi-norm in the first inequality, the multinomial theorem in the next step, and $\beta + k = \alpha$ in the last step, which immediately implies the result. □

For an application of Theorem 2.2.2 we need a suitable function $f_0 \in H_\gamma(X)$ bounding the approximation error. To this end, we first collect some facts on Gaussian RKHSs, which are a summary of [55, Theorem 4.21, Lemma 4.45, and Proposition 4.46]. By introducing the function $K_\gamma : \mathbb{R}^d \to \mathbb{R}$ defined by

$$K_\gamma(x) := \left(\frac{2}{\gamma\sqrt{\pi}} \right)^{\frac{d}{2}} \exp(-2\gamma^{-2}\|x\|^2), \qquad x \in \mathbb{R}^d,$$

the Gaussian RKHS $H_\gamma(X)$ can be characterized as the image of the convolution operator $L_2(\mathbb{R}^d) \to H_\gamma(X)$ defined by $g \mapsto K_\gamma * g$. The $H_\gamma(X)$-norm can be computed by $\|f\|_{H_\gamma(X)} = \inf\{\|g\|_{L_2(\mathbb{R}^d)} : f = K_\gamma * g\}$. Furthermore, for $0 < \gamma_1 < \gamma_2 < \infty$ the space $H_{\gamma_2}(X)$ is continuously embedded into $H_{\gamma_1}(X)$ with

$$\|\mathrm{id} : H_{\gamma_2}(X) \to H_{\gamma_1}(X)\| \leq \left(\frac{\gamma_2}{\gamma_1} \right)^{\frac{d}{2}}. \tag{2.8}$$

We will further make use of integration in spherical coordinates, see for example

[21, Theorem 2.49]. Namely, for $f \in L_1(\mathbb{R}^d)$ or $f \geq 0$ we have

$$\int_{\mathbb{R}^d} f(x) \, \mathrm{d}x = \int_0^\infty \int_{\mathbb{S}^{d-1}} f(r\omega) r^{d-1} \, \mathrm{d}\sigma(\omega) \, \mathrm{d}r, \tag{2.9}$$

where $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ and σ is the surface measure on $\mathbb{S}^{d-1}$. For radial functions f, that is $f(x) = g(\|x\|)$, Equation (2.9) simplifies to

$$\int_{\mathbb{R}^d} f(x) \, \mathrm{d}x = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty g(r) r^{d-1} \, \mathrm{d}r, \tag{2.10}$$

since $\sigma(\mathbb{S}^{d-1}) = 2\pi^{d/2}/\Gamma(d/2)$, see e.g. [21, Proposition 2.54]. Using (2.10) one can easily check, that

$$\int_{\mathbb{R}^d} (\gamma\sqrt{\pi})^{-\frac{d}{2}} K_\gamma(x) \, \mathrm{d}x = \int_{\mathbb{R}^d} \gamma^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|x\|^2\right) \, \mathrm{d}x = 1, \tag{2.11}$$

which we will rely on later. Finally, we define

$$G := \sum_{j=1}^{s} \binom{s}{j} (-1)^{1-j} (j\gamma\sqrt{\pi})^{-\frac{d}{2}} K_{j\gamma}, \tag{2.12}$$

where s will be chosen suitably later. Finally, note that by Example 1.1.1 we have

$$\mathcal{R}_{L,\mathbf{P}}(f) - \mathcal{R}^*_{L,\mathbf{P}} = \|f - f^*_{L,\mathbf{P}}\|^2_{L_2(\mathbf{P}_X)}$$

for all $f \in L_2(\mathbf{P}_X)$. The following lemma now bounds the approximation error of a Gaussian SVM using the least-squares loss.

Lemma 2.3.5. *For $f \in L_2(\mathbb{R}^d)$ with $|f|_{B^\alpha_{2,\infty}(\mathbf{P}_X)} < \infty$ we have*

$$\|G * f - f\|^2_{L_2(\mathbf{P}_X)} \leq |f|^2_{B^\alpha_{2,\infty}(\mathbf{P}_X)} 2^{-\alpha} \left(\frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^2 \gamma^{2\alpha}.$$

A similar bound as in the lemma above can be found in [16, Theorem 2.2]. Compared to [16], we provide a simpler proof leading to improved constants. The result in [16] is stated slightly more generally for the $L_p(\mathbf{P}_X)$-norm of $G * f - f$ and a (suitably modified) $B^\alpha_{q,\infty}(\mathbf{P}_X)$-semi-norm, however, our proof can be generalized easily to that case.

Proof. We set $s = \lfloor \alpha \rfloor + 1$ and compute

$$\begin{aligned} G * f(x) &= \int_{\mathbb{R}^d} \sum_{j=1}^{s} \binom{s}{j} (-1)^{1-j} (j\gamma)^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2(j\gamma)^{-2}\|y\|^2\right) f(x+y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} \sum_{j=1}^{s} \binom{s}{j} (-1)^{1-j} \gamma^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) f(x+jh)\,\mathrm{d}h. \end{aligned}$$

Using Equation (2.11) this implies

$$\begin{aligned} G * f(x) - f(x) &= G * f(x) - \int_{\mathbb{R}^d} \gamma^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) f(x)\,\mathrm{d}h \\ &= \int_{\mathbb{R}^d} \sum_{j=1}^{s} \binom{s}{j} (-1)^{1-j} \gamma^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) f(x+jh)\,\mathrm{d}h \\ &\quad - \int_{\mathbb{R}^d} \gamma^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) f(x)\,\mathrm{d}h \\ &= \int_{\mathbb{R}^d} (-1)^{1-s} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) \Delta_h^s f(x)\,\mathrm{d}h. \end{aligned}$$

With this identity we can bound our desired $L_2(\mathbf{P}_X)$-norm by

$$\begin{aligned} &\|G * f - f\|_{L_2(\mathbf{P}_X)}^2 \\ &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) \Delta_h^s f(x)\,\mathrm{d}h \right)^2 \mathrm{d}\mathbf{P}_X(x) \\ &\le \left(\int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{d} \left(\exp\left(-2\gamma^{-2}\|h\|^2\right) \Delta_h^s f(x)\right)^2 \mathrm{d}\mathbf{P}_X(x) \right)^{\frac{1}{2}} \mathrm{d}h \right)^2 \\ &= \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) \|\Delta_h^s f\|_{L_2(\mathbf{P}_X)}\,\mathrm{d}h \right)^2 \end{aligned}$$

using Minkowski's integral inequality. With our assumptions on f we can further bound this by

$$\begin{aligned} \|G * f - f\|_{L_2(\mathbf{P}_X)}^2 &\le \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} \exp\left(-\gamma^{-2}\|h\|^2\right) \omega_{s,L_2(\mathbf{P}_X)}(f, \|h\|)\,\mathrm{d}h \right)^2 \\ &\le |f|_{B_{2,\infty}^{\alpha}(\mathbf{P}_X)}^2 \left(\frac{2}{\gamma^2\pi}\right)^{d} \left(\int_{\mathbb{R}^d} \exp\left(-2\gamma^{-2}\|h\|^2\right) \|h\|^{\alpha}\,\mathrm{d}h \right)^2 \end{aligned}$$

which leaves us with computing the integral in the last step. This is done using spherical coordinates, which gives us

$$\begin{aligned}\int_{\mathbb{R}^d} \exp\left(-2\gamma^{-2}\|h\|^2\right)\|h\|^\alpha \,\mathrm{d}h &= \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty \exp\left(-2\gamma^{-2}r^2\right) r^{\alpha+d-1}\,\mathrm{d}r \\ &= \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty \frac{1}{2}\left(\frac{\gamma}{\sqrt{2}}\right)^{\alpha+d} \mathrm{e}^{-u}\, u^{\frac{\alpha+d}{2}-1}\,\mathrm{d}u \\ &= \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \left(\frac{\gamma}{\sqrt{2}}\right)^{\alpha+d} \Gamma\left(\frac{\alpha+d}{2}\right).\end{aligned}$$

Combining these considerations we get

$$\|G * f - f\|^2_{L_2(\mathbf{P}_X)} \leq |f|^2_{B^\alpha_{2,\infty}(\mathbf{P}_X)} 2^{-\alpha} \left(\frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^2 \gamma^{2\alpha}.$$

□

The following lemma bounds the regularization term.

Lemma 2.3.6. *For $f \in L_2(\mathbb{R}^d)$ we have $\|G * f\|_{H_\gamma(X)} \leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} 2^s \|f\|_{L_2(\mathbb{R}^d)}$.*

Proof. Because of the embedding property (2.8) we have

$$\begin{aligned}\|G * f\|_{H_\gamma(X)} &\leq \sum_{j=1}^s \binom{s}{j} (j\gamma\sqrt{\pi})^{-\frac{d}{2}} \|K_{j\gamma} * f\|_{H_\gamma(X)} \\ &\leq \sum_{j=1}^s \binom{s}{j} (\gamma\sqrt{\pi})^{-\frac{d}{2}} \|K_{j\gamma} * f\|_{H_{j\gamma}(X)} \\ &\leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} \|f\|_{L_2(\mathbb{R}^d)} \sum_{j=1}^s \binom{s}{j} \\ &\leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} 2^s \|f\|_{L_2(\mathbb{R}^d)}.\end{aligned}$$

□

With all these preparations completed we can now state our first main result of this section, which is the basis for the subsequent results.

Theorem 2.3.7. *Assume $\mathbf{P}$ satisfies Assumption 2.1.1 with parameters $C_{\mathrm{box}}, \varrho$ and that $Y \subset [-M, M]$. Further assume that $f^*_{L,\mathbf{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ as well as*

$|f^*_{L,\mathbf{P}}|_{B^\alpha_{2,\infty}(\mathbf{P}_X)} < \infty$. *Then for all* $\tau > 0$, $n > 1$, $\lambda \in (0,1)$ *and* $\gamma \in (0,1)$ *we have*

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq & c_1 \|f^*_{L,\mathbf{P}}\|^2_{L_2(\mathbb{R}^d)} \lambda\gamma^{-d} + c_2 |f^*_{L,\mathbf{P}}|^2_{B^\alpha_{2,\infty}(\mathbf{P}_X)} \gamma^{2\alpha} \\ & + c_3 K \lambda^{-1/\log n} \gamma^{-\varrho} n^{-1} \log^{d+1} n + c_4 \frac{\tau}{n}\end{aligned} \tag{2.13}$$

with probability $\mathbf{P}^n$ *not less than* $1 - 3\mathrm{e}^{-\tau}$, *where* $c_1 = 9\pi^{-d/2} 4^s$,

$$c_2 = 9 \left(\frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \right)^2 2^{-\alpha}, \quad c_3 = \max\left\{16M^2, 1\right\} \max\left\{C_{\mathrm{box}}, 4M^2\right\},$$
$$c_4 = 3456M^2 + 15 \max\{(2^s \|f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} + M)^2, 4M^2\}$$

with $s = \lfloor \alpha \rfloor + 1$ *and* K *is a constant independent of* $\mathbf{P}, n, \lambda$ *and* γ.

Proof. For $Y = [-M, M]$ the least-squares loss satisfies the supremum/variance bound for the constants $B = 4M^2, V = 16M^2$ and $\vartheta = 1$ by Example 1.3.2. Theorem 2.2.2 therefore gives us for $\lambda > 0, \gamma \in (0,1)$, and $p \in (0, 1/2]$

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq & 9(\lambda \|f_0\|^2_{H_\gamma(X)} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}}) \\ & + C_{\mathbf{P}} K p^{-d-1} \gamma^{-\varrho} \lambda^{-p} n^{-1} + \frac{(3456M^2 + 15B_0)\tau}{n}\end{aligned}$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$. To bound the approximation error we set $f_0 := G * f^*_{L,\mathbf{P}}$, where G is defined by (2.12) for $s = \lfloor \alpha \rfloor + 1$. First we determine B_0, i.e. a bound on $\sup_{(x,y)\in X\times Y} |y - f_0(x)|^2$. By Young's convolution inequality we have

$$\|f_0\|_{L_\infty(\mathbb{R}^d)} = \|G * f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} \leq \|f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} \cdot \|G\|_{L_1(\mathbb{R}^d)}$$

and by using Equation (2.11) we get

$$\|G\|_{L_1(\mathbb{R}^d)} \leq \sum_{j=1}^s \binom{s}{j} (j\gamma\sqrt{\pi})^{-\frac{d}{2}} \|K_{j\gamma}\|_{L_1(\mathbb{R}^d)} = \sum_{j=1}^s \binom{s}{j} \leq 2^s.$$

Consequently, we get

$$\sup_{(x,y)\in X\times Y} |f_0(x) - y|^2 \leq \left(2^s \|f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} + M\right)^2,$$

i.e. we can set

$$B_0 := \max\left\{\left(2^s \|f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} + M\right)^2, 4M^2\right\}.$$

Using Lemma 2.3.5 and Lemma 2.3.6 with $s = \lfloor \alpha \rfloor + 1$ we can bound the regularization term and the approximation error as stated in the theorem. To determine a bound on $C_{\mathbf{P}}$ first note that $|L_{\mathrm{LS}}|_{M,1} \leq 4M$ by Example 1.3.6. Some calculations then show $C_{\mathbf{P}} \leq \max\{16M^2, 1\} \max\{C_{\mathrm{box}}, 4M^2\}$. Finally, the desired inequality follows by setting

$$p = \frac{\log 2}{2 \log n} \leq 1/2.$$

□

Using the theorem above we can easily derive learning rates by choosing specific values for the regularization parameter λ and the bandwidth γ.

Corollary 2.3.8. *Let the assumptions of Theorem 2.3.7 be satisfied with the bounds* $\|f^*_{L,\mathbf{P}}\|_{L_2(\mathbb{R}^d)} \leq C_1$, $\|f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} \leq C_2$ *and* $|f^*_{L,\mathbf{P}}|_{B^\alpha_{2,\infty}(\mathbf{P}_X)} \leq C_3$. *Choosing* $\gamma_n = n^{-1/(2\alpha+\varrho)}$ *and* $\lambda_n = n^{-b}$ *for some* $b \geq (2\alpha+d)/(2\alpha+\varrho)$ *there then exists a constant* $C > 0$ *only depending on* $C_{\mathrm{box}}, C_1, C_2, C_3$, *and* M *such that for all* $n > 1$ *and* $\tau \geq 1$ *we have*

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda_n,\gamma_n}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{2\alpha}{2\alpha+\varrho}} \log^{d+1} n$$

with probability $\mathbf{P}^n$ *not less than* $1 - \mathrm{e}^{-\tau}$.

Proof. Theorem 2.3.7 gives us

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\left(\lambda\gamma^{-d} + \gamma^{2\alpha} + \lambda^{-1/\log n}\gamma^{-\varrho} n^{-1} \log^{d+1} n + \frac{\tau}{n}\right)$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$ for all $n \in \mathbb{N}$ and a constant C only depending on $C_{\mathrm{box}}, C_{1,2,3}$ and M. With the choices of λ_n and γ_n as stated in the corollary we get

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} &\leq C\left(n^{-b} n^{\frac{d}{2\alpha+\varrho}} + n^{-\frac{2\alpha}{2\alpha+\varrho}} + \mathrm{e}^b n^{-\frac{2\alpha}{2\alpha+\varrho}} \log^{d+1} n + \frac{\tau}{n}\right) \\ &\leq C\left(2n^{-\frac{2\alpha}{2\alpha+\varrho}} + \mathrm{e}^b n^{-\frac{2\alpha}{2\alpha+\varrho}} \log^{d+1} n + \frac{\tau}{n}\right)\end{aligned}$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$ for all $n \in \mathbb{N}$. A substitution of τ then easily proves the assertion. □

Remark 2.3.9. It is also possible to formulate Theorem 2.3.7 and Corollary 2.3.8 under alternative regularity assumptions. To briefly elaborate this, recall that in [66] the authors consider the case where X is a compact, connected and smooth submanifold of $\mathbb{R}^d$ without boundary and consider a convolution-type operator $S_\gamma : L_2(\mu) \to H_\gamma(X)$, where μ is the measure on X defined by the Riemannian volume form and derive the bounds

$$\|S_\gamma f\|^2_{H_\gamma(X)} \leq C_1 \|f\|^2_{L_2(\mu)} \gamma^{-d} \tag{2.14}$$

$$\|S_\gamma f - f\|^2_{L_2(\mu)} \leq C_2 \|f\|^2_{W^{2,2}(X)} \gamma^4 \tag{2.15}$$

$$\|S_\gamma f\|_{\ell_\infty(X)} \leq C_3 \|f\|_{\ell_\infty(X)} \tag{2.16}$$

[66, Lemma 4, Theorem 2, and Lemma 2], where $W^{2,2}(X)$ denotes the Sobolev space on X. Using these results one can easily derive a modification of Theorem 2.3.7 and Corollary 2.3.8 under the assumption that $f^*_{L,\mathbf{P}} \in W^{2,2}(X)$ is bounded and $\mathbf{P}_X$ has a bounded density with respect to μ and prove learning rates of the form $n^{-\frac{4}{4+\varrho}} \log^{d+1} n$. This is done by simply using (2.15) instead of Lemma 2.3.5, (2.14) instead of Lemma 2.3.6 and the supremum bound (2.16) instead of an analogous bound we derive in the proof of Theorem 2.3.7. Unfortunately, the authors also point out, that the order of approximation cannot be improved if we assume $f \in W^{m,2}(X)$ for some $m > 2$ using the operator S_γ.

