Learning with High-Dimensional Data

Simon Fischer

University of Stuttgart
Germany

Learning with High-Dimensional Data

Von der

Fakultät 8 • Mathematik und Physik

der

Universität Stuttgart

zur Erlangung der Würde eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigte Abhandlung

Vorgelegt von

Simon Fischer

aus Heilbronn

Hauptberichter: Prof. Dr. Ingo Steinwart
Mitberichter: Prof. Dr. Andreas Christmann
Prof. Dr. Tino Ullrich

Tag der mündlichen Prüfung

01. Februar 2022

Institut für Stochastik und Anwendungen

der

Universität Stuttgart

2022

Preface

Many modern real-world applications of machine learning algorithms can be modeled as high-dimensional statistical learning problems, such as image classification or medical diagnostics based on gene expressions have thousands of dimensions. It is well-known that such learning problems are challenging both practically as well as theoretically. The practical challenges include the memory and time consumption of the learning algorithms, however, we focus on the theoretical challenges. To this end, let us briefly sketch some well-known theoretical results for least-squares (LS) regression and their problems in high-dimensional scenarios. In [79] it is shown that the optimal decay rate of the L_2*-generalization error* for an increasing sample size n, the so-called *learning rate*, is

$$n^{-\frac{\alpha}{2\alpha+d}} \quad , \tag{0.1}$$

where $d \geq 1$ denotes the dimension of the input values and $\alpha > 0$ the smoothness of the *regression function.* Since then it was shown by many authors that this optimal rate (up to a logarithmic term) is achieved by various learning algorithms, see e.g. [79, 50, 78, 27, 34] and references therein. Learning rates with a similar dependence on the input dimension were obtained for other learning scenarios, see e.g. [3, 76] for binary classification. If an algorithm achieves a learning rate of this form then the common interpretation is as follows: Since the generalization error decreases polynomially in n, the learning problem is feasible and the algorithm solves it data efficient. But the polynomial order of the learning rate vanishes for increasing d and hence the algorithm suffers from the *curse of dimensionality.* However, this heuristic interpretation has its weaknesses:

(i) In (0.1) only the asymptotic behavior with respect to the sample size n is specified, but in most cases the error bound is actually of the form

$$C_{\alpha,d} \cdot n^{-\frac{\alpha}{2\alpha+d}} \quad , \tag{0.2}$$

with some constant $C_{\alpha,d}$ independent of n, but depending on the dimension d and the smoothness α. Unfortunately, most contributions providing learning rates do not investigate the dependence of $C_{\alpha,d}$ on d and α. If we, for example, assume the highly unlikely case that $C_{\alpha,d} = \exp(-d)$ then the error bound in (0.2) would have a very pleasing behavior for an increasing dimension d.

(ii) At first sight the regularity assumption expressed by α seems to be independent of d but this is not completely true. For $d = 1$ a smoothness of $\alpha \in \mathbb{N}$ implies the existence of α derivatives. For $d = 2$ the same smoothness already implies $(\alpha + 2)(\alpha + 1)/2 - 1$ different (partial) derivatives. Moreover, for general $d, \alpha \in \mathbb{N}$ the number of derivatives is expressed by the binomial coefficient $\binom{\alpha+d}{d} - 1$ and hence grows like d^α for $d \to \infty$. This shows that the smoothness assumption depends on the dimension. Moreover, if α expresses (only) the Sobolev smoothness of the regression function, as in [38, 27], the situation is even more obscure. To be more precise, in low-dimensional spaces $d < 2\alpha$ the Sobolev embedding theorem, see e.g. [1, Theorem 4.12], guaranties the continuity of the regression function whereas in high-dimensional spaces $d > 2\alpha$ the regression functions can be even discontinuous.

Both issues indicate that the learning rate alone is insufficient to decide whether a problem suffers from the curse of dimensionality or not. For these reasons, we use a different approach to model high-dimensional learning problems which is motivated by the following idea: Many practical applications are not only high-dimensional learning problems they are even intrinsically infinite-dimensional. For example, if we have gray scale images as input values x, it is natural to model them as infinite-dimensional objects, namely functions $x\colon [0, W] \times [0, H] \to [0, 1]$ over a rectangle, where $x(s, t)$ represents the brightness of the image at the position $(s, t) \in [0, W] \times [0, H]$. For the simple reason that we cannot store infinite-dimensional objects on a computer the *original* infinite-dimensional learning problem is practically not feasible. However, in the moment when we load the infinite-dimensional data on a computer each data point gets discretized to a high-dimensional but finite-dimensional object. In the case of images they get discretized, e.g. via camera or scanner, to raster images consisting of finitely many pixels and hence become finite-dimensional objects.

Inspired by this interpretation, we model a high-dimensional learning problem as an infinite-dimensional one in which our algorithm has only access to a projection of the data into a d-dimensional space. This allows us to apply assumptions on the infinite-dimensional learning problem and consider projections into a d-dimensional space for all $d \geq 1$. As a result, this approach fixes Issue (ii). In a way, this means that the dimension d is a hyper parameter of the learning algorithm. Lower values for d corresponds to less information demand per data point and hence lower values of d are preferable. In order to solve Issue (i), there is no way around tracking the dependence on the dimension d precisely in all bounds.

The goal of this thesis is to identify infinite-dimensional learning problems and corresponding projections onto $\mathbb{R}^d$ in which standard algorithms can learn with a polynomial rate in n even if $d = d_n$ increases with n. In doing so, we present a class of infinite-dimensional classification problems in which histograms can learn with a polynomial rate.

Abstract

In Part I, after a brief introduction to statistical learning theory in Chapter 1, we introduce in Chapter 2 a framework that allows us to investigate learning scenarios with restricted access to the data. In Chapter 3 we use this framework to model high-dimensional learning scenarios as an infinite-dimensional one in which the learning algorithm has only access to some finite-dimensional projections. Afterwards, in Chapter 4 we provide a prototypical example of such an infinite-dimensional classification problem. Furthermore, we investigate learning rates of histograms in this scenario. In Chapter 5 we generalize the prototypical example from the previous chapter to various particular classes of infinite-dimensional classification problems. These generalizations illustrate some peculiarities of high-dimensional learning problems. We close this part with Chapter 6, in which we sketch further research directions and briefly discuss other approaches in the existing literature.

In Part II we present some individual results that might by useful for the investigation of kernel-based learning methods in high- or infinite-dimensional learning scenarios. To be more precise, after a brief introduction to Gaussian kernels in Section 7 we present log-covering number bounds for Gaussian *reproducing kernel Hilbert spaces (RKHSs)* on general bounded subsets of the Euclidean space $\mathbb{R}^d$ in Section 8. These bounds explicitly provide the dependence on crucial parameters such as the dimension d and improve already known bounds. In Section 9 we generalize the log-covering number bounds from the previous section to Gaussian kernels defined on

special infinite-dimensional compact subsets of the sequence space ℓ_2. More precisely, the considered domains are given by the image of the unit ℓ_∞-ball under some diagonal operator.

The investigations of the compactness properties of diagonal operators from ℓ_p to ℓ_q in Part III were initially thought as a preparation for Section 9. However, it turned out that we do not need them for Section 9, but we include these results anyway as they contribute to the still incomplete picture of the compactness properties of diagonal operators.

In the appendix, we provide proofs and introductions to topics, which are not directly related to *learning with high-dimensional data*, but may support the understanding of this thesis.

Zusammenfassung

Den Teil I dieser Arbeit beginnen wir in Kapitel 1 mit einer kurzen Einführung zur statistischen Lerntheorie. In Kapitel 2 stellen wir dann ein Framework vor, das es uns ermöglicht, Lernszenarien mit eingeschränktem Zugang zu den Daten zu betrachten. Dieses Framework verwenden wir in Kapitel 3, um hochdimensionale Lernszenarien als unendlich dimensionale Szenarien zu modellieren, bei denen der Lernalgorithmus nur Zugang zu endlich dimensionalen Projektionen hat. Anschließend stellen wir in Kapitel 4 ein Modellbeispiel für ein solches unendlich dimensionales Klassifikationsproblem vor. Ferner untersuchen wir das Histogrammverfahren auf Lernraten in diesem Szenario. In Kapitel 5 verallgemeinern wir das Modellbeispiel aus dem vorigen Kapitel auf mehrere spezielle Klassen unendlich dimensionaler Klassifikationsprobleme. Diese Verallgemeinerungen veranschaulichen einige Eigenheiten von hochdimensionalen Lernproblemen. Zum Abschluss des ersten Teils skizzieren wir in Kapitel 6 weiterführende Forschungsrichtungen und diskutieren alternative Ansätze aus der Literatur.

In Teil II stellen wir Ergebnisse vor, die zur Untersuchung von kernbasierten Lernmethoden in hoch- oder unendlich dimensionalen Lernszenarien nützlich sein könnten. Genauer präsentieren wir, nach einer kurzen Einführung in Kapitel 7 zu Gaußkernen, in Kapitel 8 Schranken an die log-Überdeckungszahlen für Gauß'sche *reproduzierende Kern-Hilberträume (RKHS)* auf allgemeinen beschränkten Teilmengen des euklidischen Raums $\mathbb{R}^d$. Dabei geben wir die Abhängigkeit von wichtigen Parametern, wie der Dimension d, explizit an. Ferner verbessern diese Schranken bereits bekannte Resultate aus der Literatur. In Kapitel 9 verallgemeinern wir diese Schranken auf Gaußkerne, die auf speziellen unendlich dimensionalen kompakten Teilmengen des Folgenraums ℓ_2 definiert sind. Genauer gesagt, betrachten wir Gebiete, die als das Bild der ℓ_∞-Einheitskugel unter einem Diagonaloperator gegeben sind.

Die Untersuchungen der Kompaktheitseigenschaften von Diagonaloperatoren von ℓ_p nach ℓ_q in Teil III waren ursprünglich als Vorbereitung für Kapitel 9 gedacht. Es hat sich jedoch herausgestellt, dass wir sie für Kapitel 9 nicht benötigen. Dennoch haben wir diese Ergebnisse in diese Arbeit aufgenommen, da sie einen Beitrag zu dem noch immer unvollständigen Bild der Kompaktheitseigenschaften von Diagonaloperatoren leisten.

Der Anhang enthält Beweise und Einführungen zu Themen, die zwar nicht im direkten Zusammenhang zum Thema *Lernen mit hochdimensionalen Daten* stehen, jedoch grundlegend für das Verständnis dieser Arbeit sind.

Veröffentlichungen

Die folgenden Publikationen sind im Rahmen dieses Promotionsvorhabens entstanden. Wobei die Ergebnisse aus [38] bis in die Masterarbeit zurückreichen und daher nicht Teil dieses Prüfungsverfahrens sind.