The learning rates in Corollary 2.3.8 can only be achieved, if we know the intrinsic dimension ϱ of the data, as well as the regularity α of the Bayes decision function. However, this is highly unrealistic in practice. The following theorem therefore shows, that a TV-SVM with appropriately chosen candidate sets Λ_n and Γ_n achieves the same rate without knowledge on ϱ and α.

Theorem 2.3.10. *Let A_n be a minimal $1/\log n$-net of $(0,1]$ with $1 \in A_n$ and let B_n be a minimal $1/\log n$-net of $[1,d]$ with $d \in B_n$. Set $\Gamma_n := \{n^{-a} : a \in A_n\}$ and $\Lambda_n := \{n^{-b} : b \in B_n\}$. Let the assumptions of Theorem 2.3.7 be satisfied with $\|f^*_{L,\mathbf{P}}\|_{L_2(\mathbb{R}^d)} \leq C_1, \|f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} \leq C_2$ and $|f^*_{L,\mathbf{P}}|_{B^\alpha_{2,\infty}(\mathbf{P}_X)} \leq C_3$ and assume $\varrho \geq 1$. Then there exists a constant $C > 0$ only depending on $C_{\text{box}}, C_1, C_2, C_3$, and M such that for all $n > 1$ and $\tau \geq 1$ the TV-SVM using Λ_n and Γ_n satisfies*

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{2\alpha}{2\alpha+\varrho}} \log^{d+1} n$$

with probability $\mathbf{P}^n$ not less than $1 - \mathrm{e}^{-\tau}$.

Proof. We define $\gamma_n := n^{-1/(2\alpha+\varrho)}$ and $\lambda_n := n^{-(2\alpha+d)/(2\alpha+\varrho)}$. By [55, Theorem 7.2], which states an oracle inequality for empirical risk minimization over finite hypothesis sets, we have

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}^*_{L,\mathbf{P}} &\le 6 \min_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \left(\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}}\right) \\ &\quad + \frac{512M^2(\tau + \log(1 + |\Lambda_n \times \Gamma_n|))}{n-m} \\ &\le 6 \left(\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_0,\gamma_0}) - \mathcal{R}^*_{L,\mathbf{P}}\right) \\ &\quad + \frac{2048M^2(\tau + \log(1 + |\Lambda_n \times \Gamma_n|))}{n}\end{aligned}$$

with probability $\mathbf{P}^{n-m}$ not less than $1 - \mathrm{e}^{-\tau}$, where $m = \lfloor n/2 \rfloor + 1$ and in the last step we picked $\gamma_0 := n^{-a} \in \Gamma_n$ and $\lambda_0 := n^{-b} \in \Lambda_n$ for values a and b, which we will specify in a moment. An application of Theorem 2.3.7 combined with $n \le 2m$ gives us

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_0,\gamma_0}) - \mathcal{R}^*_{L,\mathbf{P}} &\le C \left(\lambda_0\gamma_0^{-d} + \gamma_0^{2\alpha} + b^{d+1}\gamma_0^{-\varrho} m^{-1} \log^{d+1} n + \frac{\tau}{m}\right) \\ &\le C \left(\lambda_0\gamma_0^{-d} + \gamma_0^{2\alpha} + 2b^{d+1}\gamma_0^{-\varrho} n^{-1} \log^{d+1} n + \frac{2\tau}{n}\right)\end{aligned}$$

with probability $\mathbf{P}^m$ not less than $1 - 3\mathrm{e}^{-\tau}$. Now let $\lambda_0 = n^{-d}$ and let $a \in A_n$ satisfy $1/(2\alpha+\varrho) \le a \le 1/(2\alpha+\varrho) + 1/\log n$, which implies

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_0,\gamma_0}) - \mathcal{R}^*_{L,\mathbf{P}} &\le C \left(\mathrm{e}^d\lambda_n\gamma_n^{-d} + \gamma_n^{2\alpha} + 2d^{d+1}\mathrm{e}^{\varrho}\gamma_n^{-\varrho} n^{-1} \log^{d+1} n + \frac{2\tau}{n}\right) \\ &= C \left(\mathrm{e}^d n^{-\frac{2\alpha}{2\alpha+\varrho}} + n^{-\frac{2\alpha}{2\alpha+\varrho}} + 2d^{d+1}\mathrm{e}^{\varrho} n^{-\frac{2\alpha}{2\alpha+\varrho}} \log^{d+1} n + \frac{2\tau}{n}\right)\end{aligned}$$

with probability $\mathbf{P}^m$ not less than $1 - 3\mathrm{e}^{-\tau}$. Combining these inequalities and using that $|\Lambda_n \times \Gamma_n| \in \mathcal{O}(\log^2 n)$ we get

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}^*_{L,\mathbf{P}} \le c_1 \left(n^{-\frac{2\alpha}{2\alpha+\varrho}} \log^{d+1} n + \frac{\tau}{n}\right) + c_2 \left(\frac{\tau}{n} + \frac{\log n}{n}\right)$$

with probability $\mathbf{P}^n$ not less than $(1 - \mathrm{e}^{-\tau})(1 - 3\mathrm{e}^{-\tau}) \ge 1 - 4\mathrm{e}^{-\tau}$ for all $n > 2$. □

Remark 2.3.11. The proof of Theorem 2.3.10 shows that the statement also holds if we pick as candidate set for λ the singleton $\Lambda_n = \{n^{-d}\}$. We decided to formulate the theorem as it is, because this choice is closer to the practical usage of the training

validation approach. To see why a singleton Λ_n is sufficient recall that to achieve optimal rates the regularization parameter λ_n only needs to satisfy $\lambda_n = n^{-b}$ for some $b \geq (2\alpha + d)/(2\alpha + \varrho)$. If we assume $\varrho \geq 1$ this bound on b is satisfied for $b = d$. This also shows that with no lower bound on ϱ the regularization parameter λ_n possibly needs to decay arbitrarily fast as $(2\alpha + d)/(2\alpha + \varrho)$ is unbounded for $\alpha, \varrho > 0$. On the one hand this shows why we need the additional constraint $\varrho \geq 1$ in Theorem 2.3.10, on the other hand we want to mention that the case $\varrho < 1$ is of little practical interest anyway. Nevertheless, the statement of Theorem 2.3.10 still holds with the rate $n^{-2\alpha/(2\alpha+1)} \log^{d+1} n$ in the case $\varrho < 1$.

2.4 Binary Classification

In Section 1.3 we already introduced Assumption 1.3.3, a regularity assumption for binary classification for bounding the statistical error by providing a variance bound. In order to prove the results of this section, we also need to introduce a second regularity assumption for bounding the approximation error for Gaussian SVMs using the hinge loss. To this end, first note that in binary classification it is intuitively hard to make a correct prediction for the label of $x \in X$ whenever $\eta(x) \approx 1/2$, where $\eta : X \to [0, 1]$ is the conditional class probability function, see Example 1.1.2. Consequently, Assumption 1.3.3 captures this intuition by restricting the mass of points $x \in X$ where $\eta(x)$ is close to $1/2$. Additionally, classifying $x \in X$ may also be hard if x is close to the decision boundary, which is incorporated by our second regularity assumption. Therefore, we need the following definition:

Definition 2.4.1. Let $X_{-1} := \{x \in X : \eta(x) < 1/2\}$ and $X_1 := \{x \in X : \eta(x) > 1/2\}$ and define

$$\Delta(x) := \begin{cases} \operatorname{dist}(x, X_1) & \text{if } x \in X_{-1}, \\ \operatorname{dist}(x, X_{-1}) & \text{if } x \in X_1, \\ 0 & \text{else,} \end{cases}$$

where $\operatorname{dist}(x, A) := \inf_{y \in A} \|x - y\|$.

Our second central regularity assumption, which restricts the mass *and* location of points $x \in X$ with $\eta(x) \approx 1/2$, then reads as follows:

Assumption 2.4.2. There exist constants $C_{**} > 0$ and $\beta > 0$ such that

$$\int_{\{x \in X : \Delta(x) < t\}} |2\eta(x) - 1| \, \mathrm{d}\mathbf{P}_X(x) \leq C_{**} t^{\beta}$$

for all $t \geq 0$.

The condition in Assumption 2.4.2 was introduced (in a slightly different version) in [57] to prove fast classification rates for Gaussian SVMs. The authors used the *term geometric noise condition* to describe this assumption. Assumption 2.4.2 was further adopted in [6] for the analysis of a histogram based classifier. Also, in [62] the authors point out that their *probabilistic Lipschitzness* condition in Definition 1 is closely related to Assumption 2.4.2. Intuitively, Assumption 2.4.2 is satisfied for a large exponent β if $\mathbf{P}_X$ has only a low concentration in the vicinity of the decision boundary, or if $\mathbf{P}$ is particularly noisy in this region. For example, in the extreme case, in which X_{-1} and X_1 have positive distance, we may choose arbitrarily large β. To give a better intuition on Assumptions 1.3.3 and 2.4.2, as well as on their interplay, we provide some explicit examples below.

Example 2.4.3. In the following let $X = [-1, 1]^2$.

(i) Assume $\mathbf{P}_X$ restricted to $A := [-1/2, 1/2] \times [-1, 1]$ is proportional to a measure with Lebesgue-density $|x_1|^{\sigma} \mathrm{d}\lambda(x_1, x_2)$ for some $\sigma > 0$ and on $X \backslash A$ is proportional to the uniform distribution. Further assume η is the sawtooth function

$$\eta(x_1, x_2) = \begin{cases} 2(1 - x_1) & \text{for } x_1 \in (1/2, 1], \\ 2x_1 & \text{for } x_1 \in [-1/2, 1/2], \\ 2(-1 - x_1) & \text{for } x_1 \in [-1, -1/2). \end{cases}$$

Then by the behavior of $\mathbf{P}_X$ and η near $x_1 = \pm 1$ the optimal exponent in Assumption 1.3.3 is given by $q = 1$. By the low concentration of $\mathbf{P}_X$ near the decision boundary Assumption 2.4.2 is fulfilled for the exponent $\beta = \sigma + 2$.

(ii) Assume that $\mathbf{P}_X$ is uniform on $\{x \in [-1, 1]^2 : |x_2| \leq |x_1|^{\zeta}\}$ for some $\zeta > 0$ and that $2\eta(x) - 1 = x_1$, i.e. the classes X_1, X_{-1} only meet at the point $(0, 0)$ and not along a one-dimensional curve. Then Assumption 1.3.3 is fulfilled for $q = \zeta + 1$ and Assumption 2.4.2 for $\beta = \zeta + 2$.

The first example shows that Assumption 1.3.3 describes the global amount of mass close to the critical level $\eta = 1/2$, while Assumption 2.4.2 can additionally benefit from low concentration of $\mathbf{P}_X$ in the vicinity of the decision boundary. The second example shows that Assumption 2.4.2 can also benefit from geometrical assumptions on the decision boundary, respectively the decision classes, and especially that the exponents in Assumption 1.3.3 and 2.4.2 can be simultaneously large. Also note that the exponent from Assumption 1.3.3 can be deteriorated by the behavior of $\mathbf{P}_X$ or η far away from the decision boundary, as seen in the first example, while Assumption 2.4.2 is robust to such perturbations. Further note that also a combination of the effects demonstrated by the examples above is possible.

Another common regularity condition in binary classification is to impose smoothness assumption on η. This type of regularity assumption is particularly popular for the analysis of plug-in classifiers, which implicitly treat the binary classification problem as a regression problem by first computing an estimate $\widehat{\eta}$ of the conditional class probability function η and then predicting labels using the function $\operatorname{sgn}(2\widehat{\eta}-1)$. The next proposition, which is the content of [55, Lemma 8.23], helps us to compare our results to results in the literature that use a smoothness assumption on η in some cases.

Proposition 2.4.4. *Assume there exist constants $c, \alpha > 0$ such that $|2\eta(x) - 1| \leq c\Delta^{\alpha}(x)$ for $\mathbf{P}_X$-almost all $x \in X$ and that Assumption 1.3.3 is satisfied for constants C_* and q. Then Assumption 2.4.2 is satisfied for $\beta = \alpha(q+1)$ and some constant C_{**} only depending on c and C_*.*

For $\alpha \leq 1$ the assumption in the proposition above can be seen as a substantially weaker form of α-Hölder regularity for η, since for some $x_0 \in \{x : \eta(x) = 1/2\}$ attaining minimum distance to $x \in X$ this condition can be rewritten as $|\eta(x) - \eta(x_0)| \leq c|x - x_0|^{\alpha}/2$. That is, the Hölder condition does not need to be satisfied for *arbitrary* $x, x_0 \in X$ but only where one of the points considered is in $\{x : \eta(x) = 1/2\}$.

Theorem 2.4.5. *Assume $\mathbf{P}$ satisfies Assumption 2.1.1 with parameters $C_{\mathrm{box}}, \varrho$ as well as Assumptions 1.3.3 and 2.4.2 with parameters C_*, q and C_{**}, β respectively. Then for the SVM using the hinge loss $L = L_{\mathrm{hinge}}$ we have for all $\tau > 0, n > 1, \lambda \in$*

$(0,1)$ and $n^{-1/\varrho} \leq \gamma \leq 1$

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq & c_1\lambda\gamma^{-d} + c_2 C_{**}\gamma^{\beta} + c_3 K \lambda^{-1/\log n}\left(\frac{\gamma^{-\varrho}}{n}\right)^{\frac{q+1}{q+2}} \log^{d+1} n \\ & + 3C_*^{\frac{q}{q+2}}\left(\frac{432\tau}{n}\right)^{\frac{q+1}{q+2}} + 30\frac{\tau}{n}\end{aligned} \tag{2.17}$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$, where $c_1 = 3^{d+2}/\Gamma(d/2+1)$,

$$c_2 = 9\frac{2^{1-\beta/2}\Gamma\left(\frac{\beta+d}{2}\right)}{\Gamma(d/2)}, \quad c_3 = \max\left\{C_*^{q/(q+1)}, 4\right\}\max\left\{C_{\mathrm{box}}, 2\right\}$$

and K is a constant independent of $\mathbf{P}, n, \lambda$ and γ.

Proof. The distribution $\mathbf{P}$ satisfies the supremum bound for $B = 2$ and the variance bound is satisfied for $V = 6C_*^{q/(q+1)}$ and $\vartheta = q/(q+1)$, see Example 1.3.4, where C_* is the constant from Assumption 1.3.3. Given this value for ϑ, the exponent in Theorem 2.2.2 then reads

$$\begin{aligned}\frac{1}{2-p-\vartheta+\vartheta p} &= \frac{1}{2-p-\frac{q}{q+1}(1-p)} = \frac{q+1}{(2-p)(q+1)-q(1-p)} \\ &= \frac{q+1}{2q+2-pq-p-q+pq} = \frac{q+1}{q+2-p}.\end{aligned}$$

and Theorem 2.2.2 then gives us for $\lambda > 0, \gamma \in (0,1)$, and $p \in (0, 1/2]$ that

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq & \, 9\left(\lambda\|f_0\|^2_{H_\gamma(X)} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}}\right) \\ & + C_{\mathbf{P}} K\left(\frac{p^{-d-1}\lambda^{-p}\gamma^{-\varrho}}{n}\right)^{\frac{q+1}{q+2-p}} + 3\left(\frac{432 C_*^{\frac{q}{q+1}}\tau}{n}\right)^{\frac{q+1}{q+2}} + \frac{15B_0\tau}{n}\end{aligned}$$

with probability not less than $1-3\mathrm{e}^{-\tau}$. In [55, Theorem 8.18] a function $f_0 \in H_\gamma(X)$ with $\|f_0\|_\infty \leq 1$ and

$$\lambda\|f_0\|^2_{H_\gamma(X)} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}} \leq c_1\lambda\gamma^{-d} + c_2 C_{**}\gamma^{\beta},$$

is constructed, where $c_1 = 3^d/\Gamma(d/2+1)$, $c_2 = 2^{1-\beta/2}\Gamma((\beta+d)/2)/\Gamma(d/2)$ and C_{**} is the constant from Assumption 2.4.2. Further, we have $|L_{\mathrm{hinge}}|_{1,1} = 1$ by Example 1.3.7 and since $\|f\|_\infty \leq 1$ we can choose $B_0 = 2$. Simple calculations yield $C_{\mathbf{P}} \leq \max\{4, C_*^{q/(q+1)}\}\max\{2, C_{\mathrm{box}}\}$. Finally, choosing $p = \log 2/(2\log n) \leq 1/2$

and some simple estimates prove the result. □

Corollary 2.4.6. *Let the assumptions of Theorem 2.3.7 be satisfied and set* $\gamma_n = n^{-a}$ *and* $\lambda_n = n^{-b}$ *with*

$$a = \frac{q+1}{\beta(q+2)+\varrho(q+1)} \quad \text{and} \quad b \geq \frac{(d+\beta)(q+1)}{\beta(q+2)+\varrho(q+1)}.$$

Then there exists a constant $C > 0$ *only depending on* C_{box}, C_* *and* C_{**} *such that for all* $n > 1$ *and* $\tau \geq 1$ *we have*

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda_n,\gamma_n}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{\beta(q+1)}{\beta(q+2)+\varrho(q+1)}} \log^{d+1} n$$

with probability $\mathbf{P}^n$ *not less than* $1 - \mathrm{e}^{-\tau}$.