[37] S. Fischer. Some new bounds on the entropy numbers of diagonal operators. *J. Approx. Theory*, 251:105343, 2020.

[38] S. Fischer und I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *J. Mach. Learn. Res.*, 21:1–38, 2020.

[77] I. Steinwart und S. Fischer. A closer look at covering number bounds for Gaussian kernels. *J. Complexity*, 62:101513, 2021.

Danksagungen

In besonderem Maße danke ich meinem Betreuer und Prüfer Herrn Prof. Dr. Ingo Steinwart für die Vergabe dieses Themas sowie die mathematische und persönliche Betreuung. Ebenso gilt mein Dank allen weiteren Mitgliedern des Prüfungsausschusses für ihr Mitwirken sowie der *International Max Planck Research School for Intelligent Systems (IMPRS-IS)* für ihre Unterstützung.

Außerdem danke ich allen Korrekturlesern für ihren Einsatz und allen, die an mich geglaubt haben oder in irgendeiner anderen Form diese Arbeit unterstützt haben.

Contents

Lists and Notation

In the following we list all the used abbreviations, the included figures and tables as well as the most important notation.

List of Abbreviations

ALP	at least polynomial decay
AMP	at most polynomial decay
ERM	empirical risk minimizer
EXP	exponential decay
LS	least-squares
ONB	orthonormal basis
ONS	orthonormal system
RKHS	reproducing kernel Hilbert space

List of Figures and Tables

All synthetic data shown in the following figures were generated using [71].

List of Notation

In the following we summarize the notation used throughout this thesis. We start with some general notation sorted by the mathematical field and end with some specific notation sorted by the part where it gets used.

Sets and Functions. For sets M and N we use the standard symbols: $\in$ member, $\subseteq$ subset, $\cup$ union, $\cap$ intersection, $\setminus$ difference, and $\emptyset$ empty set. Moreover, we use the notation:

M^c	$:= N \setminus M$ complement of M (in N) if $M \subseteq N$.
$M \uplus N$	union of M and N, where M and N are disjoint, i.e. $M \cap N = \emptyset$.
$M \triangle N$	$:= M \setminus N \cup N \setminus M$ symmetric difference of M and N.
$\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$	set of positive integers, non-negative integers, and real numbers.
$[a,b]$, (a,b)	$\subseteq \mathbb{R}$ closed and open interval for $-\infty \leq a \leq b \leq \infty$.
$[d]$	$:= \{1, 2, \ldots, d\}$ set of the first d positive integers.
$\mathcal{F}(M)$	$:= \{N \subseteq M : \ \lvert N \rvert < \infty\}$ set of all finite subsets.
id, id_M	$: M \to M$ identity function.
sgn	$: \mathbb{R} \to \{\pm 1\}$ sign function with $\operatorname{sgn}(0) = 1$.
$\mathbb{1}_M$	indicator function of a subset M.
t_+	$:= \max\{0, t\}$ for $t \in \mathbb{R}$.
log	natural logarithm.
$\binom{t}{k}$	(generalized) binomial coefficient for $t > 0$ and $k \in \mathbb{N}$.
$\Gamma(t)$	$:= \int_0^\infty x^{t-1} e^{-x} \,\mathrm{d}x$ Gamma function for $t > 0$.
W_{-1}, W_0	Lambert's W-functions.

Asymptotic Equivalence. For real-valued functions f, g, which are defined in some neighborhood of a point a, we use the following asymptotic expressions for $x \to a$:

$f \preccurlyeq g$	i.e. there is a constant $c > 0$ and a neighborhood U of a with $f(x) \leq c g(x)$ for all $x \in U$.
$f \asymp g$	(weak) asymptotic equivalence, i.e. $f \preccurlyeq g$ and $f \succcurlyeq g$.

$f \sim g$ strong asymptotic equivalence, i.e. $f \cdot g > 0$ is positive in some neighborhood of a and $f(x)/g(x) \to 1$ for $x \to a$.

$o(g)$ small o Landau symbol, i.e. $o(g)$ denotes a function f with $f(x)/g(x) \to 0$ for $x \to a$.

We use the same notation for sequences and $a = \infty$.

σ-Algebras and Measures. For measurable spaces $(X, \mathcal{B})$, $(\bar{X}, \bar{\mathcal{B}})$ and some measures ν and μ on $\mathcal{B}$ we use the notation:

$\mathcal{B}|_A$ trace σ-algebra on $A \subseteq X$.

$\mathcal{B}(X, \tau)$ Borel σ-algebra of X if (X, τ) is a topological space. If there is no risk of confusion, we write $\mathcal{B}(X)$.

$\sigma(f)$ initial σ-algebra on X of a function $f\colon X \to \bar{X}$.

$\mathcal{B} \otimes \bar{\mathcal{B}}$ product σ-algebra on $X \times \bar{X}$.

$\nu \otimes \bar{\nu}$ product measure on $\mathcal{B} \otimes \bar{\mathcal{B}}$ if $\bar{\nu}$ is a measure on $\bar{\mathcal{B}}$.

$\nu \circ f^{-1}$ push-forward measure of ν with a measurable function $f\colon X \to \bar{X}$.

$|\nu|$ total variation of ν if ν is a signed measure.

$\operatorname{supp} \nu$ support of ν if X is a Hausdorff space and ν is a Radon measure.

$\nu \perp \mu$ ν and μ are singular.

$\nu \ll \mu$ ν is absolute continuous with respect to μ.

$\mathrm{d}\nu/\mathrm{d}\mu$ μ-density of ν if $\nu \ll \mu$.

λ^d, λ d-dimensional and one-dimensional Lebesgue measure.

δ_x Dirac measure at the point $x \in X$.

$\operatorname{unif}(M)$ uniform distribution over a set M.

$\mathcal{N}(\mu, \Sigma)$ normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$.

For a *probability measure* ν we additionally use the notation:

$Z \sim \nu$ Z is a X-valued random variable with distribution ν.

$\mathbb{E}_\nu(f)$ $= \mathbb{E}_{x \sim \nu} f(x)$ expectation of a function $f\colon X \to \mathbb{R}$.

$\mathbb{E}_\nu(f|\mathcal{A})$ conditional expectation of a ν-integrable function $f\colon X \to \mathbb{R}$ and a sub-σ-algebra $\mathcal{A} \subseteq \mathcal{B}$.

Metric Spaces. For a subset $M \subseteq X$ of a pseudo-metric or quasi-metric space (X, d) we use the following notation:

κ_X	≥ 1 quasi-triangle constant.
$\overline{M}$, $\mathring{M}$	closure and interior of M.
$B_X(x, r)$	closed ball with center $x \in X$ and radius $r \geq 0$.
$\mathring{B}_X(x, r)$	open ball with center $x \in X$ and radius $r \geq 0$.
$\operatorname{dist}(x, M)$	$:= \inf_{x' \in M} d(x, x')$ distance between $x \in X$ and M.
$\operatorname{diam}(M)$	$:= \sup_{x, x' \in M} d(x, x')$ diameter of M.
$\mathcal{N}(M, r)$	covering number of M for $r > 0$.
$\mathcal{P}(M, r)$	packing number of M for $r > 0$.
$\varepsilon_n(M)$	entropy number of M for $n \geq 1$.
$\mathcal{H}(M, r)$	$:= \log \mathcal{N}(M, r)$ log-covering number of M for $r > 0$.
$e_n(M)$	$:= \varepsilon_{2^{n-1}}(M)$ (dyadic) entropy number of M for $n \geq 1$.

For a bounded linear operator $R \colon U \to V$ between (quasi-)Banach spaces the quantities $\mathcal{N}(R, r)$, $\mathcal{P}(R, r)$, $\varepsilon_n(R)$, $\mathcal{H}(R, r)$, and $e_n(R)$ are analogously defined with $M = RB_U$.

Vector Spaces and Linear Operators. For a linear operator $R \colon U \to V$ between two real vector spaces U and V we use the following notation:

Id, Id_U	identity operator on some vector space and on U.
$\operatorname{span} M$	$:= \{\sum_{i=1}^n c_i x_i : \ n \in \mathbb{N},\ c_i \in \mathbb{R},\ x_i \in M\}$ linear span of a subset $M \subseteq U$.
$\operatorname{ran} R$	$:= RU$ range of R.
$\operatorname{rank} R$	rank of R, i.e. dimension of $\operatorname{ran} R$.
$\\|\cdot\\|$, $\\|\cdot\\|_U$	(quasi-)norm on some vector space and on U.
B_U, $\mathring{B}_U$	closed and open unit ball if U is a (quasi-)normed vector space.
$H_1 \hat{\otimes}_{\mathrm{hs}} H_2$	tensor product of Hilbert spaces, i.e. Hilbert-Schmidt operators $H_1 \to H_2$.

Function and Sequence Spaces. For a $(X, \mathcal{B}, \nu)$ measure space and a measurable function $f\colon X \to \mathbb{R}$ we denote the corresponding ν-equivalence class by $[f]_\nu$. Moreover, let (Y, τ) be a topological space. Then we define the following function spaces:

$\mathcal{L}_0(X, \mathcal{B})$ set of measurable functions $f\colon X \to \mathbb{R}$. If there is no risk of confusion, we write $\mathcal{L}_0(X)$.

$\mathcal{L}_p(\nu)$ set of functions $f \in \mathcal{L}_0(X, \mathcal{B})$ with finite (quasi-)norm

$$\|f\|_{\mathcal{L}_p(\nu)} := \begin{cases} \left(\int_X |f|^p \,\mathrm{d}\nu\right)^{1/p}, & 0 < p < \infty \\ \inf\{a > 0 :\ \nu(|f| > a) = 0\}, & p = \infty \end{cases}$$

$L_p(\nu)$ $:= \{[f]_\nu :\ f \in \mathcal{L}_p(\nu)\}$ set of ν-equivalence classes of function equipped with the (quasi-)norm $\|[f]_\nu\|_{L_p(\nu)} := \|f\|_{\mathcal{L}_p(\nu)}$.

$\ell_p(I)$ L_p space on a set I equipped with the counting measure.

ℓ_p $:= \ell_p(\mathbb{N})$ sequence space of p-summable sequences.

ℓ_p^d $:= \ell_p([d])$ d-dimensional space equipped with the p-(quasi-)norm.

$C_b(Y)$ set of bounded continuous functions $f\colon Y \to \mathbb{R}$ equipped with the uniform norm $\|f\|_{C_b(Y)} := \sup_{y \in Y} |f(y)|$.

$C_c(Y)$ $\subseteq C_b(Y)$ set of continuous functions with compact support.

Part I: Learning Scenario. In Part I we deal with learning scenarios which we usually describe as follows:

$(X, \mathcal{B})$ measurable space used as input space.

$Y \subseteq \mathbb{R}$ closed subset used as output space.

π_X, π_Y projections from $X \times Y$ onto X and Y.

P probability measure on $X \times Y$.

ν $:= P \circ \pi_X^{-1}$ marginal distribution of P on X.

D $\in (X \times Y)^n$ data set as well as the corresponding empirical measure.

δ $:= D \circ \pi_X^{-1}$ marginal distribution of D on X.

L $: Y \times \mathbb{R} \to [0, \infty)$ (supervised) loss function.

$\mathcal{R}_{L,P}(f)$ $:= \mathbb{E}_{(x,y)\sim P} L(y, f(x))$ L-risk of a function f.

$\mathcal{R}^*_{L,P,H}$ $:= \inf_{f \in H} \mathcal{R}_{L,P}(f)$ minimal L-risk over a hypothesis class $H \subseteq \mathcal{L}_0(X)$.

$\mathcal{R}^*_{L,P}$ $:= \mathcal{R}^*_{L,P,\mathcal{L}_0(X)}$ Bayes L-risk.

$f^*_{L,P}$ $\in \operatorname{argmin}_{f \in \mathcal{L}_0(X)} \mathcal{R}_{L,P}(f)$ Bayes decision function.

For a *(binary) classification problem* $Y = \{\pm 1\}$ we additionally use the following notation:

p_+, p_- probability for observing a positive and negative label.

ν_+, ν_- marginal distributions of the positive and negative labeled data points.

$\eta(x)$ $:= p_+ \mathrm{d}\nu_+ / \mathrm{d}\nu(x)$ probability for observing a positive label at $x \in X$.

If there is a *pseudo-metric* d on X then we use the following notation:

Δ_d distance to the decision boundary w.r.t. P and d.

M_d, MN_d margin and margin-noise function w.r.t. P and Δ_d.

For a (measurable and countable) partition $\mathcal{A} = (A_k)_{k \in K}$ of X we consider *histograms* and therefore we use the following notation:

$\operatorname{diam}(\mathcal{A})$ $:= \sup_{k \in K} \operatorname{diam}(A_k)$ diameter of the partition.

$h_{P,\mathcal{A}}$, $h_{D,\mathcal{A}}$ population and empirical histogram.

$H(\mathcal{A}, Y)$ $:= \left\{ \sum_{k \in K} c_k \mathbb{1}_{A_k} \ : \ c_k \in Y \right\}$ set of Y-valued functions which are constant on cells of $\mathcal{A}$.

$\mathcal{A}_\nu$ $:= \left\{ k \in K : \ \nu(A_k) > 0 \right\}$ indexes of relevant cells.

$\mathcal{A}_M$ $:= \left\{ k \in K : \ M \cap A_k \neq \emptyset \right\}$ indexes of relevant cells for a subset $M \subseteq X$.

If $X = \prod_{i \geq 1} X_i$ is a *sequence space*, i.e. (X_i, d_i) are Polish spaces, and $\mathcal{A} = (\mathcal{A}_i)_{i \geq 1}$ a partition of $(X_i)_{i \geq 1}$ then we use the following notation, for $I \in \mathcal{F}(\mathbb{N})$:

X_I $:= \prod_{i \in I} X_i$.

d_I $:= \max_{i \in I} d_i$

π_I $: X \to X_I$ projection onto the features in I.

P_I $:= P \circ (\pi_I, \mathrm{id}_Y)^{-1}$ transformed distribution of P on $X_I \times Y$.

ν_I $:= \nu \circ \pi_I^{-1}$ marginal distribution of P_I on X_I.

$\mathcal{A}_I$ product partition of X_I.

$\mathcal{A}_{I,\nu}$ $:= (\mathcal{A}_I)_{\nu_I}$ indexes of the relevant cells of the product partition.

Δ_I distance to the decision boundary on X w.r.t. P and the pull-back pseudo-metric of d_I.

M_I, MN_I margin and margin-noise function w.r.t. P and Δ_I.

$h_{P,\mathcal{A},I}$, $h_{D,\mathcal{A},I}$ population and empirical histogram using the features specified in I.

$h_{P,r,I}$, $h_{D,r,I}$ population and empirical histogram using the features specified in I and a predefined cubic partition $\mathcal{A}$ with radius r if $X_i = \mathbb{R}^{p_i}$ for all $i \geq 1$.

Part II: Gaussian Kernels. For an index set I we use in Part II the following notation:

k_σ, $k_{\boldsymbol{\sigma}}$ isotropic Gaussian kernel with width $\sigma > 0$ and anisotropic Gaussian kernel with width vector $\boldsymbol{\sigma} \in \ell_\infty(I)$.

$H_\sigma(X)$ Gaussian RKHS on $X \subseteq \ell_2(I)$.

I_σ, $I_\sigma[X]$ $: H_\sigma(X) \to \ell_\infty(X)$ ℓ_∞-embedding of the Gaussian RKHS on $X \subseteq \ell_2(I)$ and $I_\sigma := I_\sigma[B_{\ell_2^d}]$.

$P(U)$ $: H_\sigma(X) \to H_\sigma(X)$ orthogonal projection onto a subspace $U \subseteq H_\sigma(X)$.

$H_1 \otimes H_2$ tensor product of RKHSs H_1 and H_2.

Part III: Diagonal Operators. For $0 < p, q \leq \infty$, $k \geq 1$ and a sequence $\sigma = (\sigma_n)_{n\geq 1}$ we use in Part III the following notation:

D_σ $: \ell_p \to \ell_q,\ (x_n)_{n\geq 1} \mapsto (\sigma_n x_n)_{n \geq 1}$ diagonal operator.

$D_{p,q}^k$ $: \ell_p^k \to \ell_q^k,\ (x_n)_{n=1}^k \mapsto (\sigma_1 x_1, \ldots, \sigma_k x_k)$ k-dimensional part of D_σ.

$\mathrm{Id}_{p,q}^k$ $: \ell_p^k \to \ell_q^k$ identity operator.

P_p^k $: \ell_p \to \ell_p^k,\ (x_n)_{n\geq 1} \mapsto (x_1, \ldots, x_k)$ projection onto the first k coordinates.

I_p^k $: \ell_p^k \to \ell_p,\ (x_n)_{n=1}^k \mapsto (x_1, \ldots, x_k, 0, 0, \ldots)$ embedding.

Part I

Learning with High-Dimensional Data

This is the main part of the thesis. In this part we develop a framework which allows us to model a high-dimensional learning problem as a subset or projection of an infinite-dimensional learning problem. Finally, we use this framework to prove polynomial learning rates for histograms on various particular classes of infinite-dimensional (binary) classification problems.

Chapter 1

Introduction to Learning Theory

In this chapter we give a brief introduction to statistical learning theory and introduce the basic notation. Most of the introductory material and the notation is taken from [76], see also [25, 43] for additional information on learning theory.

1.1 Definitions and Basic Properties

Our *learning scenario* or *learning problem* of interest is described by the measurable space $(X, \mathcal{B})$ used as *input space*, the closed subset $Y \subseteq \mathbb{R}$ used as *output space*, and the *unknown* probability distribution P on $X \times Y$. Moreover,

$$D = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \sim P^n$$

denotes a data set of length $n \geq 1$ independently sampled according to P. For the marginal distribution of P on X we write $\nu \coloneqq P \circ \pi_X^{-1}$, where $\pi_X \colon X \times Y \to X$ is the projection onto X. If there is no risk of confusion, we write $D \coloneqq \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$ for the *empirical measure* corresponding to the data set D, where $\delta_{(x_i, y_i)}$ denotes the Dirac measure at (x_i, y_i). The goal is to find a (measurable) function $f \colon X \to \mathbb{R}$ such that $f(x)$ is a *good* prediction of y for an unseen pair $(x, y) \sim P$. In this context $f \colon X \to \mathbb{R}$ is called *decision function*.

In order to measure the quality of a decision function we use loss functions and risks. A *loss (function)* is a measurable function $L \colon X \times Y \times \mathbb{R} \to [0, \infty)$

and the corresponding *(L-)risk* is given by

$$\mathcal{R}_{L,P}(f) \coloneqq \int_{X \times Y} L\big(x, y, f(x)\big) \, \mathrm{d}P(x,y)$$

for all measurable functions $f\colon X \to \mathbb{R}$. The minimal L-risk over the set $\mathcal{L}_0(X)$ of measurable functions $f\colon X \to \mathbb{R}$ is called *Bayes (L-)risk* and denoted by

$$\mathcal{R}^*_{L,P} \coloneqq \inf_{f \in \mathcal{L}_0(X)} \mathcal{R}_{L,P}(f) \ .$$

In this sense, the Bayes risk is a lower bound on the quality of any decision function. The minimal L-risk on a subset $H \subseteq \mathcal{L}_0(X)$ is denoted by $\mathcal{R}^*_{L,P,H} \coloneqq \inf_{f \in H} \mathcal{R}_{L,P}(f)$. In this context H is called *hypothesis class.* The discrepancy between the risk of a decision function $f\colon X \to \mathbb{R}$ and the Bayes risk

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}$$

is called *excess (L-)risk.* Furthermore, all decision functions that achieve the Bayes L-risk

$$f^*_{L,P} \in \operatorname*{argmin}_{f \in \mathcal{L}_0(X)} \mathcal{R}_{L,P}(f)$$

are called *(L) Bayes functions.*

We call a loss function L *supervised* if it does not depend on $x \in X$, i.e. $L\colon Y \times \mathbb{R} \to [0, \infty)$. In this thesis we are mainly interested in the following supervised loss functions:

(i) the *LS loss* $L(y,t) \coloneqq (y-t)^2$ used for regression problems and

(ii) the *classification loss* $L(y,t) \coloneqq \mathbb{1}_{(-\infty,0]}\big(y \operatorname{sgn}(t)\big)$ used for classification problems. Here sgn denotes the sign function with the convention $\operatorname{sgn}(0) \coloneqq 1$.

In the case of the LS loss we write $\mathcal{R}_{\mathrm{LS},P}(f)$, $\mathcal{R}^*_{\mathrm{LS},P}$, and $f^*_{\mathrm{LS},P}$ for the LS-risk, the Bayes LS-risk, and a LS Bayes function, respectively. Analogously, we adapt the notation for the classification loss. For further popular loss functions see e.g. [76, Chapter 2].