If η satisfies the condition in Proposition 2.4.4 for some $\alpha > 0$ the exponent in the rate in the corollary above is given by $\alpha(q+1)/(\alpha(q+2)+\varrho)$. For $\alpha \leq 1$ we see, using the short remark after Proposition 2.4.4, that by [2, Theorem 4.1] the rate we get from the corollary above is optimal up to a logarithmic factor and that we achieve this optimal rate for a substantially larger class of distributions than the class, for which the exact optimal rate was established in [2]. As the proof of Corollary 2.4.6 merely consists of plugging in the specified values, we will skip it at this point.

Theorem 2.4.7. *Let* A_n *be a minimal* $1/\log n$*-net of* $(0,1]$ *with* $1 \in A_n$ *and let* B_n *be a minimal* $1/\log n$*-net of* $(0,d]$ *with* $d \in B_n$. *Set* $\Gamma_n := \{n^{-a} : a \in A_n\}$ *and* $\Lambda_n := \{n^{-b} : b \in B_n\}$. *Let the assumptions of Theorem 2.4.5 be satisfied and assume* $\varrho \geq 1$. *Then there exists a constant* $C > 0$ *only depending on* C_{box}, C_* *and* C_{**} *such that the TV-SVM using* Λ_n *and* Γ_n *satisfies for all* $n > 1$ *and* $\tau \geq 1$,

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{\beta(q+1)}{\beta(q+2)+\varrho(q+1)}} \log^{d+1} n$$

with probability $\mathbf{P}^n$ *not less than* $1 - \mathrm{e}^{-\tau}$.

Proof. Recall that our optimal choice for γ_n and λ_n was given by $\gamma_n = n^{-a}$ and $\lambda_n = n^{-b}$ with

$$a = \frac{q+1}{\beta(q+2)+\varrho(q+1)} \quad \text{and} \quad b \geq \frac{(d+\beta)(q+1)}{\beta(q+2)+\varrho(q+1)}.$$

Now note that for $\varrho \geq 1$ we have $a \leq 1$ and the choice $b = d$ is admissible. That is, by construction Γ_n and Λ_n cover a possible choice of γ_n and λ_n, which achieve optimal rates. The statement can now be proven exactly as in the proof of Theorem 2.3.10 by using Theorem 2.4.5. □

3 Learning Rates for Local SVMs

A major drawback of SVMs are their computational costs, which are quadratic in space and at least quadratic in time, which makes this method infeasible for large scale datasets. A popular approach to circumvent this issue is the so-called localization technique, which consists of constructing a partition of the input space into m disjoint cells and computing a local SVM decision function on each cell using only the samples contained in that cell. Prediction for a new point $x \in X$ is then performed by only considering the local decision function of the cell in which x is contained. So far, existing results on partitioning methods, such as [7, 41, 43], consider an a-priori fixed partition of the input space satisfying some technical assumptions, whereas we consider a fully data dependent partition based on the farthest first traversal algorithm. In this chapter we show that localized Gaussian SVMs using this partitioning scheme achieve the same learning rates as in Chapter 2 under a slightly stronger notion of fractal dimension and a mild assumption on the distribution $\mathbf{P}_X$. Additionally, we will again show that an analogous training validation scheme for hyperparameter selection achieves the same learning rates, as long as the number of the cells in the partition does not grow too fast with the sample size. A similar approach for speeding up kernel methods is random chunking, see for example [67], where the dataset is *randomly* split into m subsets that are used for computing m separate decision functions and are then averaged to define the final decision function, which gives this method a higher computational cost for inference compared to ours. Other popular approaches for speeding up kernel methods are Nyström subsampling [63], where a low rank approximation of the kernel matrix is used or random Fourier features [46] where for translation invariant kernels a low dimensional, randomized approximation of the feature map is computed by utilizing Bochner's theorem [47, Theorem IX.9]. The results of this chapter will be published in [25].

3.1 Intrinsic Dimension Assumption

For the formulation of our assumptions of this chapter, we need to introduce yet another covering quantity.

Definition 3.1.1. Given $A \subset \mathbb{R}^d$, for $m \in \mathbb{N}$ the m-th inner entropy number is defined as

$$\varepsilon_m(A) := \inf\{\varepsilon > 0 : \text{there exists an } \varepsilon\text{-net } N \subset A \text{ of } A \text{ with } |N| = m\}.$$

Note that for some technical reasons that will become obvious later, in the definition of $\varepsilon_m(A)$ the ε-net is required to be contained *inside* of the set A. Our notion of intrinsic dimension in this chapter is based on the Assouad dimension of $S := \operatorname{supp} \mathbf{P}_X$, which is defined as the infimum over all $\varrho > 0$ such that there exists a finite constant C_S such that

$$\sup_{x \in S} \mathcal{N}_{\ell_2^d}\big(B_r(x) \cap S, \varepsilon\big) \leq C_S \left(\frac{\varepsilon}{r}\right)^{-\varrho},$$

see [22, Section 2.1]. The definition of Assouad dimension generalizes straightforward to general metric spaces and is used to characterize metric spaces that can be bi-Lipschitz embedded in a Euclidean space, see [36]. Again, the exponent ϱ in Assumption 3.1.2 is consistent with classical notions of dimensions, e.g. the dimension of Euclidean spaces and manifolds, which is a consequence of the basic properties of the Assouad dimension summarized in [22, Section 2.4]. Note that by choosing $r > 0$ sufficiently large, Assumption 3.1.2 especially implies that $\mathcal{N}(S, \varepsilon) \in \mathcal{O}(\varepsilon^{-\varrho})$ as $\varepsilon \to 0$ and by some basic properties of covering and entropy numbers the latter is equivalent to $\varepsilon_m(S) \in \mathcal{O}(m^{-1/\varrho})$ as $m \to \infty$. As we often have to switch between bounds on entropy numbers and covering numbers, we also want to assume for convenience that the the previously stated bound on the asymptotic of $\varepsilon_m(S)$ is satisfied for the same constant C_S and that this bound on $\varepsilon_m(S)$ is sharp.

Assumption 3.1.2. The set S is bounded and there exist constants $C_S \geq 1$ and $\varrho > 0$ such that

$$\sup_{x \in S} \mathcal{N}_{\ell_2^d}\big(B_r(x) \cap S, \varepsilon\big) \leq C_S \left(\frac{\varepsilon}{r}\right)^{-\varrho} \tag{3.1}$$

for all $0 < \varepsilon \leq r$ as well as

$$C_S^{-1} m^{-\frac{1}{\varrho}} \leq \varepsilon_m(S) \leq C_S m^{-\frac{1}{\varrho}} \tag{3.2}$$

for all $m \in \mathbb{N}$.

We further have to make an assumption on the small ball probabilities of $\mu := \mathbf{P}_X$.

Assumption 3.1.3. There exist constants $C_\mu \geq 1$ and $\delta > 0$ such that

$$\inf_{x \in S} \mu(B_r(x)) \geq C_\mu^{-1} r^\delta$$

for all $0 < r \leq \operatorname{diam} S$.

To give a quick example on typical values of the constant δ in Assumption 3.1.3, assume that $X = [-1, 1]^d$ and that μ has a density with respect to the uniform distribution on X bounded away from 0. Then Assumption 3.1.3 is satisfied for $\delta = d$. Note that for this example not only the density assumption on μ is crucial, but also the geometry of the support of μ. For example, if μ is the uniform distribution on a domain X with cusps, then in general Assumption 3.1.3 is not fulfilled, at least not for $\delta = d$. Assumptions similar to Assumptions 3.1.3 are common in level set estimation, see for example [10] for a survey or [1, Remark 1] for an explicit construction of probability measures on sets $S \subset \mathbb{R}^d$, that are the image of a compact set $K \subset \mathbb{R}^{d'}, d' \leq d$ under a Lipschitz map satisfying Assumption 3.1.3 for $\delta = d'$. More generally, connections between properties of metric spaces described by their covering numbers (such as Assumption 3.1.2) and properties of measures on that space (especially how they act on balls) is a well-studied field in fractal geometry, see for example [28, Chapter 1]. Particularly interesting for us is that, as a consequence of [28, Theorem 13.5], if Assumption 3.1.2 is satisfied for some ϱ, then for every $\delta > \varrho$ there exists a measure μ on S satisfying Assumption 3.1.3 for this respective δ. That is, also in the general case where Assumption 3.1.2 is fulfilled for some non-integer ϱ, there exist distributions satisfying Assumption 3.1.3.

3.2 Localized Kernels and Construction of Partition

The approach of dividing the input space into disjoint cells and solving the initial learning problem independently on each cell with the data points contained in the respective cell is especially convenient for kernel methods from a mathematical perspective, since this procedure can be described by simply using a modified kernel,

which we will explain in the following. Although this construction easily generalizes to arbitrary kernels, we will only state it for Gaussian kernels so we can immediately incorporate the additional bandwidth parameter into our notation.

Given a partition $\mathcal{A} = (A_j)_{j=1,\dots,m}$ of the input space X and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m) \in (0, \infty)^m$ let $H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$ be the space of functions $f : X \to \mathbb{R}$ such that $f|_{A_j} \in H_{\gamma_j}(A_j)$ for all $j = 1, \dots, m$ equipped with the norm

$$\|f\|^2_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} := \sum_{j=1}^{m} \lambda_j \|f|_{A_j}\|^2_{H_{\gamma_j}(A_j)}.$$

Then $H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$ is a Hilbert space where the inner product is given by

$$\langle f, g\rangle_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} = \sum_{j=1}^{m} \lambda_j \langle f|_{A_j}, g|_{A_j}\rangle_{H_{\gamma_j}(A_j)}, \quad f, g \in H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A}).$$

Moreover, $H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$ is an RKHS. To see this, we define $k : X \times X \to \mathbb{R}$ by

$$k(x, y) := \sum_{j=1}^{m} \lambda_j^{-1} \mathbf{1}_{A_j}(x) k_{\gamma_j}(x, y) \mathbf{1}_{A_j}(y).$$

and verify the reproducing property

$$\langle f, k(x, \cdot)\rangle_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} = \sum_{j=1}^{m} \mathbf{1}_{A_j}(x) \langle f|_{A_j}, k_{\gamma_j}(x, \cdot)|_{A_j}\rangle_{H_{\gamma_j}(A_j)} = f(x).$$

For the RKHS $H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$ and a convex loss function $L : Y \times \mathbb{R} \to [0, \infty)$ we now consider the regularized empirical risk minimizer

$$f_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}} := \operatorname*{arg\,min}_{f \in H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} \|f\|^2_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} + \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)), \tag{3.3}$$

where $D = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ is a dataset. We will call the resulting learning method a localized support vector machine (LSVM) Note that compared to the global objective (1.6), in (3.3) the regularization parameter λ is now a component of the RKHS norm and can be chosen individually on each cell. If we define $I_j = \{i : x_i \in A_j\}$ for $j = 1, \dots, m$, we can rewrite the learning objective as

$$f_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}} = \operatorname*{arg\,min}_{f \in H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} \sum_{j=1}^{m} \left(\lambda_j \|f|_{A_j}\|^2_{H_{\gamma_j}(A_j)} + \frac{1}{n} \sum_{i \in I_j} L\left(y_i, f|_{A_j}(x_i)\right) \right)$$

and we see that the learning objective is minimized for $f \in H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$ if and only if

$$f|_{A_j} = \operatorname*{arg\,min}_{g \in H_{\gamma_j}(A_j)} \frac{n\lambda_j}{n_j} \|g\|^2_{H_{\gamma_j}(A_j)} + \frac{1}{n_j} \sum_{i \in I_j} L(y_i, g(x_i)), \tag{3.4}$$

where $n_j := |I_j|$ and we assume that $n_j \geq 1$ for all $j = 1, \ldots, m$, i.e. every cell contains at least one data point. Since the minimizing function in (3.4) is unique, we can conclude that $f_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}|_{A_j} = f_{D_j,\lambda'_j,\gamma_j}$, where $D_j = ((x_i, y_i))_{i \in I_j}$, $\lambda'_j = n\lambda_j/n_j$ and

$$f_{D_j,\lambda'_j,\gamma_j} = \operatorname*{arg\,min}_{f \in H_{\gamma_j}(A_j)} \lambda'_j \|f\|_{H_{\gamma_j}(A_j)} + \frac{1}{n_j} \sum_{i \in I_j} L(y_i, f(x_i))$$

is the regularized empirical risk minimizer using the standard Gaussian RKHS and the dataset D_j.

Regarding the construction of the partition, we will consider a Voronoi partition of the input space X, based on a set of center points $C = \{c_1, \ldots, c_m\}$, where the center points are a subset of the feature vectors $V = \{x_1, \ldots, x_n\}$ of our dataset $((x_1, y_1), \ldots, (x_n, y_n))$. To this end, recall that in a Voronoi partition $\mathcal{A} = (A_j)_{j=1,\ldots,n}$ with respect to the centers $C = \{c_1, \ldots, c_m\}$ each cell A_j consists of all the points $x \in X$ that have c_j as the closest center, where we break ties in favor of a smaller index j of the center c_j. We consider a set of centers C constructed by the farthest first traversal (FFT) algorithm, see Algorithm 1. We will denote the learning method (3.3) using the partition into m cells constructed in this manner by $\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}$.

Algorithm 1 Farthest First Traversal

Require: $V = \{v_1, \ldots, v_n\} \subset \mathbb{R}^d, k \leq n$
  $C \leftarrow \{v_1\}$
  while $|C| < k$ **do**
    $C \leftarrow C \cup \{c\}$ for $c \in V \backslash C$ with maximum distance to C
  end while
  return C

The farthest first traversal algorithm has a time complexity of $\mathcal{O}(kn)$ and gives an approximate solution to the metric k-center problem [27, Theorem 4.3]. Recall that the task in the metric k-center problem is to find a set $C \subset V$ with $|C| = k$, which minimizes

$$\max_{v \in V} \min_{c \in C} \|v - c\|, \tag{3.5}$$

that is, to find a set of k centers such that the maximum distance from any $v \in V$

to its closest center is minimized. Solving the metric k-center exactly is NP-hard, the solution computed by FFT is within a factor of 2 of the optimal value of (3.5). Translated to the language of entropy numbers, FFT constructs a covering of the finite set V, whose radii are bounded by $2\varepsilon_m(V)$. A key step in the analysis of the statistical error of the estimator (3.3), see Corollary 3.3.7, will be to show, that the center points C, which are by construction a $2\varepsilon_m(V)$-net of V, are also a covering of the order of $\varepsilon_m(V)$ of the *whole* support S. More precisely, we will show that under Assumption 3.1.3 $\varepsilon_m(S)$ is with sufficiently high probability within a constant range of $\varepsilon_m(V)$, when V is sampled from μ^n, see Corollary 3.3.5.

Furthermore, we will again consider a training validation procedure for adaptive hyperparameter selection, which we perform for LSVMs independently on each cell. To this end, recall that we split our dataset into a training set $D_1 := ((x_1, y_1), \ldots, (x_l, y_l))$ and a validation set $D_2 := ((x_{l+1}, y_{l+1}), \ldots, (x_n, y_n))$, where $l := \lfloor n/2 \rfloor + 1$ and pick finite sets of candidate values Λ_n, Γ_n for λ_j and γ_j. We then compute the decision functions $f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}$ for all $\boldsymbol{\lambda} \in \Lambda_n^m, \boldsymbol{\gamma} \in \Gamma_n^m$ using the training set D_1 and pick the hyperparameters $\boldsymbol{\lambda}_{D_2}, \boldsymbol{\gamma}_{D_2}$ which perform best on the validation set D_2, that is our final decision function $\widehat{f}_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}$ is defined by

$$\sum_{i=l+1}^{n} L\left(y_i, \widehat{f}_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}(x_i)\right) = \min_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in \Lambda_n^m \times \Gamma_n^m} \sum_{i=l+1}^{n} L\left(y_i, \widehat{f}_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}(x_i)\right). \tag{3.6}$$

We call the resulting learning method a TV-LSVM. Note that, since the validation step is executed independently on each cell, this amounts to a total number of $m|\Lambda_n \times \Gamma_n|$ training runs that need to be performed, instead of $|\Lambda_n \times \Gamma_n|^m$. In Sections 3.4 and 3.5 we will show that it is sufficient for the candidate sets Λ_n, Γ_n to grow logarithmically in the sample size n in order to achieve optimal learning rates. In contrast, [7, 41] consider a similar training validation procedure for kernel partitioning methods, but they require the candidate sets to grow at least linearly in n, which makes the validation step computationally infeasible.

Remark 3.2.1. In the results of Sections 3.3, 3.4, and 3.5 the reader will notice, that the regularization parameters $\lambda_1, \ldots, \lambda_m$ and the bandwidths $\gamma_1, \ldots, \gamma_m$ are chosen identically on each cell, i.e. $\lambda_1 = \ldots = \lambda_m$ and $\gamma_1 = \ldots = \gamma_m$. The reason for this is, that the asymptotically optimal choices for the parameters $\lambda_1, \ldots, \lambda_m$ and $\gamma_1, \ldots, \gamma_m$ are determined by the *global* regularity properties of the data generating distribution $\mathbf{P}$. We illustrate this in the case of least squares regression, where regularity is measured by the smoothness of the Bayes decision function $f^*_{L,\mathbf{P}}$. As-

sume that $f^*_{L,\mathbf{P}} \in C^\alpha(X)$, but restricted to some subset $X' \subset X$ the Bayes function $f^*_{L,\mathbf{P}}|_{X'}$ is α'-times differentiable with $\alpha' \gg \alpha$. Then for cells A_j contained in X', this property of $f^*_{L,\mathbf{P}}$ can be utilized by an individual adjustment of the hyperparameters on these cells, and in turn improve the overall generalization performance of the estimator. However, the *asymptotic* behavior of the excess risk (on which we focus in our theoretical results) can not be improved by this individual choice and is bottlenecked by the global degree of smoothness of $f^*_{L,\mathbf{P}}$. Additionally, an identical choice of $\lambda_1, \dots, \lambda_m$ and $\gamma_1, \dots, \gamma_m$ across all cells greatly simplifies the expressions in our theoretical analysis. For an example of a set of regularity assumptions in binary classification, where an individual choice of the hyperparameters on the cells actually leads to an improved asymptotic behavior of the learning rate, we refer to [7].

3.3 A General Oracle Inequality

The proof of our general oracle inequality in this section requires some preliminary results. The following three lemmas collect some basic results of the (inner) entropy numbers $\varepsilon_m(A)$. Note that by definition of $\varepsilon_m(A)$, where $A \subset \mathbb{R}^d$ is some non-empty set, for every $\epsilon > \varepsilon_m(A)$ there exists an ϵ-net $N \subset A$ of A. The content of the next lemma is that for compact sets A this also holds for $\epsilon = \varepsilon_m(A)$.

Lemma 3.3.1. *Let $A \subset \mathbb{R}^d$ be compact. Then for every $m \in \mathbb{N}$ there exists an $\varepsilon_m(A)$-net $N \subset A$ of A with $|N| = m$.*

Proof. For $n \in \mathbb{N}$ let $x_{1,n}, \dots, x_{m,n} \in A$ be an $(\varepsilon_m(A) + 1/n)$-net of A. By compactness of A, each sequence $(x_{j,n})_{n\in\mathbb{N}}$ has an accumulation point $x_j \in A, j = 1, \dots, m$. These accumulation points are an $\varepsilon_m(A)$-net, since for all $x \in A$ we have

$$\begin{aligned}\min_{j=1,\dots,m} \|x - x_j\| &\leq \min_{j=1,\dots,m} \|x - x_{j,n}\| + \|x_{j,n} - x_j\| \\ &\leq \min_{j=1,\dots,m} \varepsilon_m(A) + \frac{1}{n} + \|x_{j,n} - x_j\| \\ &= \varepsilon_m(A) + \frac{1}{n} + \min_{j=1,\dots,m} \|x_{j,n} - x_j\|.\end{aligned}$$

Taking the infimum over $n \in \mathbb{N}$ then yields the assertion. □

Obviously, the outer entropy numbers are monotone with respect to inclusion of subsets. The following lemma shows that for the inner entropy numbers we still

have a slightly weaker form of this monotonicity.

Lemma 3.3.2. *For $A \subset B \subset \mathbb{R}^d$ we have $\varepsilon_m(A) \leq 2\varepsilon_m(B)$ for all $m \in \mathbb{N}$.*

Proof. Let $x_1, \ldots, x_m \in B$ be an ε-net of B. For each $j = 1, \ldots, m$ pick an $y_j \in A$ with $\|x_j - y_j\| \leq \varepsilon$, if such an y_j exists and else let $y_j \in A$ be an arbitrary point. Then, by the triangle inequality $y_1, \ldots, y_m$ is a 2ε-net of A. □

Lemma 3.3.3. *Let $A \subset \mathbb{R}^d$ be compact. Then we have $\varepsilon_m(A) \leq \operatorname{diam} A$ for all $m \in \mathbb{N}$.*

Proof. By monotonicity of $\varepsilon_m(A)$ (with respect to m) it suffices to prove the statement for $m = 1$. Let $x \in A$ with $A \subset B_\varepsilon(x)$ for $\varepsilon = \varepsilon_1(A)$, cf. Lemma 3.3.1. Then we have

$$\varepsilon = \sup_{y \in A} \|x - y\| \leq \sup_{y,z \in A} \|z - y\| = \operatorname{diam} A.$$

□

The following lemma will help us to show, that a covering of the data points $x_1, \ldots, x_n$ is with high probability also a covering of the whole support S under Assumption 3.1.3, which is the statement of the corollary thereafter.

Lemma 3.3.4. *Let Assumption 3.1.3 be satisfied for the constants $C_\mu \geq 1$ and $\delta > 0$. Then we have*

$$\mu^n \left(x_1, \ldots, x_n : \sup_{x \in S} \min_{i=1,\ldots,n} \|x_i - x\| > \tau \right) \leq m \exp\big(- C_\mu^{-1} (\tau - \varepsilon_m(S))^\delta n \big)$$

for all $\varepsilon_m(S) < \tau \leq \varepsilon_m(S) + \operatorname{diam} S$.

Proof. First note that $\min_{i=1,\ldots,n} \|x_i - x\| > \tau$ if and only if $\|x_i - x\| > \tau$ for all $i = 1, \ldots, n$, which implies

$$\mu^n \left(x_1, \ldots, x_n : \min_{i=1,\ldots,n} \|x_i - x\| > \tau \right) = \big(1 - \mu(B_\tau(x))\big)^n \tag{3.7}$$

for all $x \in S$ and $\tau > 0$. With the help of Lemma 3.3.1 let $N \subset S$ be an $\varepsilon_m(S)$-net of S with $|N| = m$. Now, for every $x \in S$ there exists an $x' \in N$ such that

$$\min_{i=1,\ldots,n} \|x_i - x\| \leq \min_{i=1,\ldots,n} \|x_i - x'\| + \|x' - x\| \leq \min_{i=1,\ldots,n} \|x_i - x'\| + \varepsilon_m(S),$$

which combined with (3.7) implies

$$\begin{aligned}\mu^n\left(\sup_{x\in S}\min_{i=1,\dots,n}\|x_i-x\|>\tau\right) &\le \mu^n\left(\max_{x\in N}\min_{i=1,\dots,n}\|x_i-x\|+\varepsilon_m(S)>\tau\right)\\ &\le \sum_{x\in N}\big(1-\mu(B_{\tau-\varepsilon_m(S)}(x))\big)^n\end{aligned}$$

for $\tau-\varepsilon_m(S)>0$. Using Assumption 3.1.3 we can further bound this by

$$\begin{aligned}\sum_{x\in N}\big(1-\mu(B_{\tau-\varepsilon_m(S)}(x))\big)^n &\le m\big(1-C_\mu^{-1}(\tau-\varepsilon_m(S))^\delta\big)^n\\ &\le m\exp\big(-C_\mu^{-1}(\tau-\varepsilon_m(S))^\delta n\big)\end{aligned}$$

for $\tau-\varepsilon_m(S)\le \operatorname{diam} S$, which proves the assertion. □

Corollary 3.3.5. *Let Assumption 3.1.3 be satisfied for the constants $C_\mu\ge 1$ and $\delta>0$ and let $\mathcal{A}=(A_j)_{j=1,\dots,m}$ be an FFT partition for some $m\le n$ with respect to the centers $c_1,\dots,c_m$. Then we have $A_j\cap S\subset B_{6\varepsilon_m(S)}(c_j)$ for all $j=1,\dots,m$ simultaneously with probability μ^n not less than*

$$1-m\exp\big(-C_\mu^{-1}n\varepsilon_m(S)^\delta\big).$$

Proof. For $x\in S$ let $c(x)\in C$ be its respective Voronoi center and let $D=\{x_1,\dots,x_n\}$. Recall, that since the FFT algorithm produces a 2-approximation of the metric k-center problem, we have $\|x_i-c(x_i)\|\le 2\varepsilon_m(D)$ for all $i=1,\dots,n$. Consequently, we can estimate

$$\begin{aligned}\|x-c(x)\| &= \min_{i=1,\dots,n}\|x-c(x_i)\| \le \min_{i=1,\dots,n}\|x-x_i\|+\|x_i-c(x_i)\|\\ &\le \min_{i=1,\dots,n}\|x-x_i\|+2\varepsilon_m(D)\le \min_{i=1,\dots,n}\|x-x_i\|+4\varepsilon_m(S),\end{aligned}$$

where in the last step we used Lemma 3.3.2. Applying Lemma 3.3.4 with $\tau=2\varepsilon_m(S)$ subsequently gives us

$$\sup_{x\in S}\|x-c(x)\|\le 6\varepsilon_m(S)$$

with probability not less than

$$1-m\exp\big(-C_\mu^{-1}n\varepsilon_m(S)^\delta\big).$$

Note that the prerequisite $\varepsilon_m(S) < \tau \leq \varepsilon_m(S) + \operatorname{diam} S$ of Lemma 3.3.4 is fulfilled for $\tau = 2\varepsilon_m(S)$ because of Lemma 3.3.3. □

The subsequent lemma is a tool for bounding the covering numbers of the composite kernel $H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$ based on the covering numbers of the individual Gaussian RKHSs $H_{\gamma_j}(A_j)$, for which we already developed the necessary bounds.

Lemma 3.3.6. *Assume $\mathcal{A} = (A_j)_{j=1,\ldots,m}$ is a partition of X such that there exist constants $a_1, \ldots, a_m > 0$ and $q > 0$ with*

$$e_i(\mathrm{id} : H_{\gamma_j}(A_j) \to \ell_\infty(A_j \cap S)) \leq a_j \, i^{-\frac{1}{q}}.$$

for all $i \in \mathbb{N}$ and $j = 1, \ldots, m$. Then we have

$$e_i(\mathrm{id} : H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A}) \to \ell_\infty(S)) \leq (3 \log 4)^{\frac{1}{q}} \left(\sum_{j=1}^m \left(\frac{a_j}{\sqrt{\lambda_j}} \right)^q \right)^{\frac{1}{q}} i^{-\frac{1}{q}} \text{ for all } i \in \mathbb{N}.$$

Proof. By Lemma 1.3.10 we have

$$\log \mathcal{N}_{\ell_\infty(A_j \cap S)} \left(B_{H_{\gamma_j}(A_j)}, \varepsilon \right) \leq \log(4) \left(\frac{a_j}{\varepsilon} \right)^q$$

for all $\varepsilon > 0$ and hence

$$\log \mathcal{N}_{\ell_\infty(A_j \cap S)} \left(\lambda_j^{-\frac{1}{2}} B_{H_{\gamma_j}(A_j)}, \varepsilon \right) \leq \log(4) \left(\frac{a_j}{\varepsilon \sqrt{\lambda_j}} \right)^q,$$

which yields

$$\begin{aligned} \log \mathcal{N}_{\ell_\infty(S)} \left(B_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})}, \varepsilon \right) &\leq \sum_{j=1}^m \log \mathcal{N}_{\ell_\infty(A_j \cap S)} \left(\lambda_j^{-\frac{1}{2}} B_{H_{\gamma_j}(A_j)}, \varepsilon \right) \\ &\leq \sum_{j=1}^m \log(4) \left(\frac{a_j}{\varepsilon \sqrt{\lambda_j}} \right)^q \end{aligned}$$

where in the first estimate we used Lemma 1.3.14. Finally, we again turn this into a bound on the dyadic entropy numbers using Lemma 1.3.10, which completes the proof. □

Corollary 3.3.7. *Let Assumptions 3.1.2 and 3.1.3 be satisfied and let $\mathcal{A}$ be an FFT partition of m cells constructed from points $x_1, \ldots, x_n$ sampled from μ^n. Then there*

exists a constant $K > 0$ such that

$$e_i(\mathrm{id} : H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A}) \to \ell_\infty(S)) \leq (3 \cdot 6^\varrho \log(4) K C_S^{\varrho+1})^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \lambda^{-\frac{1}{2}} \gamma^{-\frac{\varrho}{2p}} i^{-\frac{1}{2p}}$$

with probability not less than $1 - \exp(-C_\mu^{-1} C_S^{-\delta} n m^{-\delta/\varrho})$ for all $\gamma_1 = \ldots = \gamma_m =: \gamma \leq m^{-1/\varrho}$, $\lambda_1 = \ldots = \lambda_m =: \lambda$, and $p \in (0,1)$.

Proof. We will apply Lemma 3.3.6 for suitable constants a_j. To this end, note that by Theorem 1.3.16 we can choose

$$a_j = K^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \mathcal{N}(S \cap A_j, \gamma)^{\frac{1}{2p}}. \tag{3.8}$$

Further note that by Corollary 3.3.5 we have $S \cap A_j \subset S \cap B_r(c_j)$ with probability not less than $1 - \exp(-C_\mu^{-1} C_S^{-\delta} n m^{-\delta/\varrho})$, where $r := \max\{m^{-1/\varrho}, 6\varepsilon_m(S)\}$. Consequently, by Assumption 3.1.2 we have

$$\begin{aligned} a_j &\leq K^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \mathcal{N}(S \cap B_r(c_j), \gamma)^{\frac{1}{2p}} \leq K^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \left(C_S \left(\frac{\gamma}{r} \right)^{-\varrho} \right)^{\frac{1}{2p}} \\ &\leq \left(6^\varrho C_S^{\varrho+1} K \right)^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \gamma^{-\frac{\varrho}{2p}} m^{-\frac{1}{2p}} \end{aligned}$$

for $\gamma \leq m^{-1/\varrho} \leq r$. Lemma 3.3.6 then finally gives us

$$e_i(\mathrm{id} : H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A}) \to \ell_\infty(S)) \leq (3 \cdot 6^\varrho \log(4) K C_S^{\varrho+1})^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \lambda^{-\frac{1}{2}} \gamma^{-\frac{\varrho}{2p}} i^{-\frac{1}{2p}}$$

with probability not less than $1 - \exp(-C_\mu^{-1} C_S^{-\delta} n m^{-\delta/\varrho})$ for $\gamma \leq m^{-1/\varrho}$. □

With all the previous preparations we can now formulate and proof a general oracle inequality for the estimator (3.3) using a FFT Voronoi partition.

Theorem 3.3.8. *Assume L is a locally Lipschitz continuous loss that can be clipped at $M > 0$ and that the supremum and variance bounds are satisfied for constants $B > 0, \vartheta \in [0,1]$, and $V \geq B^{2-\vartheta}$. Furthermore, let $\mathbf{P}$ satisfy Assumptions 3.1.2 and 3.1.3 and fix an $f_0 \in H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$ and a $B_0 \geq B$ with $\|L \circ f_0\|_\infty \leq B_0$, where $H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$ is constructed using an independent sample of size n and $m < n$ cells. Then there exists a constant K such that for all $n \in \mathbb{N}, \gamma_1 = \ldots = \gamma_m =: \gamma \in (0, m^{-1/\varrho}), \lambda_1 =$*

$\ldots = \lambda_m =: \lambda > 0, p \in (0, 1/2]$ *and* $\tau > 0$ *we have*

$$\begin{aligned}
&\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \\
\leq\, &9(\|f_0\|^2_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}}) \\
&+ KC_{\mathbf{P},m}\left(\frac{p^{-d-1}\gamma^{-\varrho}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}
\end{aligned}$$

with probability not less than $1 - 3\mathrm{e}^{-\tau} - m\exp(-C_\mu^{-1}C_S^{-\delta}nm^{-\delta/\varrho})$*, where*

$$C_{\mathbf{P},m} = \max\left\{B, \left(|L|^p_{M,1}V^{\frac{1-p}{2}}\right)^{\frac{2}{2-p-\vartheta+\vartheta p}}, |L|^p_{M,1}B^{1-p}, 1\right\} \cdot \max\left\{C_S^{\frac{\varrho+1}{2p}}, B^{2p}\right\}^{\frac{1}{2-p-\vartheta+\vartheta p}}.$$

Proof. By Corollary 3.3.7 we have

$$\mathbf{E}_{D\sim\mathbf{P}^n} e_i(\mathrm{id} : H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A}) \to L_2(\mathbf{D})) \leq (3\cdot 6^\varrho \log(4) K C_S^{\varrho+1})^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \lambda^{-\frac{1}{2}} \gamma^{-\frac{\varrho}{2p}} i^{-\frac{1}{2p}}$$

with probability not less than $1 - \exp(-C_\mu^{-1}C_S^{-\delta}nm^{-\delta/\varrho})$ for $\gamma \leq m^{-1/\varrho}$. By Theorem 1.3.17 we then have

$$\begin{aligned}
\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \leq\, &9(\|f_0\|^2_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}}) \\
&+ K\left(\frac{a^{2p}}{n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}
\end{aligned}$$

with probability not less than

$$(1 - 3\mathrm{e}^{-\tau})(1 - m\exp(-C_\mu^{-1}C_S^{-\delta}nm^{-\delta/\varrho})) \geq 1 - \mathrm{e}^{-\tau} - m\exp(-C_\mu^{-1}C_S^{-\delta}nm^{-\delta/\varrho}),$$

where

$$a := p^{-\frac{d+1}{2p}}\lambda^{-\frac{1}{2}}\gamma^{-\frac{\varrho}{2p}}\max\{C_S^{\varrho+1}, B\}K_d^{\frac{1}{2p}}$$

and K satisfies

$$K \leq \tilde{K}\max\left\{B, \left(|L|^p_{M,1}V^{\frac{1-p}{2}}\right)^{\frac{2}{2-p-\vartheta+\vartheta p}}, |L|^p_{M,1}B^{1-p}, 1\right\}$$

for a universal constant $\tilde{K}$ by the short remark after Theorem 1.3.17 for all $p \in (0, 1/2]$. □

3.4 Least-Squares Regression

To simplify the presentation of the results of this sections compared to Section 2.3, we will formulate our assumptions for bounding the approximation error only in terms of classical differentiability of the Bayes function $f^*_{L,\mathbf{P}}$ instead of a Besov semi-norm. To this end, recall Definition 2.3.2 of the Hölder spaces $C^{k,\beta}(\mathbb{R}^d)$.