A *learning method* $\mathcal{L} = (\mathcal{L}_n)_{n \geq 1}$ on $X \times Y$ is a sequence of mappings

$\mathcal{L}_n \colon (X \times Y)^n \to \mathcal{L}_0(X)$ such that the map

$$\begin{aligned} (X \times Y)^n \times X &\to \mathbb{R} \\ (D, x) &\mapsto f_D(x) \coloneqq \mathcal{L}_n(D)(x) \end{aligned} \tag{1.1}$$

is measurable with respect to the product σ-algebra. Note that in the literature the measurability is sometimes defined with respect to the universal completion of product σ-algebra, see e.g. [76, Definition 6.2]. Since we do not need such a generality, we stick to the simpler definition for convenience. If there is no risk of confusion, we just write $D \mapsto f_D$ for a learning method. In other words, a learning method produces for any data set D a decision function $f_D \colon X \to \mathbb{R}$. As a result, the risk $D \mapsto \mathcal{R}_{L,P}(f_D)$, describing the quality of the learning method, is no longer a real number, but a real-valued random variable on $(X \times Y)^n$. Note that the measurability of that random variable is ensured by the measurability of (1.1) together with the non-negativity of the loss and Tonelli's theorem.

Since the comparison of random variables is not straightforward, one typically focuses on their asymptotic behavior for an increasing data set size. A learning method $\mathcal{L} = (\mathcal{L}_n)_{n \geq 1}$ with $D \mapsto f_D$ is called *L-risk consistent* for P if $\mathcal{R}_{L,P}(f_D) \to \mathcal{R}^*_{L,P}$ in probability for $n \to \infty$, i.e.

$$\lim_{n \to \infty} P^n\Big(D \in (X \times Y)^n : \ \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \leq \varepsilon\Big) = 1 \tag{1.2}$$

is satisfied for all $\varepsilon > 0$. In other words, the learning method achieves the Bayes risk in the limit $n \to \infty$.

If a learning method $\mathcal{L}^{(p)} = (\mathcal{L}^{(p)}_n)_{n \geq 1}$ depends on a hyper parameter p, then we have, for every hyper parameter sequence $(p_n)_{n \geq 1}$, a *new* learning method $(\mathcal{L}^{(p_n)}_n)_{n \geq 1}$, whose consistency properties can be investigated. If $(\mathcal{L}^{(p_n)}_n)_{n \geq 1}$ is L-risk consistent for P, then we say that $\mathcal{L}^{(p)}$ is *L-risk consistent* for P using the hyper parameter sequence $(p_n)_{n \geq 1}$. Furthermore, we call $\mathcal{L}^{(p)}$ *potentially L-risk consistent* for P if there is a hyper parameter sequence $(p_n)_{n \geq 1}$ such that the learning method $(\mathcal{L}^{(p_n)}_n)_{n \geq 1}$ is L-risk consistent for P. Note that this notion is essentially weaker than consistency since the sequence $(p_n)_{n \geq 1}$ does not need to be specified.

Both notions are qualitative descriptions of the asymptotic performance. Consequently, we introduce next the quantitative counterpart, namely learning rates. For a set of probability measures $\mathcal{P}$ on $X \times Y$ and a positive sequence $(\varepsilon_n)_{n\geq 1}$ with $\varepsilon \to 0$ for $n \to \infty$ a learning method $D \mapsto f_D$ is said to *learn with rate* $(\varepsilon_n)_{n\geq 1}$ on $\mathcal{P}$ if for every $\tau \geq 1$ there is some constant $C_\tau > 0$ and $n_\tau \geq 1$ with

$$P^n\Big(D \in (X \times Y)^n : \ \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \leq C_\tau \varepsilon_n\Big) \geq 1 - e^{-\tau}$$

for all $\tau \geq 1$, $n \geq n_\tau$, and $P \in \mathcal{P}$. Note that in the following the set of considered probability measures $\mathcal{P}$ is only implicitly defined.

As final part of this section we recall some basic properties of the LS loss function from [76, Example 2.6]. To this end, let π_X and π_Y be the projection onto X and Y, respectively, and define the average p-th moment

$$|P|_p := \|\pi_Y\|_{L_p(P)} = \left(\int_{X\times Y} |y|^p \, \mathrm{d}P(x,y)\right)^{1/p} \tag{1.3}$$

of P for $p > 0$. In the following we assume $|P|_2 < \infty$. Then a function $f\colon X \to \mathbb{R}$ is a LS Bayes function if and only if

$$f \circ \pi_X = \mathbb{E}_P(\pi_Y|\pi_X) \tag{1.4}$$

is P-almost surely satisfied. This ensures, under consideration of the factorization lemma, see e.g. [48, Corollary 1.97], that there is a LS Bayes function $f^*_{\mathrm{LS},P}$ and that all LS Bayes functions coincide ν-almost everywhere. Moreover, Jensen's inequality yields

$$\|f^*_{\mathrm{LS},P}\|_{L_2(\nu)} = \big\|\mathbb{E}_P(\pi_Y|\pi_X)\big\|_{L_2(P)} \leq \|\pi_Y\|_{L_2(P)} = |P|_2 < \infty$$

and hence $f^*_{\mathrm{LS},P}$ is in the space $\mathcal{L}_2(\nu)$ of square ν-integrable functions. In addition, the excess LS-risk is given by

$$\mathcal{R}_{\mathrm{LS},P}(f) - \mathcal{R}^*_{\mathrm{LS},P} = \|f - f^*_{\mathrm{LS},P}\|^2_{L_2(\nu)} \ . \tag{1.5}$$

Finally, note that the Bayes risk is given by the average conditional variance

$$\mathcal{R}^*_{\mathrm{LS},P} = \int_{X\times Y} \mathbb{E}_P(\pi_Y^2|\pi_X) - \big(\mathbb{E}_P(\pi_Y|\pi_X)\big)^2 \,\mathrm{d}P \ .$$

For later use the following lemma provides a basic property of the LS Bayes function.

1.1.1 Lemma *For $A \in \mathcal{B}$ the following identity is satisfied*

$$\int_A f^*_{\mathrm{LS},P}(x)\ \mathrm{d}\nu(x) = \int_{\pi_X^{-1}(A)} y\ \mathrm{d}P(x,y) \ .$$

Proof. Using the change-of-variables formula for $\nu = P \circ \pi_X^{-1}$ and the representation $f^*_{\mathrm{LS},P} \circ \pi_X = \mathbb{E}_P(\pi_Y|\pi_X)$ from (1.4) we find

$$\int_A f^*_{\mathrm{LS},P}(x)\ \mathrm{d}\nu(x) = \int_{\pi_X^{-1}(A)} f^*_{\mathrm{LS},P} \circ \pi_X\ \mathrm{d}P = \int_{\pi_X^{-1}(A)} \mathbb{E}_P(\pi_Y|\pi_X)\ \mathrm{d}P \ .$$

Since the set $\pi_X^{-1}(A) \in \sigma(\pi_X)$ is $\sigma(\pi_X)$-measurable, the assertion follows by the properties of the conditional expectation. □

1.2 Classification

For (binary) classification problems P on $X \times Y$ we assume that the labels are $+1$ and -1, i.e. $Y = \{\pm 1\}$, and define

$$p_\pm := P\big(X \times \{\pm 1\}\big)$$

as the probability for observing a positive and a negative label, respectively. Without loss of generality we assume $0 < p_\pm < 1$ since in the cases $p_\pm \in \{0,1\}$ only one class can be observed with positive probability. Moreover, we denote the marginal distribution of the positive and negative labeled data points by

$$\nu_\pm(A) := \frac{P\big(A \times \{\pm 1\}\big)}{p_\pm} \ , \tag{1.6}$$

for $A \in \mathcal{B}$. With this notation the marginal $\nu = P \circ \pi_X^{-1}$ is given by the convex combination

$$\nu = p_+\nu_+ + p_-\nu_- \quad . \tag{1.7}$$

Note that the probability distribution P, which describes our learning problem, is uniquely defined by the two probability measures $\nu_\pm$ and by $p_\pm \in (0,1)$ with $p_+ + p_- = 1$.

Using $\nu_\pm$ and $p_\pm \in (0,1)$ with $p_+ + p_- = 1$ the sampling of a data point $(x,y) \sim P$ can be described as two-stage sampling as follows: As a first step draw $y \in \{\pm 1\}$ such that $y = +1$ with probability p_+ and $y = -1$ with probability p_-. Then, as a second step, draw $x \sim \nu_+$ if $y = +1$ and draw $x \sim \nu_-$ if $y = -1$. This procedure gives a data point $(x,y) \sim P$ following the distribution P.

Since $\nu_+ \ll \nu$ holds true, we can define another important quantity

$$\eta(x) \coloneqq p_+ \frac{\mathrm{d}\nu_+}{\mathrm{d}\nu}(x) \in [0,1]$$

for $x \in X$. Note that η equals $\eta(x) = P\big(\{\pi_Y = 1\} | \pi_X = x\big)$ and hence is the probability for observing a positive label for the input value $x \in X$. To see this, we determine the following integrals for a measurable set $A \in \mathcal{B}$

$$\int_{\pi_X^{-1}(A)} \mathbb{1}_{\{\pi_Y = 1\}} \, \mathrm{d}P = \int_{X \times Y} \mathbb{1}_{A \times \{1\}} \, \mathrm{d}P = P\big(A \times \{1\}\big) \qquad \text{and}$$
$$\int_{\pi_X^{-1}(A)} \eta \circ \pi_X \, \mathrm{d}P = \int_A \eta \, \mathrm{d}\nu = p_+ \int_A \frac{\mathrm{d}\nu_+}{\mathrm{d}\nu} \, \mathrm{d}\nu = p_+\nu_+(A) \quad .$$

According to the definition of ν_+ both integrals coincide and hence

$$\eta(x) = \mathbb{E}_P\big(\mathbb{1}_{\{\pi_Y = 1\}} | \pi_X = x\big) = P\big(\{\pi_Y = 1\} | \pi_X = x\big)$$

is proven. However, η is not uniquely defined by P. To be more precise, different versions coincide only ν-almost surely and hence we fix some specific version of η in the following. Using this version we define the sets

$$X_+ \coloneqq \big\{\eta > 1/2\big\} \qquad \text{and} \qquad X_- \coloneqq \big\{\eta < 1/2\big\} \quad . \tag{1.8}$$

Note that it is more likely to observe a positive label than a negative one for $x \in X_+$ and vice verse for $x \in X_-$. Consequently, on the set X_+ the label $+1$ should be predicted and on the set X_- the label -1.

The following lemma establishes a connection to the LS Bayes function.