The following theorem contains our first main result of this section, an oracle inequality for LSVMs using the least-squares loss.

Theorem 3.4.1. *Let $\mathbf{P}$ satisfy Assumption 3.1.2 for the constants C_S and ϱ and Assumption 3.1.3 for the constants C_μ and δ. Further assume that $Y \subset [-M, M]$ as well as $f^*_{L,\mathbf{P}} \in C^{k,\beta}(\mathbb{R}^d) \cap L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ for some $k \in \mathbb{N}_0$ and $\beta \in [0,1)$ and set $\alpha := k + \beta$. Consider the estimator $\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}$ using the least-squares loss $L = L_{\mathrm{LS}}$ for some $m \leq n$ and hyperparameters $\lambda_1 = \ldots = \lambda_m =: \lambda$ and $\gamma_1 = \ldots = \gamma_m =: \gamma$. Then for all $\tau > 0, n > 1, \lambda \in (0,1)$, and $\gamma \in (0, m^{-1/\varrho}]$ we have*

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \leq & c_1 \|f^*_{L,\mathbf{P}}\|^2_{L_2(\mathbb{R}^d)} m\lambda\gamma^{-d} + c_2 |f^*_{L,\mathbf{P}}|^2_{C^{k,\beta}(\mathbb{R}^d)} \gamma^{2\alpha} \\ & + c_3 K \lambda^{-1/\log n} \gamma^{-\varrho} n^{-1} \log^{d+1} n + c_4 \frac{\tau}{n}\end{aligned}$$

with probability not less than $1 - 3\mathrm{e}^{-\tau} - m\exp(-C_\mu^{-1} C_S^{-\delta} n m^{-\delta/\varrho})$, where $c_1 = 9\pi^{-d/2} 4^{k+1}$,

$$\begin{aligned}c_2 &= 9\left(\frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^2 2^{-\alpha} d^k, c_3 = \max\left\{C_S^{\varrho+1}, 4M^2\right\} \max\left\{16M^2, 1\right\}, \\ c_4 &= 3456M^2 + 15\max\left\{(2^{k+1}\|f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} + M)^2, 4M^2\right\},\end{aligned}$$

and K is a constant independent of $\mathbf{P}, \lambda, \gamma, n$, and m.

Proof. The least-squares loss satisfies a supremum/variance bound for $B = 4M^2$, $V = 16M^2$ and $\vartheta = 1$ as well as $|L|_{M,1} = 4M$. Theorem 3.3.8 gives us

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \leq & 9(\|f_0\|^2_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}}) \\ & + K C_{\mathbf{P},m} \frac{p^{-d-1}\gamma^{-\varrho}}{\lambda^p n} + \frac{(3456M^2 + 15B_0)\tau}{n}\end{aligned}$$

with probability not less than $1 - 3\mathrm{e}^{-\tau} - m\exp(-C_\mu^{-1}C_S^{-\delta}nm^{-\delta/\varrho})$, where

$$C_{\mathbf{P},m} \leq \max\left\{16M^2, 1\right\} \max\left\{C_S^{\varrho+1}, (4M^2)^{2p}\right\}.$$

Since $n \geq 2$ we have $p = \log 2/(2\log n) \leq 1/2$ and the second factor above can be bounded by $\max\left\{C_S^{\varrho+1}, 4M^2\right\}$.Further we have

$$p^{-d-1}\lambda^{-p} \leq (2/\log 2)^{d+1}\log^{d+1} n\lambda^{-1/\log n}$$

for $\lambda \leq 1$. To complete the proof we need to pick a suitable function $f_0 \in H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$. To this end, note that for $\gamma = \gamma_j, \lambda = \lambda_j, j = 1, \ldots, m$ we have $H_\gamma(X) \subset H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$ and $\|f\|^2_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} \leq m\lambda\|f\|^2_{H_\gamma(X)}$ for all $f \in H_\gamma(X)$. By combining Lemma 2.3.5 and Proposition 2.3.3 there exists an $f_0 \in H_\gamma(X)$ with

$$\mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}} \leq \left(\frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^2 2^{-\alpha}d^k|f_0|^2_{C^{k,\beta}(\mathbb{R}^d)}\gamma^{2\alpha}$$

and $\|f_0\|^2_{H_\gamma(X)} \leq \pi^{-d/2}4^{k+1}\|f_0\|_{L_2(\mathbb{R}^d)}\gamma^{-d}$ Lemma 2.3.6 which completes the proof. □

The following corollary shows that an LSVM using the least-squares loss and an FFT partition achieves the same optimal rates of a global SVM as long as the number of cells does not grow too fast with the sample size.

Corollary 3.4.2. *Let the assumptions of Theorem 3.4.1 be satisfied with the number of cells specified as $m = \lceil n^\sigma \rceil$ for some $\sigma < 1$. Assume that $|f^*_{L,\mathbf{P}}|_{C^{k,\beta}(\mathbb{R}^d)} \leq C_1$, $\|f^*_{L,\mathbf{P}}\|_{L_2(\mathbb{R}^d)} \leq C_2$, and $\|f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} \leq C_3$ for some finite constants C_1, C_2, C_3 and that the parameters from Theorem 3.4.1 satisfy*

$$\sigma < \min\left\{\frac{\varrho}{2\alpha+\varrho}, \frac{\varrho}{\delta}\right\}.$$

Setting $\gamma = n^{-a}$ and $\lambda = n^{-b}$ with $a = 1/(2\alpha+\varrho)$ and $b \geq \sigma + (2\alpha+d)/(2\alpha+\varrho)$ there exists a constant $C > 0$ only depending on $C_1, C_2, C_3, C_\mu, C_S$ and M such that for all $n > 1$ and $\tau \geq 1$ we have

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{2\alpha}{2\alpha+\varrho}}\log^{d+1} n,$$

with probability not less than $1 - \mathrm{e}^{-\tau} - n^\sigma\exp(-C_\mu^{-1}C_S^{-\delta}n^{1-\sigma\delta/\varrho})$.

Proof. We apply Theorem 3.4.1 with the specified values for λ and γ. Examining the summands in the bound given in Theorem 3.4.1, ignoring constants for the moment, we see that for $m = \lceil n^\sigma \rceil$ and $\gamma = n^{-a}, \lambda = n^{-b}$ with $a = 1/(2\alpha + \varrho), b \geq \sigma + (2\alpha + d)/(2\alpha + \varrho)$ we have

$$m\lambda\gamma^{-d} \leq 2n^\sigma n^{-b} n^{ad} \leq 2n^{-\frac{2\alpha}{2\alpha+\varrho}},$$

$$\gamma^{2\alpha} = n^{-2\alpha a} = n^{-\frac{2\alpha}{2\alpha+\varrho}},$$

$$\lambda^{-1/\log n}\gamma^{-\varrho} n^{-1} \log^{d+1} n = n^{b/\log n} n^{a\varrho} n^{-1} \log^{d+1} n = \mathrm{e}^b n^{-\frac{2\alpha}{2\alpha+\varrho}} \log^{d+1} n.$$

That is, every summand is of the order of $n^{-2\alpha/(2\alpha+\varrho)} \log^{d+1} n$. Note that the constraint $\gamma \leq m^{-1/\varrho}$ from Theorem 3.4.1 is fulfilled since $\sigma < \varrho/(2\alpha + \varrho)$. □

Note that since $1 - \sigma\delta/\varrho > 0$ we have $n^\sigma \exp(-C_\mu^{-1} C_S^{-\delta} n^{1-\sigma\delta/\varrho}) \to 0$ in the probability in Corollary 3.4.2. Also note that the learning rates in Corollary 3.4.2 coincide with the learning rates from Section 2.3. We again show that the same rates can be achieved by an adaptive hyperparameter selection.

Theorem 3.4.3. *Let the assumptions of Theorem 3.4.1 be satisfied with $\varrho \geq 1$ and the number of cells specified as $m = \lceil n^\sigma \rceil$ for some $\sigma < 1$. Let A_n be a minimal $1/\log n$-net of $(0, 1]$ with $1 \in A_n$ and let B_n be a minimal $1/\log n$-net of $[\sigma+1, \sigma+d]$ with $\sigma + d \in B_n$ and set $\Gamma_n := \{n^{-a} : a \in A_n\}$ and $\Lambda_n := \{n^{-b} : b \in B_n\}$. Assume that $|f^*_{L,\mathbf{P}}|_{C^{k,\beta}(\mathbb{R}^d)} \leq C_1$, $\|f^*_{L,\mathbf{P}}\|_{L_2(\mathbb{R}^d)} \leq C_2$, and $\|f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} \leq C_3$ for some constants $C_{1,2,3}$ and that the parameters from Theorem 3.4.1 satisfy*

$$\sigma < \min\left\{\frac{\varrho}{2\alpha + \varrho}, \frac{\varrho}{\delta}\right\}.$$

Then there exists a constant $C > 0$ only depending on $C_1, C_2, C_3, C_\mu, C_S$, and M such that for all $n > \exp\frac{2\varrho}{\varrho-\sigma}$ and $\tau \geq 1$ we have

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{2\alpha}{2\alpha+\varrho}} \log^{d+1} n$$

with probability not less than $1 - \mathrm{e}^{-\tau} - n^\sigma \exp(-C_\mu^{-1} C_S^{-\delta} n^{1-\sigma\delta/\varrho})$ where the hyperparameters $\boldsymbol{\lambda}_{D_2}, \boldsymbol{\gamma}_{D_2}$ are selected by the training validation procedure (3.6).

Proof. By [55, Theorem 7.2], an oracle inequality for empirical risk minimization,

we have

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\hat{f}_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} &\leq 6 \min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in\Lambda_n^m\times\Gamma_n^m} \left(\mathcal{R}_{L,\mathbf{P}}(\hat{f}_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}}\right) \\ &\quad + \frac{512M^2(\tau+\log(1+|\Lambda_n^m\times\Gamma_n^m|))}{n-l} \\ &\leq 6\left(\mathcal{R}_{L,\mathbf{P}}(\hat{f}_{D_1,\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*}) - \mathcal{R}^*_{L,\mathbf{P}}\right) \\ &\quad + \frac{2048M^2(\tau+\log(1+|\Lambda_n^m\times\Gamma_n^m|))}{n}\end{aligned} \tag{3.9}$$

with probability $\mathbf{P}^{n-l}$ not less than $1-\mathrm{e}^{-\tau}$, where in the last step we picked values $\boldsymbol{\gamma}^*\in\Gamma_n^m$ and $\boldsymbol{\lambda}^*\in\Lambda_n^m$ which we will specify in a moment. We again only consider $\lambda_1=\ldots=\lambda_m=:\lambda$ and $\gamma_1=\ldots=\gamma_m=:\gamma$. Since $\sigma<\varrho/\delta$, by Theorem 3.4.1 there exists a constant $C>0$ independent of λ,γ and n (see also proof of Corollary 3.4.2) such that

$$\begin{aligned}\mathcal{R}_{L,\mathbf{P}}(\hat{f}_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} &\leq C\left(m\lambda\gamma^{-d}+\gamma^{2\alpha}+\lambda^{-1/\log n}\gamma^{-\varrho}l^{-1}\log^{d+1}n+\frac{\tau}{l}\right) \\ &\leq C\left(m\lambda\gamma^{-d}+\gamma^{2\alpha}+2\mathrm{e}^{\sigma+d}\gamma^{-\varrho}n^{-1}\log^{d+1}n+\frac{2\tau}{n}\right)\end{aligned}$$

with probability $\mathbf{P}^{n+l}$ not less than $1-3\mathrm{e}^{-\tau}-n^\sigma\exp(-C_\mu^{-1}C_S^{-\delta}n^{1-\sigma\delta/\varrho})$ for $\lambda\in\Lambda_n$ and $\gamma\in\Gamma_n\cap(0,m^{-1/\varrho})$. Since A_n is an $1/\log n$-net of $(0,1]$ we have $\Gamma_n\cap(0,m^{-1/\varrho})\neq\emptyset$ for $1-\sigma/\varrho>2/\log n$ and since $\sigma<\varrho/(2\alpha+\varrho)$ we can choose $a_*\in A_n$ such that $\gamma=n^{-a_*}\in(0,m^{-1/\varrho})$ and $1/(2\alpha+\varrho)\leq a_*\leq 1/(2\alpha+\varrho)+2/\log n$. That is, by choosing $\gamma=n^{-a_*}$ and $\lambda=n^{-\sigma-d}$ as $\boldsymbol{\gamma}^*,\boldsymbol{\lambda}^*$ we have

$$\mathcal{R}_{L,\mathbf{P}}(\hat{f}_{D_1,\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*,\mathrm{FFT}(m)}) \leq C\left(n^{-\frac{2\alpha}{2\alpha+\varrho}}+\mathrm{e}^{c+d+4\alpha}n^{-\frac{2\alpha}{2\alpha+\varrho}}\log^{d+1}n+\frac{2\tau}{n}\right)$$

with probability not less than $1-3\mathrm{e}^{-\tau}-n^\sigma\exp(-C_\mu^{-1}C_S^{-\delta}n^{1-\sigma\delta/\varrho})$. Combining this with (3.9) we get using $|\Lambda_n^m\times\Gamma_n^m|\lesssim\log^{2m}n$ that

$$\mathcal{R}_{L,\mathbf{P}}(\hat{f}_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \leq c_1\left(n^{-\frac{2\alpha}{2\alpha+\varrho}}\log^{d+1}n+\frac{\tau}{n}\right)+c_2\left(\frac{\tau}{n}+\frac{m\log\log n}{n}\right)$$

with probability not less than

$$(1-\mathrm{e}^{-\tau})(1-3\mathrm{e}^{-\tau}-n^\sigma\exp(-C_\mu^{-1}C_S^{-\delta}n^{1-\sigma\delta/\varrho}))\geq(1-4\mathrm{e}^{-\tau}-n^\sigma\exp(-C_\mu^{-1}C_S^{-\delta}n^{1-\sigma\delta/\varrho})).$$

Noting that $m/n = 2n^{\sigma-1} \leq 2n^{-2\alpha/(2\alpha+\varrho)}$ and some elementary transformations yield the assertion. □

Remark 3.4.4. The constraint on σ in Corollary 3.4.2 and Theorem 3.4.3 can be interpreted as follows: The user specifies a value for σ depending on some computational time and space constraints given by the available resources. The condition on σ then specifies the set of distributions for which we can achieve the optimal learning rate. The smaller we choose σ, the larger this class of distributions becomes but small values of σ in turn diminish the computational speed up. As a consequence we have a fundamental trade-off between computational benefit and the fastest achievable learning rate. Indeed, the fastest possible learning rate is given by $n^{\sigma-1}$, which can be seen by the bounds on the statistical error of the validation step in the proof of Theorem 3.4.3. A very similar trade-off was observed for least-squares kernel regression using random features [48], Nyström subsampling [49], and random chunking [67] as speed up strategies. In all these articles the authors consider a more abstract setting of general kernels with assumptions on the decay of the eigenvalues of the corresponding integral operator, however, they focus on the restrictive case where the Bayes decision function is assumed to be contained in the considered RKHS. None of these mentioned articles consider adaptive hyperparameter selection for achieving the same rates without knowledge on the data generating distribution.

3.5 Binary Classification

For the results of this section recall our regularity assumptions for binary classification given by Assumption 1.3.3 and 2.4.2.