1.2.1 Lemma *The following equalities are ν-almost surely satisfied*

$$2\eta - 1 = p_+ \frac{\mathrm{d}\nu_+}{\mathrm{d}\nu} - p_- \frac{\mathrm{d}\nu_-}{\mathrm{d}\nu} = f^*_{\mathrm{LS},P} \quad .$$

Proof. Using (1.7) and the definition of η we receive the first identity, namely

$$2\eta - 1 = 2\eta - \frac{\mathrm{d}(p_+\nu_+ + p_-\nu_-)}{\mathrm{d}\nu} = p_+ \frac{\mathrm{d}\nu_+}{\mathrm{d}\nu} - p_- \frac{\mathrm{d}\nu_-}{\mathrm{d}\nu} \quad .$$

The second identity is well-known, see e.g. [25, p. 11], where in this textbook the labels are $Y = \{0, 1\}$ and hence they have to be transformed to $Y = \{\pm 1\}$. □

Next, we recall some basic properties of the classification loss. With the help of the function η the Bayes classification-risk is given by

$$\mathcal{R}^*_{\mathrm{Class},P} = \int_X \min\{\eta, 1-\eta\} \, \mathrm{d}\nu \tag{1.9}$$

and a measurable function $f\colon X \to \mathbb{R}$ is a classification Bayes function if and only if

$$(2\eta - 1)\,\mathrm{sgn}(f) \geq 0$$

is satisfied ν-almost surely. Since $2\eta - 1 = f^*_{\mathrm{LS},P}$ holds true according to Lemma 1.2.1, the functions $f^*_{\mathrm{LS},P}$ and $f^*_{\mathrm{Class},P} \coloneqq \mathrm{sgn}(2\eta - 1)$ are Bayes functions for the classification loss, see [76, Example 2.4] for details. According to [9, Lemma A.1] the excess classification-risk is given by

$$\mathcal{R}_{\mathrm{Class},P}(f) - \mathcal{R}^*_{\mathrm{Class},P} = \int_{X_+ \triangle \{f \geq 0\}} |2\eta - 1| \, \mathrm{d}\nu \quad , \tag{1.10}$$

where $\triangle$ denotes the symmetric difference. Note that the right hand side does not depend on the specific version of η even if X_+, which depends on the version of η, appears.

In the next lemma we need the supports $\operatorname{supp}\nu_\pm \subseteq X$ of the probability measures $\nu_\pm$. To this end, let X be a topological space equipped with the Borel σ-algebra $\mathcal{B}(X) =: \mathcal{B}$. Then we define the support of a measure ν as

$$\operatorname{supp}\nu := \Bigg(\bigcup_{\substack{O\subseteq X \text{ open:} \\ \nu(O)=0}} O \Bigg)^c . \tag{1.11}$$

To ensure that the support is actually a set of *full measure* $\nu\big((\operatorname{supp}\nu)^c\big) = 0$ we need some additional regularity assumptions on ν or X. In the following if we need the support of a measure, we always assume that X is a *Polish space*, i.e. it is a complete separable metric space, equipped with the Borel σ-algebra $\mathcal{B}(X)$ to ensure that the support is a set of full measure. To be more precise, according to [31, Satz VIII.1.16] every locally finite measure ν on a Polish space is an (inner and outer) regular measure. Especially ν is a *Radon measure*, i.e. inner regular and locally finite, and according to [31, Lemma VIII.2.15] the support of a Radon measure is a set of full measure. Since we are only interested in probability measures ν, the support is always a set of full measure on Polish spaces. Note that there are more general spaces X which ensure that the support is a set of full measure, but we stick to Polish spaces X (equipped with the Borel σ-algebra) for convenience. In Appendix B we summarize some basic properties of the support.

In Chapter 4 classification problems *without (label) noise* are of particular interest, i.e. $\eta \in \{0,1\}$ ν-almost surely. The following lemma provides some characterizations of that property.

1.2.2 Lemma (No Noise) *For a probability measure P on $X \times \{\pm 1\}$ the following statements are equivalent:*

(i) There is no noise, i.e. $\eta \in \{0,1\}$ ν-almost surely.

(ii) $\mathcal{R}^*_{\mathrm{Class},P} = 0$.

(iii) $\nu_+ \perp \nu_-$ are singular *measures, i.e. there is a decomposition $X = A_+ \uplus A_-$ with $A_\pm \in \mathcal{B}$ and $\nu_+(A_-) = \nu_-(A_+) = 0$.*

(iv) $\nu_+(X_+) = \nu_-(X_-) = 1$.

Moreover, if X is a Polish space (equipped with the Borel σ-algebra) and $\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-$ *is a ν_+- or ν_--zero set then there is no noise.*

To see that the condition $\nu_-(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = 0$ or $\nu_+(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = 0$ is stronger than the conditions in (i)–(iv) consider the distributions $\nu_\pm$ given as convex combinations of a uniform distribution and a Dirac distribution

$$\nu_+ := \frac{\operatorname{unif}\bigl([0,1]\bigr) + \delta_{-1}}{2} \qquad \text{and} \qquad \nu_- := \frac{\operatorname{unif}\bigl([-1,0]\bigr) + \delta_{+1}}{2}$$

on $X = [-1,1]$. In this case we have $\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_- = \{-1,0,1\}$ which is neither a ν_+-zero set nor a ν_--zero set. However $(\{-1\} \cup (0,1)) \uplus ((-1,0] \cup \{1\})$ is a suitable decomposition that implies $\nu_+ \perp \nu_-$.

Proof. (i)⇔(ii) This is a direct consequence of the representation of the Bayes risk in (1.9).

(i)⇒(iii) We show that $X = \{\eta = 1\} \uplus \{\eta \neq 1\}$ is a suitable decomposition. Using the definition of η we find

$$p_-\nu_-(\eta = 1) = \int_{\{\eta=1\}} p_- \frac{\mathrm{d}\nu_-}{\mathrm{d}\nu}\ \mathrm{d}\nu = \int_{\{\eta=1\}} 1 - \eta\ \mathrm{d}\nu = 0$$

and hence $\nu_-(\eta = 1) = 0$. Analogously, we can show $\nu_+(\eta = 0) = 0$. Since $\eta \in \{0,1\}$ holds true ν-almost surely, we get $\nu_+(\eta \neq 1) = \nu_+(\eta = 0) = 0$. This proves $\nu_+ \perp \nu_-$.

(iii)⇒(iv) According to our assumption there is a decomposition $X = A_+ \uplus A_-$ with $\nu_+(A_-) = \nu_-(A_+) = 0$. Since

$$\int_{A_+} 1 - \eta\ \mathrm{d}\nu = p_- \int_{A_+} \frac{\mathrm{d}\nu_-}{\mathrm{d}\nu}\ \mathrm{d}\nu = p_-\nu_-(A_+) = 0$$

and $0 \leq \eta \leq 1$ are satisfied, we find $\nu(\{\eta \neq 1\} \cap A_+) = 0$. Together with the definition of X_+ and $\nu_+(A_+) = 1$ we get

$$\nu_+(X_+^{\mathrm{c}}) = \nu_+(\eta \leq 1/2) \leq \nu_+(\eta \neq 1) = \nu_+\bigl(\{\eta \neq 1\} \cap A_+\bigr) = 0\ \ .$$

This proves $\nu_+(X_+) = 1$. The identity $\nu_-(X_-) = 1$ follows analogously.

(iv)⇒(i) We define the function $\eta' := \mathbb{1}_{X_+}$ and show that $\eta' = p_+ \mathrm{d}\nu_+/\mathrm{d}\nu$ ν-almost surely. Note that X_+ and hence η' depends on some version of η, which we do not specify. For $A \in \mathcal{B}$ we have

$$\int_A \eta' \,\mathrm{d}\nu = \nu(A \cap X_+) = p_+\nu_+(A \cap X_+) + p_-\nu_-(A \cap X_+) = p_+\nu_+(A) \ ,$$

where we used in the last step $\nu_+(X_+) = 1$ and $\nu_-(A \cap X_+) \leq \nu_-(X_-^c) = 0$. This proves $\eta' = p_+ \mathrm{d}\nu_+/\mathrm{d}\nu$ ν-almost surely. Since $\eta' \in \{0, 1\}$ holds true, every version of η satisfies $\eta \in \{0, 1\}$ ν-almost surely.

Finally, we assume that $\nu_-(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = 0$ and show Point (iii). To this end, we consider the decomposition $X = \operatorname{supp}\nu_+ \uplus (\operatorname{supp}\nu_+)^c$. Then $\nu_+((\operatorname{supp}\nu_+)^c) = 0$ is a consequence of the fact that the support is a set of full measure. Using again this property of the support and our assumption gives $\nu_-(\operatorname{supp}\nu_+) = \nu_-(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = 0$. As a result, this is a suitable decomposition which shows Point (iii). □

In Lemma 1.2.2 we characterized learning problems without noise. Sometimes it is useful to have a more precise quantification of the amount of noise in a classification problem. To this end, we introduce the *noise function* $N\colon [0, \infty) \to [0, 1]$ given by $N(t) := \nu(|2\eta - 1| < 2t)$. Moreover, we say that the distribution P has *noise exponent* $0 \leq q \leq \infty$ if there is some constant $c_N > 0$ such that, for $t \geq 0$,

$$N(t) \leq (c_N t)^q \tag{1.12}$$

holds true, c.f. [76, Definition 8.22] and [61, Equation (4)]. This condition is also known as *Tsybakov's noise condition* in the literature. Note that if there is no noise then we have $N(t) = \mathbb{1}_{(1/2,\infty)}(t)$. Consequently, we have the noise exponent $q = \infty$ with $c_N = 2$.

If we have a pseudo-metric space (X, d), i.e. $d(x, x') = 0$ does not imply $x = x'$, as input space, we can additionally introduce the following quantities, cf. [76, Defintion 8.6 and Defintion 8.15]: The *distance to the decision* *boundary* is given by

$$\Delta_d(x) := \begin{cases} \operatorname{dist}(x, X_+), & x \in X_- \\ \operatorname{dist}(x, X_-), & x \in X_+ \\ 0, & \text{else}, \end{cases} \tag{1.13}$$

with $\operatorname{dist}(x, A) := \inf_{a \in A} d(x, a)$, for $A \subseteq X$. Note that the mapping $x \mapsto \operatorname{dist}(x, A)$, for fixed $A \subseteq X$, is Lipschitz continuous and hence $\Delta_d \colon X \to \mathbb{R}$ is measurable if the σ-algebra $\mathcal{B}$ on X contains the Borel σ-algebra $\mathcal{B}(X, d)$, that is

$$\mathcal{B} \supseteq \mathcal{B}(X, d) \ . \tag{1.14}$$

In this case we can further define the *margin function* $M_d \colon [0, \infty) \to [0, 1]$ and the *margin-noise function* $MN_d \colon [0, \infty) \to [0, 1]$ with respect to Δ_d by

$$\begin{aligned} M_d(r) &:= \nu(\Delta_d \leq 2r) \qquad \text{and} \\ MN_d(r) &:= \int_{\{\Delta_d \leq 2r\}} |2\eta - 1| \, \mathrm{d}\nu \end{aligned} \ , \tag{1.15}$$

respectively. Using Lemma 1.2.1 the margin-noise function equals

$$MN_d(r) = \big|p_+\nu_+ - p_-\nu_-\big|(\Delta_d \leq 2r)$$

where $|p_+\nu_+ - p_-\nu_-|$ denotes the total variation of the signed measure $p_+\nu_+ - p_-\nu_-$. Note that the quantities Δ_d, M_d, and MN_d additionally depend on the specific version of η. In contrast to the notation in [76] we use the metric d in the subscript of Δ_d in instead of η since we will later use different metric but a fixed version of η.