Theorem 3.5.1. *Let* $\mathbf{P}$ *satisfy Assumption 3.1.2 for the constants* C_S *and* ϱ *and Assumption 3.1.3 for the constants* C_μ *and* δ*. Further let Assumption 1.3.3 be satisfied for the constants* C_* *and* q *and Assumption 2.4.2 for the constants* C_{**} *and* β*. Consider the estimator* $\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}$ *using the hinge loss* $L = L_{\text{hinge}}$ *for some* $m \leq n$ *and hyperparameters* $\lambda_1 = \ldots = \lambda_m =: \lambda$ *and* $\gamma_1 = \ldots = \gamma_m =: \gamma$*.*

Then for all $\tau > 0, n > 1, \lambda \in (0,1]$, and $\gamma \in [n^{-1/\varrho}, m^{-1/\varrho}]$ we have

$$\begin{aligned}
&\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \\
\leq& c_1 m\lambda\gamma^{-d} + c_2 C_{**}\gamma^{\beta} + c_3 K\lambda^{-1/\log n}\left(\frac{\gamma^{-\varrho}}{n}\right)^{\frac{q+1}{q+2}} \log^{d+1} n \\
&+ 3C_*^{\frac{q}{q+2}}\left(\frac{432\tau}{n}\right)^{\frac{q+1}{q+2}} + 30\frac{\tau}{n}
\end{aligned}$$

with probability not less than $1 - 3\mathrm{e}^{-\tau} - m\exp(-C_\mu^{-1}C_S^{-\delta}nm^{-\delta/\varrho})$, where

$$c_1 = 3^{d+2}/\Gamma(d/2{+}1), \quad c_2 = 9\frac{2^{1-\beta/2}\Gamma\left(\frac{\beta+d}{2}\right)}{\Gamma(d/2)}, \quad c_3 = \max\{C_S^{\varrho+1}, 2\}\max\left\{C_*^{q/(q+1)}, 1\right\},$$

and K is a constant independent of $\mathbf{P}, \lambda, \gamma, n$, and m.

Proof. The supremum bound is obviously satisfied for $B = 2$ and by [55, Theorem 8.24] the variance bound is satisfied for $V = 6C_*^{q/(q+1)}$ and $\vartheta = q/(q+1)$. Furthermore, it is not hard to see that $|L_{\mathrm{hinge}}|_{1,1} = 1$. Given this value for ϑ, the exponent in Theorem 3.3.8 then reads

$$\frac{1}{2-p-\vartheta+\vartheta p} = \frac{q+1}{q+2-p},$$

see also the proof of Theorem 2.4.5. An application of Theorem 3.3.8 then gives us for $\lambda > 0$, $\gamma \in (0, m^{-1/\varrho})$, and $p \in (0, 1/2]$

$$\begin{aligned}
\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \leq\, & 9\left(\|f_0\|^2_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}}\right) \\
&+ C_{\mathbf{P},m}K\left(\frac{p^{-d-1}\lambda^{-p}\gamma^{-\varrho}}{n}\right)^{\frac{q+1}{q+2-p}} + 3\left(\frac{432C_*^{\frac{q}{q+1}}\tau}{n}\right)^{\frac{q+1}{q+2}} \\
&+ \frac{15B_0\tau}{n}
\end{aligned}$$

with probability not less than $1 - 3\mathrm{e}^{-\tau} - m\exp(-C_\mu^{-1}C_S^{-\delta}nm^{-\delta/\varrho})$. With the specified values for B, V, ϑ and $|L_{\mathrm{hinge}}|_{1,1}$ we see that the first factor of the constant $C_{\mathbf{P},m}$

can be bounded by

$$\begin{aligned}&\max\left\{B, \left(|L|_{M,1}^{p} V^{\frac{1-p}{2}}\right)^{\frac{2}{2-p-\vartheta+\vartheta p}}, |L|_{M,1}^{p} B^{1-p}, 1\right\}\\ =&\max\left\{2, \left(6C_*^{\frac{q}{q+1}}\right)^{\frac{(1-p)(q+1)}{q+2-p}}, 2^{1-p}\right\} \leq \max\left\{2, 6C_*^{\frac{q}{q+1}}\right\}.\end{aligned}$$

Noting that $(q+1)/(q+2-p) \leq 1$ we see that the second factor of $C_{\mathbf{P},m}$ in Theorem 3.3.8 is bounded by $\max\left\{C_S^{\varrho+1}, 2\right\}$. Choosing $p = \log 2/(2\log n) \leq 1/2$ gives us

$$C_{\mathbf{P},m} K \left(\frac{p^{-d-1}\lambda^{-p}\gamma^{-\varrho}}{n}\right)^{\frac{q+1}{q+2-p}} \leq c_3 K \lambda^{-1/\log n} \left(\frac{\gamma^{-\varrho}}{n}\right)^{\frac{q+1}{q+2}}$$

for $\gamma^{-\varrho}/n \leq 1$ with c_3 defined as in the theorem. Finally, to bound the approximation error we need to pick a suitable function $f_0 \in H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$. To this end, note that for $\gamma = \gamma_j, \lambda = \lambda_j, j = 1, \ldots, m$ we have $H_\gamma(X) \subset H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})$ with $\|f\|_{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathcal{A})}^2 \leq m\lambda \|f\|_{H_\gamma(X)}^2$ for all $f \in H_\gamma(X)$. By Equation (8.15) in [55, Proof of Theorem 8.18] there exists a function $f_0 \in H_\gamma(X)$ with $\|f_0\|_\infty \leq 1$ and

$$\lambda \|f_0\|_{H_\gamma(X)}^2 + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}_{L,\mathbf{P}}^* \leq \frac{3^d}{\Gamma\left(\frac{d}{2}+1\right)} \lambda \gamma^{-d} + \frac{2^{1-\beta/2}\Gamma\left(\frac{\beta+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} C_{**} \gamma^{\beta}.$$

since $\|f_0\|_\infty \leq 1$ we have $B_0 = 2$ which completes the proof. □

As in the previous section, we can use the general oracle inequality above to prove the same learning rates as in Section 2.4 by choosing suitable values for the hyperparameters λ and γ provided the number of cells m does not grow too fast.

Corollary 3.5.2. *Let the assumptions of Theorem 3.5.1 be satisfied with the number of cells specified as $m = \lceil n^\sigma \rceil$ for some $\sigma < 1$. Assume the parameters from Theorem 3.5.1 satisfy*

$$\sigma < \min\left\{\frac{\varrho(q+1)}{\beta(q+2)+\varrho(q+1)}, \frac{\varrho}{\delta}\right\}.$$

Setting $\gamma_n = n^{-a}$ and $\lambda_n = n^{-b}$ with

$$a = \frac{q+1}{\beta(q+2)+\varrho(q+1)} \quad \text{and} \quad b \geq \sigma + \frac{(d+\beta)(q+1)}{\beta(q+2)+\varrho(q+1)}$$

there exists a constant $C > 0$ only depending on C_*, C_{**}, C_μ, and C_S such that for

all $n > 1$ and $\tau \geq 1$ we have

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{\beta(q+1)}{\beta(q+2)+\varrho(q+1)}} \log^{d+1} n$$

with probability not less than $1 - \mathrm{e}^{-\tau} - n^{\sigma} \exp(-C_{\mu}^{-1} C_S^{-\delta} n^{1-\sigma\delta/\varrho})$.

Proof. This follows from Theorem 3.5.1 by plugging in the values for λ and γ as specified, where we only need to check that the specified γ is in the admissible range required by Theorem 3.5.1. To this end, recall that $\gamma = n^{-a}$ with

$$a = \frac{q+1}{\beta(q+2)+\varrho(q+1)}.$$

By assumption we have $\sigma/\varrho \leq a$ and obviously also $a \leq 1/\varrho$. This implies $\gamma \in [n^{-1/\varrho}, m^{-1/\varrho}]$, as required by Theorem 3.5.1. □

As usual, we also show that the learning rates from the corollary above can also be achieved adaptively.

Theorem 3.5.3. *Let the assumptions of Theorem 3.5.1 be satisfied for $\varrho \geq 1$ and the number of cells specified as $m = \lceil n^{\sigma} \rceil$ for some $\sigma < 1$. Let A_n be a minimal $1/\log n$-net of $(0,1]$ and let B_n be a minimal $1/\log n$-net of $[\sigma+1, \sigma+d]$ with $\sigma + d \in B_n$ and set $\Gamma_n := \{n^{-a} : a \in A_n\}$ and $\Lambda_n := \{n^{-b} : b \in B_n\}$. Assume that the parameters from Theorem 3.5.1 satisfy*

$$\sigma < \min\left\{ \frac{\varrho(q+1)}{\beta(q+2)+\varrho(q+1)}, \frac{\varrho}{\delta} \right\}.$$

Then there exists a constant $C > 0$ only depending on $C_{*,**}, C_{\mu}$ and C_S such that for all $n > \exp \frac{2\varrho+\log 2}{1-\sigma}$ and $\tau \geq 1$ we have

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{\beta(q+1)}{\beta(q+2)+\varrho(q+1)}} \log^{d+1} n$$

with probability not less than $1 - \mathrm{e}^{-\tau} - n^{\sigma} \exp(-C_{\mu}^{-1} C_S^{-\delta} n^{1-\sigma\delta/\varrho})$ where the hyper-parameters $\boldsymbol{\lambda}_{D_2}, \boldsymbol{\gamma}_{D_2}$ are selected by the training validation procedure (3.6).

Proof. By [55, Theorem 7.2], an inequality for empirical risk minimization, we have

$$\begin{aligned}
&\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \\
\leq\, & 6 \min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in\Lambda_n\times\Gamma_n} \left(\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}}\right) \\
& + 4\left(\frac{48C_*^{\frac{q}{q+1}}(\tau+\log(1+|\Lambda_n^m\times\Gamma_n^m|))}{n-l}\right)^{\frac{q+1}{q+2}} \\
\leq\, & 6\left(\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*}) - \mathcal{R}^*_{L,\mathbf{P}}\right) \\
& + 4\left(\frac{192C_*^{\frac{q}{q+1}}(\tau+\log(1+|\Lambda_n^m\times\Gamma_n^m|))}{n}\right)^{\frac{q+1}{q+2}}
\end{aligned} \tag{3.10}$$

with probability not less than $1-\mathrm{e}^{-\tau}$, where we picked values $\boldsymbol{\gamma}^*\in\Gamma_n^m$ and $\boldsymbol{\lambda}^*\in\Lambda_n^m$ which we will specify in a moment. We again set $\lambda_1=\ldots=\lambda_m=:\lambda$ and $\gamma_1=\ldots=\gamma_m=:\gamma$. By Theorem 3.5.1 there exists a constant $C>0$ such that

$$\begin{aligned}
&\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma},\mathrm{FFT}(m)}) - \mathcal{R}^*_{L,\mathbf{P}} \\
\leq & C\left(m\lambda\gamma^{-d}+\gamma^\beta+\lambda^{-1/\log n}\left(\frac{\gamma^{-\varrho}}{l}\right)^{\frac{q+1}{q+2}}\log^{d+1}n+\left(\frac{\tau}{l}\right)^{\frac{q+1}{q+2}}+\frac{\tau}{l}\right)
\end{aligned}$$

with probability not less than $1-3\mathrm{e}^{-\tau}-n^\sigma\exp(-C_\mu^{-1}C_S^{-\delta}n^{1-\sigma\delta/\varrho})$ for all $\lambda\in\lambda_n$ and $\gamma\in\Gamma_n\cap[l^{-1/\varrho},m^{-1/\varrho}]$. Note that

$$\left[l^{-1/\varrho},m^{-1/\varrho}\right]\supset\left[\left(\frac{2}{n}\right)^{1/\varrho},n^{-\sigma/\varrho}\right]=\left[n^{-(1-\log 2/\log n)/\varrho},n^{-\sigma/\varrho}\right].$$

Since A_n is an $1/\log n$-net of $(0,1]$ we have for

$$\left(1-\frac{\log 2}{\log n}\right)\frac{1}{\varrho}-\frac{\sigma}{\varrho}>\frac{2}{\log n},$$

which is equivalent to $n>\exp((2\varrho+\log 2)/(1-\sigma))$, that $\Gamma_n\cap[l^{-1/\varrho},m^{-1/\varrho}]\neq\emptyset$. That is, we can choose $a_*\in A_n$ such that $\gamma=n^{-a_*}$ is in the admissible range and

$$\frac{q+1}{\beta(q+2)+\varrho(q+1)}-\frac{2}{\log n}\leq a_*\leq\frac{q+1}{\beta(q+2)+\varrho(q+1)}+\frac{2}{\log n}.$$

Choosing $\gamma=n^{-a_*}$ and $\lambda=n^{-\sigma-d}$, we can finish the proof exactly as the proof of Theorem 3.4.3 by combining the inequalities above. □

4 Experimental Results

In this chapter, whose contents will be published in [25], we complement our theoretical findings by experimentally verifying that, given some dataset D, global and local SVMs achieve the same generalization performance if this dataset is non-trivially embedded in a much higher dimensional space. To this end, we define an embedding $\Phi : \mathbb{R}^d \to \mathbb{R}^{d+p}$ as follows: Sample $w_1, \ldots, w_p$ iid from the uniform distribution on $[-\pi, \pi]^d$ and define the function $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ by $\varphi_j(x) = \sin\langle x, w_j\rangle$ for $j = 1, \ldots, p$. Subsequently, sample an orthogonal matrix $T \in \mathbb{R}^{(d+p)\times(d+p)}$ from the Haar-measure and set $\Phi(x) := T(x, \varphi(x))$. Now, given a dataset $D = ((x_1, y_1), \ldots, (x_n, y_n))$ with $x_i \in [-1, 1]^d$ for $i = 1, \ldots, n$ we define the embedded dataset $D_p = (\Phi(x_i), y_i)_{i=1,\ldots,n}$. The resulting dataset lies on the rotated graph of the map $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ and therefore naturally has a non-trivial d-dimensional structure in a $(d + p)$-dimensional Euclidean space, see Figure 4.1.

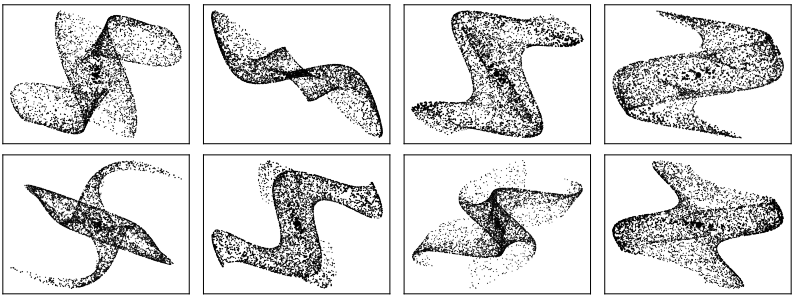

Figure 4.1: 5000 points sampled uniformly on $[-1, 1]^2$ and mapped into $\mathbb{R}^6$ with the procedure described. The figures show the orthogonal projection of the mapped datapoints onto randomly sampled two-dimensional planes.

We want to investigate experimentally how the generalization performance behaves as a function of artificially added dimensions p. Similar experiments have been conducted, for example, in [5] where the authors consider the synthetic three-dimensional dataset, where x_1 is sampled from a standard normal distribution

and $x_2 = x_1^3 + \sin(x_1) - 1$ and $x_3 = \log(x_1^2 + 1) - x_1$. The response variable is $y = \cos(x_1) + x_2 - x_3^3 + \varepsilon$, where ε is sampled from a normal distribution. They compare the performance of a local polynomial regressor as an estimator based on the whole feature vector (x_1, x_2, x_3) against an estimator having only access to the only true feature x_1. In [64] the authors consider datapoints lying on the two-dimensional *swiss roll* manifold in $\mathbb{R}^3$, which they map into $\mathbb{R}^{100}$ via a random 100×3-matrix and modeled the response variables as a function of the features plus noise. They only state, that their estimator "has a relatively fast convergence rate even though the dimension of the ambient space is large", but do not compare it to the performance of their estimator using the dataset in the original three-dimensional space, or a dataset in which the feature vectors contain only the two necessary parameters to describe the manifold. In [44] the authors conduct similar experiments using deep neural neetworks with ReLU activation function and the least-squares loss. They sample points from a uniform distribution on a d'-dimensional sphere in a d-dimensional space and modeled the response using a predefined function plus noise and examine the performance of a neural network for varying values of d' and d and different sample sizes. The hypothesis of low intrinsic dimensionality is especially prevalent for image datasets and convolutional neural network being able to exploit these structures. Although these highly specific datasets are not readily comparable to our setting, [45] consider a conceptually similar experimental setup where they keep the intrinsic dimension a dataset fixed and investigate the generalization performance for varying ambient dimensions.

For our purposes, we collected 32 regression datasets and 32 binary classification datasets from the UCI Repository [15] summarized in Tables 4.1 and 4.2. For the respective 16 smallest datasets we used a global kernel (i.e. no partition), for the remaining datasets we used a partition such that each cell contains at most 4000 samples. For each dataset we performed training runs with the embedding described above for $p = 0, \ldots, 50$, where 20% of each dataset was left out for testing and each run was repeated 50 times. For training and testing we used the command line version of liquidSVM [58], which implements a partitioning method. For hyperparameter selection we used 5-fold cross validation over a default 10×10-grid chosen by liquidSVM based on some characteristics of the dataset, which has been empirically verified to yield competitive performance. The results are summarized in Figures 4.2, 4.3, 4.4, and 4.5 for regression with global kernels, classification with global kernels, regression with local kernels, and classification

with local kernels respectively. We can divide the results in roughly three categories:

(i) In accordance with our theoretical findings, the generalization performance is independent of the number of artificially added dimensions. That is, the test error for the datasets $D_p, p = 1, \ldots, 50$, is similar as for the original dataset D. The datasets in this category constitute a clear majority.

(ii) After an initial increase of the test error, it quickly levels out at a moderately higher level, which is still well below the naive error. We still see this as a partial verification of our theoretical findings since at least after a certain point, the test error is independent of the further artificially added dimensions. Examples of datasets in this category are `bike_sharing_casual`, `bike_sharing_total`, `gas_sensor_drift_class`, `gas_sensor_drift_conc`, `sml2010_dining`, `sml2010_room`, `thyroid_ann`, and `travel_review_ratings`.

(iii) On a few rare exceptions, the test error grows significantly, as for `chess`, `crowd_sourced_mapping`, and `electrical_grid_stability_simulated`.

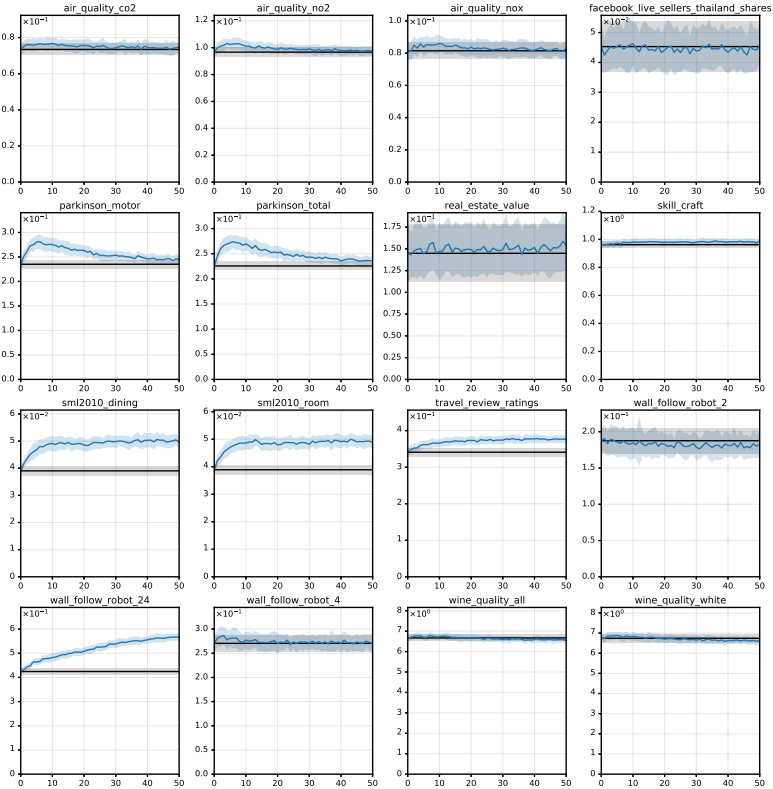

Figure 4.2: Test root mean-squared errors (y-axis) for global kernels. The x-axis contains the number of artificially added dimensions. The shaded blue area corresponds to the standard deviation across the different runs. For comparison, the horizontal black line shows the test error for the original dataset. The shaded grey area corresponds to the standard deviation across the different runs for the original dataset.

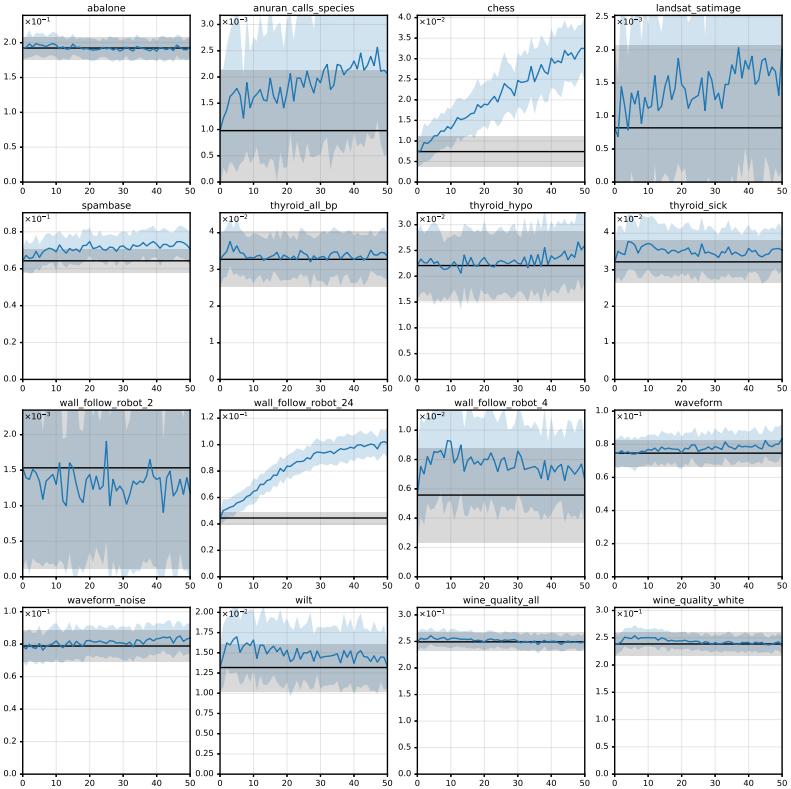

Figure 4.3: Test classification errors (y-axis) for global kernels. The x-axis contains the number of artificially added dimensions. The shaded blue area corresponds to the standard deviation across the different runs. For comparison, the horizontal black line shows the test error for the original dataset. The shaded grey area corresponds to the standard deviation across the different runs for the original dataset.

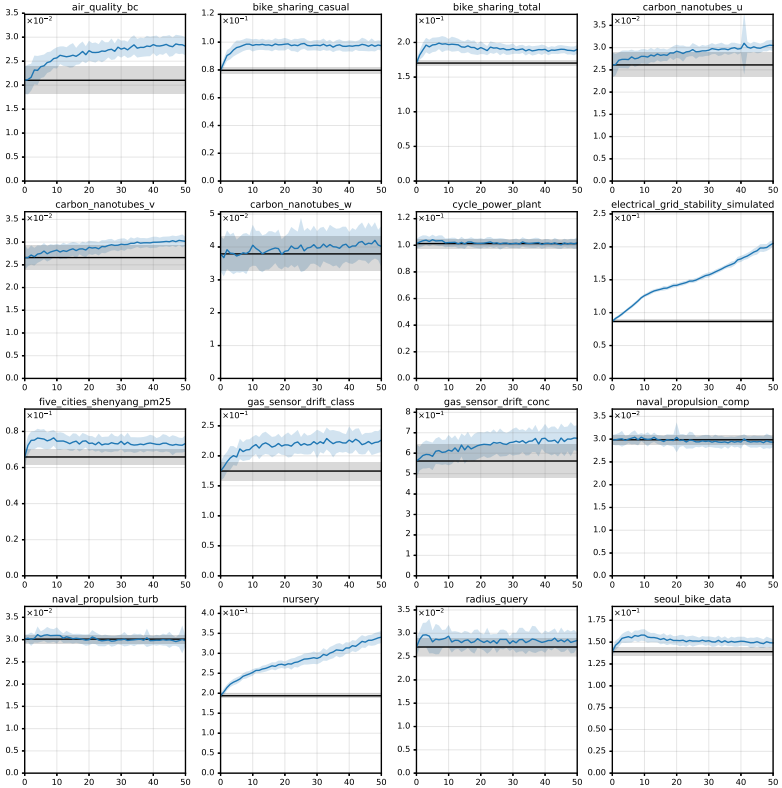

Figure 4.4: Test root mean-squared errors (y-axis) for local kernels. The x-axis contains the number of artificially added dimensions. The shaded blue area corresponds to the standard deviation across the different runs. For comparison, the horizontal black line shows the test error for the original dataset. The shaded grey area corresponds to the standard deviation across the different runs for the original dataset.

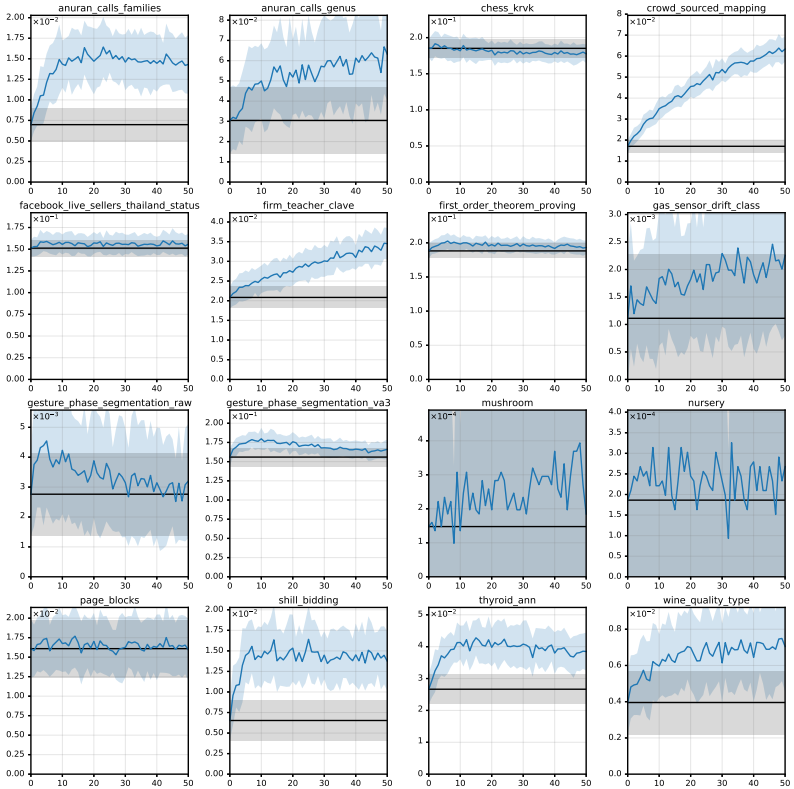

Figure 4.5: Test classification errors (y-axis) for local kernels. The x-axis contains the number of artificially added dimensions. The shaded blue area corresponds to the standard deviation across the different runs. For comparison, the horizontal black line shows the test error for the original dataset. The shaded grey area corresponds to the standard deviation across the different runs for the original dataset.

Name	Samples	Dimension	Naive Error	Base error
air_quality_bc	8991	10	0.2343	0.0212
air_quality_co2	7674	10	0.2463	0.0751
air_quality_no2	7715	10	0.2862	0.0976
air_quality_nox	7718	10	0.2884	0.0806
bike_sharing_casual	17379	12	0.2687	0.0801
bike_sharing_total	17379	12	0.3717	0.1707
carbon_nanotubes_u	10721	5	0.6304	0.0268
carbon_nanotubes_v	10721	5	0.6311	0.0270
carbon_nanotubes_w	10721	5	0.5782	0.0382
cycle_power_plant	9568	4	0.4521	0.0992
electrical_grid_stability_simulated	10000	12	0.3883	0.0871
facebook_live_sellers_thailand_shares	7050	9	0.0769	0.0504
five_cities_shenyang_pm25	19038	14	0.1306	0.0653
gas_sensor_drift_class	13910	128	1.7285	0.1702
gas_sensor_drift_conc	13910	128	0.3432	0.0542
naval_propulsion_comp	11934	14	0.5888	0.0299
naval_propulsion_turb	11934	14	0.6000	0.0302
nursery	12960	8	1.2356	0.1923
parkinson_motor	5875	19	0.4716	0.2316
parkinson_total	5875	19	0.4459	0.2246
radius_query	10000	3	0.3755	0.0270
real_estate_value	414	6	0.2473	0.1521
seoul_bike_data	8760	14	0.3627	0.1414
skill_craft	3338	18	1.4480	0.9727
sml2010_dining	4137	17	0.3769	0.0386
sml2010_room	4137	17	0.3790	0.0393
travel_review_ratings	5456	23	0.6278	0.3442
wall_follow_robot_2	5456	2	1.0047	0.1882
wall_follow_robot_24	5456	24	1.0047	0.4310
wall_follow_robot_4	5456	4	1.0047	0.2747
wine_quality_all	6497	12	0.8732	0.6739
wine_quality_white	4898	11	0.8855	0.6802

Table 4.1: Regression datasets.

Name	Samples	Dimension	Naive Error	Base error
abalone	2870	8	0.4676	0.1868
anuran_calls_families	6585	22	0.3288	0.0075
anuran_calls_genus	5743	22	0.2774	0.0028
anuran_calls_species	4599	22	0.2437	0.0013
chess	3196	36	0.4778	0.0050
chess_krvk	8747	22	0.4795	0.1782
crowd_sourced_mapping	9003	28	0.1659	0.0195
facebook_live_sellers_thailand_status	6622	9	0.3525	0.1574
firm_teacher_clave	8606	16	0.4997	0.0215
first_order_theorem_proving	6118	51	0.4175	0.1904
gas_sensor_drift_class	5935	128	0.4930	0.0009
gesture_phase_segmentation_raw	5719	19	0.4842	0.0040
gesture_phase_segmentation_va3	5691	32	0.4816	0.1560
landsat_satimage	3041	36	0.4959	0.0010
mushroom	8124	111	0.4820	0.0000
nursery	8588	8	0.4967	0.0003
page_blocks	5242	10	0.0628	0.0155
shill_bidding	6321	9	0.1068	0.0069
spambase	4601	57	0.3940	0.0629
thyroid_all_bp	3621	31	0.0434	0.0352
thyroid_ann	7034	21	0.0523	0.0271
thyroid_hypo	2700	25	0.0504	0.0220
thyroid_sick	3621	31	0.0621	0.0314
wall_follow_robot_2	4302	2	0.4874	0.0010
wall_follow_robot_24	4302	24	0.4874	0.0443
wall_follow_robot_4	4302	4	0.4874	0.0043
waveform	3353	21	0.4942	0.0739
waveform_noise	3347	40	0.4945	0.0754
wilt	4839	5	0.0539	0.0150
wine_quality_all	4974	12	0.4298	0.2481
wine_quality_type	6497	11	0.2461	0.0048
wine_quality_white	3655	11	0.3986	0.2311

Table 4.2: Classification datasets.

5 Final Remarks

In this final chapter we give an outlook on possible future research and generalizations based on the results of this thesis. We further summarize existing results in the literature similar to ours for comparison. Contents of this chapter were published in [24].

5.1 Outlook

Optimality of Rates

We briefly want to discuss the optimality of the rates in Section 2.3 and 3.4. The classical result of Stone [60] considers the case where $\mathbf{P}_X$ is the uniform distribution on $[0,1]^d$ and states that $n^{-\frac{2\alpha}{2\alpha+d}}$ is the optimal rate of convergence. This statement can directly be generalized to the case where $\mathbf{P}_X$ is the uniform d'-dimensional distribution on the cube in the first d' axes of $\mathbb{R}^d$, where $d' \in \{1,\dots,d\}$. Note that in this case $\mathbf{P}$ satisfies Assumption 2.1.1 and 3.1.2 for all $\varrho \geq d'$. From this we can conclude that in the case $\varrho \in \{1,\dots,d\}$ the rates in the respective sections are optimal up to the logarithmic factor. In the general non-integer case $\varrho \geq 1$ we can still deduce a lower bound of order $2\alpha/(2\alpha + \lfloor \varrho \rfloor)$, however there is no immediate argument that this is the optimal lower bound. We strongly hypothesize that the general optimal lower bound is of order $2\alpha/(2\alpha + \varrho)$ but we will leave this as a conjecture for possible future research.

Other Loss Functions

As our general oracle inequalities in Sections 2.2 and 3.3 already indicate, our technique is flexible enough to handle additional learning scenarios using other loss functions than the least-squares loss or the hinge loss. For example, the conditional

τ-quantile function can be estimated using the pinball loss

$$L_{\text{pin}}(y,t) := \begin{cases} (1-\tau)(t-y), & \text{if } y < t, \\ \tau(y-t), & \text{if } t \geq y. \end{cases}$$

Using a variance bound, see Definition 1.3.1, and a calibration inequality for the pinball loss from [56] and imposing some standard regularity assumptions on $f^*_{L_{\text{pin}},\mathbf{P}}$ we can thus derive learning rates for the pinball loss, as well as bounds on the $L_q(\mathbf{P}_X)$ distance from $f_{D,\lambda,\gamma}$ to $f^*_{L_{\text{pin}},\mathbf{P}}$ under one of our set of intrinsic dimensionality assumptions from Chapter 2 or 3. This would generalize some of the results from [16] in the sense, that we can substitute d with ϱ in their learning rates. The same is true for the conditional expectile function, which is estimated by the asymmetric least-squares loss

$$L_{\text{ALS}}(y,t) := \begin{cases} (1-\tau)(t-y)^2, & \text{if } y < t, \\ \tau(y-t)^2, & \text{if } t \geq y, \end{cases}$$

where $\tau \in (0,1)$, since in [20] the necessary variance bound and a calibration inequality for L_{ALS} are derived.