1.3 Histograms

A histogram is a learning method that is based on a given partition $\mathcal{A} = (A_k)_{k \in K}$ of the input space X, where K is some index set. We call A_k a *cell* of $\mathcal{A}$ and say that the partition $\mathcal{A}$ is *measurable* if every cell is measurable, i.e. $A_k \in \mathcal{B}$ for all $k \in K$. Moreover, we say that the partition $\mathcal{A}$ is *countable* if K is countable. If (X, d) is a pseudo-metric space, we define the *diameter* of $\mathcal{A}$ by $\operatorname{diam}(\mathcal{A}) := \sup_{k\in K} \operatorname{diam}(A_k)$, where $\operatorname{diam}(A) := \sup_{x,x'\in A} d(x, x')$ for $A \subseteq X$. For a measurable partition $\mathcal{A} = (A_k)_{k\in K}$ we denote the indexes of *relevant cells* by

$$\mathcal{A}_\nu := \big\{k \in K : \ \nu(A_k) > 0\big\} \ . \tag{1.16}$$

For a distribution P on $X \times Y$ with $|P|_2 < \infty$ and a measurable and countable partition $\mathcal{A} = (A_k)_{k\in K}$ the corresponding *(population or infinite-sample) histogram* $h_{P,\mathcal{A}} \colon X \to \mathbb{R}$ is given by

$$h_{P,\mathcal{A}} := \sum_{k\in\mathcal{A}_\nu} \mathbb{1}_{A_k} \cdot \frac{1}{P(A_k \times Y)} \int_{A_k\times Y} y \,\mathrm{d}P(x, y) \ . \tag{1.17}$$

Using Lemma 1.1.1 the histogram can be written as

$$h_{P,\mathcal{A}} = \sum_{k\in\mathcal{A}_\nu} \mathbb{1}_{A_k} \cdot \frac{1}{\nu(A_k)} \int_{A_k} f^*_{\mathrm{LS},P} \,\mathrm{d}\nu = \mathbb{E}_\nu\big(f^*_{\mathrm{LS},P}|\sigma(\mathcal{A})\big) \ , \tag{1.18}$$

where we used a well-known representation of the conditional expectation with respect to a σ-algebra generated by a measurable and countable partition, see [4, Equation (15.3)]. Consequently, histograms are constant on each cell and take there the average value of $f^*_{\mathrm{LS},P}$. For a data set $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ the corresponding *(empirical) histogram* is given by

$$h_{D,\mathcal{A}}(x) = \sum_{k\in\mathcal{A}_\delta} \mathbb{1}_{A_k}(x) \cdot \frac{\sum_{i=1}^n y_i \mathbb{1}_{A_k}(x_i)}{\sum_{i=1}^n \mathbb{1}_{A_k}(x_i)} \ ,$$

where $\delta := D \circ \pi_X^{-1}$ denotes the marginal distribution of the empirical distribution D. As a result, $D \mapsto h_{D,\mathcal{A}}$ defines a learning method in the sense of (1.1) in which the predefined partition $\mathcal{A}$ is a hyper parameter.

In the case $X = \mathbb{R}^p$ for some $p \geq 1$ the partition $\mathcal{A} = (A_k)_{k\in K}$ of $\mathbb{R}^p$ is called *cubic* with radius $r > 0$ (or width $2r$) if for every $k \in K$, there is some $z_k \in \mathbb{R}^p$ with $A_k = z_k + (-r, r]^p$. Note that cubic partitions are automatically Borel measurable and satisfy $\operatorname{diam}(\mathcal{A}) = 2r$ with respect to the ℓ_∞^d-norm. Moreover, cubic partitions are automatically countable, see the discussion after Lemma 1.3.7 below.

Next, we introduce two oracle inequalities for the histogram learning method. The proofs are deferred to Appendix A.

1.3.1 Lemma (LS-Risk Oracle Inequality) *Let P be a distribution on $X \times Y$ with $Y = [-M, M]$ for some $M > 0$ and marginal distribution ν. Furthermore, let $\mathcal{A} = (A_k)_{k \in K}$ be a measurable and countable partition of X. Then the following bound is satisfied, for all $0 < \varepsilon \leq M$, $\tau \geq 1$, and $n \geq 1$,*

$$\begin{aligned}\mathcal{R}_{\mathrm{LS},P}(h_{D,\mathcal{A}}) - \mathcal{R}^*_{\mathrm{LS},P} < 4\big\|h_{P,\mathcal{A}} - f^*_{\mathrm{LS},P}\big\|^2_{L_2(\nu)} \\ + 20M\varepsilon \\ + 1536M^2 \log(3M/\varepsilon) \cdot \frac{\tau|\mathcal{A}_\nu|}{n}\end{aligned}$$

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

1.3.2 Lemma (Classification-Risk Oracle Inequality) *Let (X, d) be a pseudo-metric space equipped with a σ-algebra $\mathcal{B} \supseteq \mathcal{B}(X, d)$ containing the Borel σ-algebra and P be a distribution on $X \times \{\pm 1\}$ with marginal distribution ν and noise exponent $0 \leq q \leq \infty$ (and constant $c_N > 0$) defined in (1.12). Furthermore, let $r > 0$ and $\mathcal{A} = (A_k)_{k \in K}$ be a measurable and countable partition of X with $\operatorname{diam}(\mathcal{A}) \leq 2r$. Then there is a constant $C > 0$, depending only on q and c_N, such that the following bound is satisfied, for all $\tau \geq 1$ and $n \geq 1$,*

$$\mathcal{R}_{\mathrm{Class},P}(h_{D,\mathcal{A}}) - \mathcal{R}^*_{\mathrm{Class},P} < 6MN_d(r) + C\left(\frac{\tau|\mathcal{A}_\nu|}{n}\right)^{\frac{q+1}{q+2}}$$

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

For both oracle inequalities, the number of relevant cells $|\mathcal{A}_\nu|$ is a key quantity. Therefore, we provide basic properties of $\mathcal{A}_\nu$ in the following. The indexes in $\mathcal{A}_\nu$ specify the relevant cells with respect to the measure ν, however, often it is easier to investigate the relevant cells $\mathcal{A}_M$ with respect to some subset $M \subseteq X$, namely

$$\mathcal{A}_M := \big\{ k \in K : \; A_k \cap M \neq \emptyset \big\} \; . \tag{1.19}$$

Both notions of relevant cells are closely related as we will see in the following. But we start with some basic properties.

1.3.3 Lemma (Basic Properties) *Let $\mathcal{A} = (A_k)_{k\in K}$ be a partition of X. Then the following statements are true:*

(i) $\mathcal{A}_N \subseteq \mathcal{A}_M$ for subsets $N \subseteq M \subseteq X$.

(ii) $\mathcal{A}_M = \mathcal{A}_{M_+} \cup \mathcal{A}_{M_-}$ for subsets $M_\pm \subseteq X$ with $M := M_+ \cup M_-$.

If $\mathcal{A}$ is measurable then the following statements are true:

(iii) $\mathcal{A}_\mu \subseteq \mathcal{A}_\nu$ for absolute continuous measures $\mu \ll \nu$ on X.

(iv) $\mathcal{A}_\nu = \mathcal{A}_{\nu_+} \cup \mathcal{A}_{\nu_-}$ for measures $\nu_\pm$ on X with $\nu := \nu_+ + \nu_-$.

Proof. (i) and (iii) follow directly from the definitions in (1.19) and (1.16), respectively.

(ii) For $k \in K$ we have $k \in \mathcal{A}_M$ if and only if $A_k \cap M_+ \neq \emptyset$ or $A_k \cap M_- \neq \emptyset$. The latter is equivalent to $k \in \mathcal{A}_{M_+} \cup \mathcal{A}_{M_-}$.

(iv) For $k \in K$ we have $k \in \mathcal{A}_\nu$ if and only if $\nu_+(A_k) > 0$ or $\nu_-(A_k) > 0$. The latter is equivalent to $k \in \mathcal{A}_{\nu_+} \cup \mathcal{A}_{\nu_-}$. □

The next lemma establishes a relation between $\mathcal{A}_\nu$ and $\mathcal{A}_M$.

1.3.4 Lemma ($\mathcal{A}_\nu$ vs. $\mathcal{A}_M$) *Let $\mathcal{A} = (A_k)_{k\in K}$ be a measurable partition of X, ν be a measure on X, and $M \subseteq X$ be a measurable subset. If M is a set of full measure, i.e. $\nu(M^c) = 0$, then $\mathcal{A}_\nu \subseteq \mathcal{A}_M$ is satisfied.*

If ν is a Radon measure on a Polish space X, this lemma directly gives us $\mathcal{A}_\nu \subseteq \mathcal{A}_{\operatorname{supp}\nu}$.

Proof. Let $k \in \mathcal{A}_\nu$ be fixed. Since M is a set of full measure, we have $\nu(A_k \cap M) = \nu(A_k) > 0$ and this implies $A_k \cap M \neq \emptyset$. As a result, we find $k \in \mathcal{A}_M$. □

For the next result we need covering numbers. For $r > 0$, the *covering number* $\mathcal{N}(M, r)$ of a subset $M \subseteq X$ of a metric space is defined as the minimum number of closed balls of radius r needed to cover the set M. For basic properties of covering numbers see e.g. Appendix C.

1.3.5 Lemma (Lower Bound for $\mathcal{A}_M$ in Metric Spaces) *Let X be a metric space, $\mathcal{A} = (A_k)_{k\in K}$ be a partition of X and $M \subseteq X$ be a subset. If there are $b_k \in X$ and a $r > 0$ such that A_k is a subset of the closed ball with center b_k and radius r, i.e.*

$$A_k \subseteq B_X(b_k, r) \quad , \tag{1.20}$$

for all $k \in \mathcal{A}_M$ then $|\mathcal{A}_M| \geq \mathcal{N}(M, r)$ is satisfied.

Proof. The assumption (1.20) implies

$$M \subseteq \bigcup_{k\in\mathcal{A}_M} A_k \subseteq \bigcup_{k\in\mathcal{A}_M} B_X(b_k, r)$$

and hence the points b_k form an r-net of M. Since $\mathcal{N}(M, r)$ is the cardinality of a minimal r-net, we get the desired lower bound. □

The next corollary transfers this lower bound for $\mathcal{A}_{\operatorname{supp}\nu}$ to $\mathcal{A}_\nu$.

1.3.6 Corollary (Lower Bound for $\mathcal{A}_\nu$ in Metric Spaces) *Let X be a Polish space, ν be a probability measure on X, and $\mathcal{A} = (A_k)_{k\in K}$ be a measurable and countable partition of X satisfying the condition in* (1.20) *for some $r > 0$. Then the following bound is satisfied*

$$|\mathcal{A}_\nu| \geq \mathcal{N}(\operatorname{supp}\nu, r) \quad .$$

Proof. We define the set $M := \bigcup_{k\in\mathcal{A}_\nu} A_k$. Since $\mathcal{A}$ is countable, the set $M^c = \bigcup_{k\notin\mathcal{A}_\nu} A_k$ is as countable union of ν-zero sets a ν-zero set. In other words, M is a set of full measure that additionally satisfies $\mathcal{A}_M = \mathcal{A}_\nu$. Now, Lemma 1.3.5 gives us

$$|\mathcal{A}_\nu| = |\mathcal{A}_M| \geq \mathcal{N}(M, r) = \mathcal{N}(\overline{M}, r) \quad ,$$

where we used Point (iv) of Lemma C.4 in the last step. Since $\overline{M}$ is a closed set of full measure, we have $\operatorname{supp}\nu \subseteq \overline{M}$ and the assertion is proven. □

1.3.7 Lemma (Upper Bound for $\mathcal{A}_M$ in $\mathbb{R}^d$) *Let $X = \mathbb{R}^d$ be equipped with some norm $\|\cdot\|$ and B_X the closed unit ball with respect to that norm, $\mathcal{A} = (A_k)_{k\in K}$ be a measurable partition of X, and $M \subseteq X$ be a subset. If there are $a_k \in \mathbb{R}^d$ and $0 < r_0 \leq r$ with*

$$A_k \supseteq a_k + r_0\mathring{B}_X \qquad \textit{and} \qquad \operatorname{diam}(A_k) \leq 2r \tag{1.21}$$

for all $k \in \mathcal{A}_M$ then the following bound is satisfied, for $\varepsilon > 0$,

$$|\mathcal{A}_M| \leq \bigl(\varepsilon/r_0 + 2r/r_0\bigr)^d \cdot \mathcal{N}(M, \varepsilon)\ .$$

Since $\mathcal{N}([-b,b]^d, r) < \infty$ is finite for every $b > 0$ and $\mathbb{R}^d = \bigcup_{b\in\mathbb{N}}[-b,b]^d$ holds true, this lemma proves that every measurable partition satisfying (1.21) is countable.

Note that for a cubic partition with radius $r > 0$ and $\|\cdot\| = \|\cdot\|_{\ell_\infty^d}$ we can choose $r_0 = r$ and hence $|\mathcal{A}_M| \leq 3^d \cdot \mathcal{N}(M, r)$.

Proof. We use a volume argument to prove this lemma. To this end, we denote the Lebesgue measure on $\mathbb{R}^d$ by λ^d. Let $\varepsilon > 0$ be fixed. Since there is nothing to prove in the case $\mathcal{N}(M,\varepsilon) = \infty$, we can assume that $n := \mathcal{N}(M,\varepsilon) < \infty$ is finite. Moreover, let $L \subseteq \mathcal{A}_M$ be a finite subset. Since $a_k + r_0\mathring{B}_{r_0} \subseteq A_k$ for $k \in \mathcal{A}_M$ and the cells A_k are disjoint for $k \in K$, we find

$$\begin{aligned} |L| \cdot r_0^d \cdot \lambda^d\bigl(\mathring{B}_X\bigr) &= \sum_{k\in L} \lambda^d(a_k + r_0\mathring{B}_{r_0}) \\ &\leq \sum_{k\in L} \lambda^d(A_k) \\ &= \lambda^d\biggl(\bigcup_{k\in L} A_k\biggr)\ . \end{aligned} \tag{1.22}$$

For every $k \in \mathcal{A}_M$ there is some $y_k \in M \cap A_k$ and hence $\operatorname{diam}(A_k) \leq 2r$ ensures that $A_k \subseteq y_k + 2rB_X$ holds true. This shows that A_k is a subsets of $M + 2rB_X$ for all $k \in \mathcal{A}_M$, i.e.

$$\bigcup_{k\in L} A_k \subseteq M + 2rB_X \quad .$$

Now, let $x_1, \ldots, x_n \in X$ be a minimal ε-net of M, i.e. $M \subseteq \bigcup_{i=1}^n x_i + \varepsilon B_X$. This gives

$$\bigcup_{k\in L} A_k \subseteq \bigcup_{i=1}^{n} \bigl(x_i + \varepsilon B_X + 2rB_X\bigr) = \bigcup_{i=1}^{n} \bigl(x_i + (\varepsilon + 2r)B_X\bigr) \quad ,$$

where we used the convexity of B_X in the last step. Finally, if we plug this into the right hand side of (1.22) we find

$$|L| \cdot r_0^d \cdot \lambda^d(\mathring{B}_X) \le n \cdot (\varepsilon + 2r)^d \cdot \lambda^d(B_X) \quad .$$

Note that we do not apply the Lebesgue measure to the set M. For that reason, we do not need the measurability of M. Since B_X is a convex set, [31, Satz II.7.7] gives us $\lambda^d(\mathring{B}_X) = \lambda^d(B_X)$. Consequently, we find the upper bound $|L| \le (\varepsilon/r_0 + 2r/r_0)^d \cdot n$. Since this bound is satisfied by every finite subset $L \subseteq \mathcal{A}_M$, the set $\mathcal{A}_M$ itself is finite and satisfies the claimed inequality. □

Finally, the following corollary summarizes our findings for $X = \mathbb{R}^d$.

1.3.8 Corollary (Relevant Cells in $\mathbb{R}^d$) *Let $X = \mathbb{R}^d$ be equipped with some norm $\|\cdot\|$, $\mathcal{A} = (A_k)_{k\in K}$ be a measurable partition of X satisfying the conditions in* (1.20) *and* (1.21) *for some $0 < r_0 \le r$, and ν be a probability measure on X. Then the following bounds are satisfied, for $\varepsilon > 0$,*

$$\mathcal{N}(\operatorname{supp}\nu, r) \le |\mathcal{A}_\nu| \le |\mathcal{A}_{\operatorname{supp}\nu}| \le \bigl(\varepsilon/r_0 + 2r/r_0\bigr)^d \cdot \mathcal{N}(\operatorname{supp}\nu, \varepsilon) \quad .$$

Note that the condition in (1.20) already yields $\operatorname{diam}(A_k) \le 2r$ which is the second part of the condition in (1.21).

Proof. The remark after Lemma 1.3.7 implies that the partition $\mathcal{A}$ is countable. As a result, the first inequality is a consequence of Corollary 1.3.6.

Since $\operatorname{supp}\nu$ is a set of full measure, Lemma 1.3.4 gives us the second inequality $\mathcal{A}_\nu \subseteq \mathcal{A}_{\operatorname{supp}\nu}$. Together with Lemma 1.3.7 for $M = \operatorname{supp}\nu$ we find the third inequality. □

Chapter 2

Transformed Learning Scenarios

In this chapter we provide a framework that allows us to model a learning scenario with restricted access to the data set. Based on this framework we describe high-dimensional learning scenarios as a subset or a projection of an infinite-dimensional learning scenario in Chapter 3 below. Since this framework is possibly of interest in its own right, we devote this chapter to it. For the special case of classification, some of the following ideas can already be found in [25, Chapter 32], see also the references therein.

2.1 Definitions and Basic Properties

In real world applications we often do not have access to the original data set $D = \big((x_1, y_1), \dots, (x_n, y_n)\big) \sim P^n$. Instead our algorithm learns from the data set

$$s(D) \coloneqq \big((\bar{x}_1, y_1), \dots, (\bar{x}_n, y_n)\big) \ ,$$

where $\bar{x}_i \coloneqq s(x_i)$, for $i = 1, \dots, n$, with some measurable function

$$s \colon X \to \bar{X}$$

which maps $(X, \mathcal{B})$ into a further measurable space $(\bar{X}, \bar{\mathcal{B}})$. In other words, for all the input values x_i, the learning method can only access the information preserved by the mapping s. Since the data set $s(D) \sim \bar{P}^n$ follows the probability distribution

$$\bar{P} := P \circ (s, \mathrm{id}_Y)^{-1} \tag{2.1}$$

on $\bar{X} \times Y$, this defines a new learning scenario which we call *transformed learning scenario* of P under s. In the transformed scenario we denote all quantities with a bar, e.g. we write $\bar{D} := s(D)$, $\bar{\nu}$, etc. In this context we call the learning scenario given by P *original learning scenario*. Note that, for $\bar{X} = X$ and $s = \mathrm{id}_X$, the transformed learning scenario coincides with the original scenario.

In contrast to the original scenario, which is given by the application at hand, the measurable space $\bar{X}$ and the transformation $s \colon X \to \bar{X}$ can be chosen—at least in some applications—by the user. For example, in some image classification tasks the camera and its resolution can be selected by the user. In this case each camera and resolution corresponds to a different transformation s.

In order to measure the quality of a decision function $\bar{f} \colon \bar{X} \to \mathbb{R}$ in the transformed scenario we pull back the decision function $\bar{f} \circ s \colon X \to \mathbb{R}$ and measure its quality using risks in the original scenario, that is

$$\mathcal{R}_{L,P}(\bar{f} \circ s) - \mathcal{R}^*_{L,P} \quad . \tag{2.2}$$

Analogously, for a learning method $\bar{\mathcal{L}} = (\bar{\mathcal{L}}_n)_{n \geq 1}$ on $\bar{X} \times Y$, in the sense of (1.1), we define the *pull-back learning method* $\mathcal{L} = (\mathcal{L}_n)_{n \geq 1}$ on $X \times Y$ given by

$$\mathcal{L}_n(D) := \bar{\mathcal{L}}_n\big(s(D)\big) \circ s \quad .$$

In other words, we transform the data set $D \in (X \times Y)^n$ to $s(D) \in (\bar{X} \times Y)^n$, apply the learning method $\bar{\mathcal{L}}$ on $\bar{X} \times Y$ using $s(D)$, and pull back the produced decision function on X. We say that the learning method $\bar{\mathcal{L}}$ is *(potentially) L-risk consistent* for P and s if the corresponding pull-back learning method $\mathcal{L}$ is (potentially) L-risk consistent for P in the sense of (1.2). Analogously, we define the notion of learning rates for $\bar{\mathcal{L}}$.

The next lemma provides basic properties and relations between the original and the transformed learning scenario.