Other Kernels

Another possibility to further generalize our results is to consider a larger class of kernels. For example Theorem 1.3.16, a main tool for bounding the statistical error, can be generalized with the same techniques to anisotropic Gaussian kernels, i.e. Gaussian kernels that have a different bandwidth parameter in each covariate. Anisotropic Gaussian SVMs are for example considered in [26] where the authors show that, compared to a regular Gaussian kernel, the anisotropic one has an improved performance for regression functions that have a varying degree of smoothness in each covariate. A generalization to other classes of kernels is however not obvious to us, as we make use of many properties that are exclusive to the Gaussian RKHS.

Boundedness Assumption in Regression

The boundedness assumption $Y \subset [-M, M]$ in Sections 2.3 and 3.4 can be relaxed to an exponential decay of the distribution of the noise variable $y - f^*_{L,\mathbf{P}}(x)$. Under this assumption one can show, that by choosing a sequence of logarithmically growing clipping values $M = M_n$, the resulting learning algorithms achieve the same rates as in the Sections 2.3, 2.4, 3.4, and 3.5. This generalization can be proven by showing that the clipping value M_n is correct with sufficiently high probability. As the details are a little bit technical and would distract from the actual conclusions of our results, we refer to [16, Theorem 3.6].

5.2 Review of Existing Results

In this section we summarize existing results on learning rates for various learning methods under the assumption that the data has a low-dimensional intrinsic structure. This summary highlights the contribution of the results of this thesis:

(i) The most common assumption to describe the intrinsic dimensionality of the data is to assume that the data generating distribution is supported on a smooth manifold. We considerably weaken this assumption by considering the fractal dimension of the support of the generating distribution instead.

(ii) Adaptivity to the intrinsic dimensionality of the data is rarely considered or only under additional assumptions. In contrast, our results show that a simple training validation approach achieves the same learning rates without knowledge on the intrinsic dimension of the data or the regularity of the target function.

(iii) So far there exist no results on learning rates that depend on the intrinsic dimensionality of the data for a speed-up procedure of a common learning method, such as our results of Chapter 3.

Least-Squares Regression

We first summarize results on least-squares regression.

SVMs. In [65] the authors consider the case, where the input space $X \subset \mathbb{R}^d$ is a compact, smooth manifold of dimension ϱ and that the Bayes decision function is α-Hölder continuous with respect to the geodesic distance on X for

some $\alpha \in (0, 1]$. They derive learning rates of the form $(\log^2(n)/n)^{\alpha/(8\alpha+4\varrho)}$. Under these assumptions on X, the assumption that the Bayes function is α-Hölder continuous with respect to the geodesic distance is equivalent to α-Hölder continuity with respect to the Euclidean distance, which is a consequence of [65, Lemma 1]. That is, Corollary 2.3.8 gives a significant improvement of the result in [65] under much less restrictive assumptions. Additionally, they do not consider adaptive parameter selection, i.e. the dimension ϱ of X and α need to be known.

Bayesian Regression with Gaussian processes. In [64] the authors consider the case of a compact ϱ-dimensional manifold $X \subset \mathbb{R}^d$, which is sufficiently regular and a C^α Bayes function, where $\alpha \leq 2$. For Bayesian regression using Gaussian processes with squared exponential covariance they derive the learning rate $n^{-2\alpha/(2\alpha+\varrho)} \log^{\varrho+1} n$, which is identical to ours, but under much more restrictive assumptions on both α and $\operatorname{supp} \mathbf{P}_X$. Additionally, they present a training validation scheme for choosing the hyperparameters of the prior distribution. Under some additional technical assumption, which is hard to verify, they prove that this method achieves the same rates adaptively. Also note that Bayesian GP regression is related to Gaussian least-squares SVMs in the sense that the posterior mean function of the Gaussian process coincides with the SVM solution provided the prior distribution is suitably chosen, see [29, Proposition 3.6]. The choice of the prior distribution in [64] however does not lead to the same decision function as the one chosen by a Gaussian least-squares SVM.

Neural Networks. In [44] the authors consider deep neural networks with ReLU activation. They show that if the number of layers L, the number of non-zero weights W, and the sup-norm B of the weights are chosen appropriately, for an α-Hölder regular target function a neural network achieves a learning rate of $n^{-\frac{2\alpha}{2\alpha+\varrho+\varepsilon}}(1+\log n)^2$ where ϱ is the upper box-counting dimension of $\operatorname{supp} \mathbf{P}_X$ and $\varepsilon > 0$ can be chosen arbitrarily small. However, the choice of L, W, and B requires knowledge on the unknown parameters α and ϱ. A further drawback in their result is that the estimator they consider is not computable in practice as they consider an exact minimizer of a non-convex optimization problem.

Local Polynomial Regression. In [5] the authors consider local linear regression

in the setting of a differentiable ϱ-dimensional manifold $X \subset \mathbb{R}^d$ and a twice differentiable Bayes function. They further assume, that $\mathbf{P}_X$ has a differentiable density w.r.t. local charts and prove the learning rate $n^{-4/(4+\varrho)}$. They also state, that the result can be extended if the Bayes function is α-times differentiable using polynomials of degree $\alpha-1$ to achieve the rate $n^{-2\alpha/(2\alpha+\varrho)}$. They propose a training validation scheme for bandwidth selection, but give no theoretical guarantees.

Tree-Based Regressor. In [32] the authors consider a tree-based locally constant regressor. Their notion of intrinsic dimension is based on the doubling dimension of the input space. The doubling dimension of a metric space is the smallest constant c, such that every ball of radius $r > 0$ in X can be covered by 2^c balls of radius $r/2$. Based on the doubling dimension c of the input space they prove that for a Lipschitz continuous target function their estimator achieves the learning rate $n^{-2/(2+k)}$ for a constant $k \in \mathcal{O}(c \log c)$. They achieve this rate adaptively using a training validation scheme, as well as a stopping criterion for building the tree. Although the doubling dimension has the same favorable properties as Assumption 2.1.1 in the sense that it does not require any differentiable structure of the input space, it has the great disadvantage that its value is usually much larger than the exponent ϱ in Assumption 2.1.1. To illustrate this, let $X \subset \mathbb{R}^d$ be a bounded set with doubling dimension c. For simplicity, let us assume that $X \subset B_{\ell_2^d}$, that is $\mathcal{N}_{\ell_2^d}(X, 1) = 1$. By assumption, X can be covered by 2^c balls of radius $1/2$ which in turn can each be covered by 2^c balls of radius $1/4$. Inductively this gives us $\mathcal{N}(X, 1/2^k) \leq 2^{ck}$ for all $k \in \mathbb{N}_0$. A simple argument then gives us $\mathcal{N}(X, \varepsilon) \leq 2^c \varepsilon^{-c}$ for all $\varepsilon \in (0, 1)$, and hence we have $c \geq \varrho$, whenever 2.1.1 is fulfilled exactly for ϱ. In fact, we often have $c > \varrho$. For example, as a consequence of [4, Satz 2] the optimal value for the doubling dimension of the unit disc $B_{\ell_2^2}$ is given by $\log_2 7 \approx 2.81$, while its box-counting dimension is 2. Doubling dimensions of sets, that actually have a standard notion of dimension, can rarely be computed explicitly. For example, a ϱ-dimensional manifold has doubling dimension $\mathcal{O}(\varrho)$, where the proportionality constant depends on the curvature of X, see e.g. [11, Theorem 22].

k-Nearest Neighbor. In [35] the authors show, that under Assumption 2.1.1 and for an α-Hölder continuous target function, where $0 < \alpha \leq 1$, the k-NN rule achieves a learning rate of $n^{-2\alpha/(2\alpha+\varrho)}$. They only achieve this rate with knowledge on both, ϱ and α.

Binary Classification

Finally, we summarize results on binary classification, although they are much scarcer than for least-squares regression.

SVMs. In [66] the authors assume that the input space $X \subset \mathbb{R}^d$ is a compact, connected, smooth ϱ-dimensional manifold without boundary and $\mathbf{P}_X$ is the normalized surface measure on X. They further assume, that $\operatorname{sgn}(2\eta - 1)$ is contained in the interpolation space $(L_1(X), W^{2,1}(X))_{\theta,\infty}$ for some $\theta \in (0, 1]$, where $W^{2,1}(X)$ is a Sobolev-space on the manifold X and derive the learning rate $(\log^2(n)/n)^{\theta/(6\theta+\varrho)}$ for Gaussian SVMs. Of course, such a regularity assumption for a discrete-valued function is very restrictive, especially for small values of ϱ. Exemplarily, for $\varrho = 1$ and $\theta \geq 1/2$ their assumptions actually imply the pathological case $\operatorname{sgn}(2\eta - 1) \equiv 1$ or $\operatorname{sgn}(2\eta - 1) \equiv -1$, since by the embedding theorem in [61, 7.4.2 (iv)] the space $(L_1(X), W^{2,1}(X))_{\theta,\infty}$ then consists of continuous functions. In addition, their fastest possible rate is given by $n^{-1/7}$, while we derive learning rates up to n^{-1} in Sections 2.4 and 3.5.

Dyadic Decision Trees. In [53] dyadic decision trees are considered. They assume, that for a partition $\mathcal{P}_m$ of the input space $X = [0, 1]^d$ into cubes of sidelength $1/m$, where m is a dyadic integer, every $A \in \mathcal{P}_m$ satisfies $\mathbf{P}_X(A) \leq c_0 m^{-\varrho}$. Additionally they assume, that the number of cubes in $\mathcal{P}_m$, that intersect the decision boundary $\{x \in X : \eta(x) = 1/2\}$ is bounded by $c_1 m^{\varrho-1}$, that is they impose an assumption similar to 2.1.1 on the decision boundary. They derive a learning rate of $(\log(n)/n)^{1/\varrho}$. Remarkably, the optimal choice of their only hyperparameter, the depth of the tree, does not depend on ϱ.

Bibliography

[1] L. Ambrosio, A. Colesanti, and E. Villa. Outer Minkowski content for some classes of closed sets. *Math. Ann.*, 342:727–748, 2008.

[2] J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35:608–633, 2007.

[3] J. Bergh and J. Löfström. *Interpolation Spaces. An introduction*, Springer-Verlag, Berlin-New York, 1976.

[4] K. Bezdek. Über einige Kreisüberdeckungen. *Beiträge Algebra Geom.*, 14: 7–14, 1983.

[5] P. J. Bickel and B. Li. Local polynomial regression on unknown manifolds. In *Complex datasets and inverse problems*, volume 54 of *IMS Lecture Notes Monogr. Ser.*, pages 177–186. Inst. Math. Statist., Beachwood, OH, 2007.

[6] I. Blaschzyk and I. Steinwart. Improved classification rates under refined margin conditions. *Electron. J. Stat.*, 12:793–823, 2018.

[7] I. Blaschzyk and I. Steinwart. Improved classification rates for localized SVMs. *arXiv*, 1905.01502, 2019.

[8] O. Bousquet. New approaches to statistical learning theory. *Ann. Inst. Statist. Math.*, 55:371–389, 2003.

[9] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*, Cambridge University Press, Cambridge, 1990.

[10] A. Cuevas. Set estimation: Another bridge between statistics and geometry. *Bol. Estad. Investig. Oper.*, 25:71–85, 2009.

[11] S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. In *STOC'08*, pages 537–546. ACM, New York, 2008.

[12] S. De Vito, M. Piga, L. Martinotto, and G. Di Francia. CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic Bayesian regularization. *Sensors and Actuators B: Chemical*, 143:182 – 191, 2009.

[13] L. Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE transactions on pattern analysis and machine intelligence*, 4:154–157, 1982.

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

[15] D. Dua and C. Graff. UCI Machine Learning Repository, 2017. URL `http://archive.ics.uci.edu/ml`.

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

[17] L. C. Evans. *Partial Differential Equations*, American Mathematical Society, Providence, RI, second edition, 2010.

[18] K. Falconer. *Fractal Geometry*, John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2003.

[19] J. D. Farmer, E. Ott, and J. A. Yorke. The dimension of chaotic attractors. *Phys. D*, 7:153–180, 1983.

[20] M. Farooq and I. Steinwart. Learning rates for kernel-based expectile regression. *Mach. Learn.*, 108:203–227, 2019.

[21] G. B. Folland. *Real Analysis*, John Wiley & Sons, Inc., New York, second edition, 1999.

[22] J. M. Fraser. *Assouad Dimension and Fractal Geometry*, Cambridge University Press, 2020.

[23] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York, 2002.

[24] T. Hamm and I. Steinwart. Adaptive learning rates for support vector machines working on data with low intrinsic dimension. *Ann. Statist.*, 49:3153–3180, 2021.

[25] T. Hamm and I. Steinwart. Intrinsic dimension adaptive partitioning for kernel methods. *SIAM J. Math. Data Sci.*, 2022 (accepted).

[26] H. Hang and I. Steinwart. Optimal learning with anisotropic Gaussian SVMs. *Appl. Comput. Harmon. Anal.*, 55:337–367, 2021.

[27] S. Har-Peled. *Geometric Approximation Algorithms*, American Mathematical Society, Providence, RI, 2011.

[28] J. Heinonen. *Lectures on Analysis on Metric Spaces*, Springer-Verlag, New York, 2001.

[29] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv*, 1805.08845v1, 2018.

[30] A. Klenke. *Probability Theory*, Springer, London, second edition, 2014.

[31] S. Kpotufe. k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems 24*, pages 729–737. Curran Associates, Inc., 2011.

[32] S. Kpotufe and S. Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *J. Comput. System Sci.*, 78:1496–1515, 2012.

[33] S. Kpotufe and V. Garg. Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems 26*, pages 3075–3083. Curran Associates, Inc., 2013.

[34] T. Kühn. Covering numbers of Gaussian reproducing kernel Hilbert spaces. *J. Complexity*, 27:489–499, 2011.

[35] S. R. Kulkarni and S. E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inform. Theory*, 41:1028–1039, 1995.

[36] J. Lehrbäck and H. Tuominen. A note on the dimensions of Assouad and Aikawa. *J. Math. Soc. Japan*, 65:343–356, 2013.

[37] E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17, pages 777–784. MIT Press, 2005.

[38] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27:1808–1829, 1999.

[39] M. J. McGuinness. The fractal dimension of the Lorenz attractor. *Phys. Lett. A*, 99:5–9, 1983.

[40] A. McRae, J. Romberg, and M. Davenport. Sample complexity and effective dimension for regression on manifolds. In *Advances in Neural Information Processing Systems*, volume 33, pages 12993–13004. Curran Associates, Inc., 2020.

[41] M. Meister and I. Steinwart. Optimal learning rates for localized SVMs. *J. Mach. Learn. Res.*, 17:Paper No. 194, 44, 2016.

[42] J. Milnor. On the concept of attractor. *Comm. Math. Phys.*, 99:177–195, 1985.

[43] N. Mücke. Reducing training time by efficient localized kernel regression. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2603–2610. PMLR, 2019.

[44] R. Nakada and M. Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.*, 21: Paper No. 174, 38, 2020.

[45] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

[46] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[47] M. Reed and B. Simon. *Methods of Modern Mathematical Physics II. Fourier Analysis, Self-Adjointness*, Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1975.

[48] A. Rudi and L. Rosasco. Generalization Properties of Learning with Random Features. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[49] A. Rudi, R. Camoriano, and L. Rosasco. Less is More: Nyström Computational Regularization. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[50] D. Ruelle. Strange attractors. *Math. Intelligencer*, 2:126–137, 1979/80.

[51] I. J. Schoenberg. Metric spaces and completely monotone functions. *Ann. of Math. (2)*, 39:811–841, 1938.

[52] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA, 2002.

[53] C. Scott and R. D. Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inform. Theory*, 52:1335–1353, 2006.

[54] E. M. Stein. *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, N.J., 1970.

[55] I. Steinwart and A. Christmann. *Support Vector Machines*, Springer, New York, 2008.

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

[57] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35:575–607, 2007.

[58] I. Steinwart and P. Thomann. liquidSVM: A fast and versatile SVM package. *arXiv*, 1702.06899, 2017.

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

[60] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10:1040–1053, 1982.

[61] H. Triebel. *Theory of Function Spaces II*, Birkhäuser Verlag, Basel, 1992.

[62] R. Urner and S. Ben-David. Probabilistic Lipschitzness: A niceness assumption for deterministic labels. *Learning Faster from Easy Data Workshop@NIPS*, 2013.

[63] C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001.

[64] Y. Yang and D. B. Dunson. Bayesian manifold regression. *Ann. Statist.*, 44: 876–905, 2016.

[65] G.-B. Ye and D.-X. Zhou. Learning and approximation by Gaussians on Riemannian manifolds. *Adv. Comput. Math.*, 29:291–310, 2008.

[66] G.-B. Ye and D.-X. Zhou. SVM learning and L^p approximation by Gaussians on Riemannian manifolds. *Anal. Appl. (Singap.)*, 7:309–339, 2009.

[67] Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16:3299–3340, 2015.