2.1.1 Lemma (Basic Properties) *Let P be a probability distribution on $X \times Y$, $s\colon X \to \bar{X}$ be a measurable function, and $\bar{P} = P \circ (s, \mathrm{id}_Y)^{-1}$ be the distribution of the transformed learning scenario. Then the following statements are true:*

(i) The marginal distributions $\nu = P \circ \pi_X^{-1}$ and $\bar{\nu} = \bar{P} \circ \pi_{\bar{X}}^{-1}$ of P on X and $\bar{P}$ on $\bar{X}$, respectively, satisfy $\bar{\nu} = \nu \circ s^{-1}$.

(ii) For a data set $D \in (X \times Y)^n$ and $\bar{D} := s(D) \in (\bar{X} \times Y)^n$ the corresponding empirical measures satisfy $\bar{D} = D \circ (s, \mathrm{id}_Y)^{-1}$.

(iii) For $0 < p < \infty$ the average p-th moments, defined in (1.3), coincide $|P|_p = |\bar{P}|_p$.

(iv) For a supervised loss function $L\colon Y \times \mathbb{R} \to [0, \infty)$ and a measurable function $\bar{f}\colon \bar{X} \to \mathbb{R}$ the L-risk is given by

$$\mathcal{R}_{L,\bar{P}}(\bar{f}) = \mathcal{R}_{L,P}(\bar{f} \circ s) \enspace .$$

(v) For another measurable function $\bar{s}\colon \bar{X} \to X'$ into another measurable space $(X', \mathcal{B}')$ the transformed learning scenario $P' = \bar{P} \circ (\bar{s}, \mathrm{id}_Y)^{-1}$ of $\bar{P}$ under $\bar{s}$ equals the transformed learning scenario of P under $\bar{s} \circ s$.

Note that Point (iv) only considers supervised loss functions. For general loss functions, the loss must be transformed as well. Since we are mainly interested in the LS loss and the classification loss, which are supervised, we stick to supervised loss functions for convenience.

Point (v) states that building transformed learning scenarios is transitive.

Proof. (i) The definitions of $\bar{\nu}$ and $\bar{P}$ ensure

$$\bar{\nu} = \bar{P} \circ \pi_{\bar{X}}^{-1} = P \circ (s, \mathrm{id}_Y)^{-1} \circ \pi_{\bar{X}}^{-1} = P \circ \big(\pi_{\bar{X}} \circ (s, \mathrm{id}_Y)\big)^{-1} \enspace .$$

Using the identity $\pi_{\bar{X}} \circ (s, \mathrm{id}_Y) = s \circ \pi_X$ we find $\bar{\nu} = P \circ \pi_X^{-1} \circ s^{-1} = \nu \circ s^{-1}$.

(ii) For a data point $(x, y) \in X \times Y$ and a measurable set $A \subseteq \bar{\mathcal{B}} \times Y$ the corresponding Dirac measures satisfy

$$\begin{aligned}\delta_{(s(x),y)}(A) &= \mathbb{1}_A \circ (s, \mathrm{id}_Y)(x,y) \\ &= \mathbb{1}_{(s,\mathrm{id}_Y)^{-1}(A)}(x,y) \\ &= \delta_{(x,y)} \circ (s, \mathrm{id}_Y)^{-1}(A) \ \ .\end{aligned}$$

This proves the desired identity $\delta_{(s(x),y)} = \delta_{(x,y)} \circ (s, \mathrm{id}_Y)^{-1}$ for a single data point. Using

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} \delta_{(s(x_i),y_i)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i,y_i)} \circ (s, \mathrm{id}_Y)^{-1} = D \circ (s, \mathrm{id}_Y)^{-1} \ \ ,$$

we get the assertion for general data sets.

(iii) Using the change-of-variables formula for $\bar{P} = P \circ (s, \mathrm{id}_Y)^{-1}$ we get the assertion

$$|\bar{P}|_p^p = \int_{\bar{X}\times Y} |y|^p \ \mathrm{d}\bar{P}(\bar{x}, y) = \int_{X\times Y} |y|^p \ \mathrm{d}P(x,y) = |P|_p^p \ \ .$$

(iv) Using the change-of-variables formula for $\bar{P} = P \circ (s, \mathrm{id}_Y)^{-1}$ we get the assertion

$$\begin{aligned}\mathcal{R}_{L,\bar{P}}(\bar{f}) &= \int_{\bar{X}\times Y} L\big(y, \bar{f}(\bar{x})\big) \ \mathrm{d}\bar{P}(\bar{x}, y) \\ &= \int_{X\times Y} L\big(y, \bar{f}(s(x))\big) \ \mathrm{d}P(x,y) \\ &= \mathcal{R}_{L,P}(\bar{f} \circ s) \ \ .\end{aligned}$$

(v) This is a direct consequence of $(\bar{s} \circ s)^{-1} = s^{-1} \circ \bar{s}^{-1}$. To be more precise,

$$P \circ (\bar{s} \circ s)^{-1} = P \circ s^{-1} \circ \bar{s}^{-1} = \bar{P} \circ \bar{s}^{-1} = P'$$

proves the assertion. □

The next lemma compares the Bayes risks of the original and the transformed scenario.

2.1.2 Lemma (Bayes Risk) *Let P be a probability distribution on $X \times Y$, $s\colon X \to \bar{X}$ be a measurable function, and $\bar{P} = P \circ (s, \mathrm{id}_Y)^{-1}$ be the distribution of the transformed learning scenario. Furthermore, let $L\colon Y \times \mathbb{R} \to [0, \infty)$ be a supervised loss function. Then the following statements are true:*

(i) The Bayes risks satisfy $\mathcal{R}^*_{L,\bar{P}} = \mathcal{R}^*_{L,P,\mathcal{L}_0(X,\sigma(s))} \geq \mathcal{R}^*_{L,P}$.

(ii) There is a $\sigma(s)$-measurable Bayes function $f^*_{L,P}\colon X \to \mathbb{R}$ if and only if there is a Bayes function $f^*_{L,\bar{P}}\colon \bar{X} \to \mathbb{R}$ and $\mathcal{R}^*_{L,\bar{P}} = \mathcal{R}^*_{L,P}$ holds true.

The inequality $\mathcal{R}^*_{L,\bar{P}} \geq \mathcal{R}^*_{L,P}$ from Point (i) states that the transformed learning scenario provides at most as much relevant information as the original one. Here relevance is specified by the loss function L. If the Bayes risks coincide $\mathcal{R}^*_{L,\bar{P}} = \mathcal{R}^*_{L,P}$, we interpret this in such a way that no relevant information gets lost when considering the transformed learning scenario instead of the original one. Moreover, the representation of $\mathcal{R}^*_{L,\bar{P}}$ shows that $\sigma(s) = \mathcal{B}$ is a sufficient condition to ensures $\mathcal{R}^*_{L,\bar{P}} = \mathcal{R}^*_{L,P}$ for all distributions P on $X \times Y$ and for all supervised loss functions L.

Point (ii) shows that if a Bayes functions $f^*_{L,\bar{P}}$ exists then $\mathcal{R}^*_{L,\bar{P}} = \mathcal{R}^*_{L,P}$ is equivalent to the existence of a Bayes function $f^*_{L,P}$ satisfying a certain symmetry condition. Moreover, Point (ii) can be found in [25, Theorem 32.5] for the special case of binary classification.

Proof. (i) Note that according to [48, Corollary 1.97] the set of $\sigma(s)$-measurable functions is given by $\mathcal{L}_0(X, \sigma(s)) = \{\bar{f} \circ s \ : \ \bar{f} \in \mathcal{L}_0(\bar{X}, \bar{\mathcal{B}})\}$. Consequently, Point (iv) of Lemma 2.1.1 gives us the claimed equality, namely

$$\begin{aligned} \mathcal{R}^*_{L,\bar{P}} &= \inf_{\bar{f} \in \mathcal{L}_0(\bar{X}, \bar{\mathcal{B}})} \mathcal{R}_{L,\bar{P}}(\bar{f}) \\ &= \inf_{\bar{f} \in \mathcal{L}_0(\bar{X}, \bar{\mathcal{B}})} \mathcal{R}_{L,P}(\bar{f} \circ s) \\ &= \inf_{\bar{f} \in \mathcal{L}_0(X, \sigma(s))} \mathcal{R}_{L,P}(f) \ . \end{aligned}$$

The claimed inequality is a direct consequence of $\sigma(s) \subseteq \mathcal{B}$ which is ensured by the measurability of s.

(ii) First we assume that there is a $\sigma(s)$-measurable Bayes function $f^*_{L,P}\colon X \to \mathbb{R}$. Then [48, Corollary 1.97] gives us a measurable function $\bar{f}\colon \bar{X} \to \mathbb{R}$ with $f^*_{L,P} = \bar{f} \circ s$. Together with Point (i) we find

$$\mathcal{R}^*_{L,\bar{P}} \geq \mathcal{R}^*_{L,P} = \mathcal{R}_{L,P}(f^*_{L,P}) = \mathcal{R}_{L,P}(\bar{f} \circ s) = \mathcal{R}_{L,\bar{P}}(\bar{f}) \geq \mathcal{R}^*_{L,\bar{P}} \ .$$

This shows that $f^*_{L,\bar{P}} := \bar{f}$ is a Bayes function and $\mathcal{R}^*_{L,\bar{P}} = \mathcal{R}^*_{L,P}$. For the converse implication we use

$$\mathcal{R}^*_{L,P} = \mathcal{R}^*_{L,\bar{P}} = \mathcal{R}_{L,\bar{P}}(f^*_{L,\bar{P}}) = \mathcal{R}_{L,P}(f^*_{L,\bar{P}} \circ s) \ .$$

This shows that $f^*_{L,P} := f^*_{L,\bar{P}} \circ s$ is a $\sigma(s)$-measurable Bayes function. □

The following lemma investigates sufficient and necessary conditions that ensure $\sigma(s) = \mathcal{B}$. To this end, we recall the definition of the trace σ-algebra: For a subset $A \subseteq X$ the *trace σ-algebra* $\mathcal{B}|_A$ is given by

$$\mathcal{B}|_A := \big\{ A \cap B : \ B \in \mathcal{B} \big\} \ .$$

Note that A does not need to be measurable as subset of X.

2.1.3 Lemma (No Loss of Information) *Let $s\colon X \to \bar{X}$ be a measurable function and consider the following statements:*

(i) $\sigma(s) = \mathcal{B}$.

(ii) $s(A) \in \bar{\mathcal{B}}|_{s(X)}$ *for all* $A \in \mathcal{B}$.

(iii) s *is injective.*

(iv) $\{x\} \in \mathcal{B}$ *for all* $x \in X$.

Then the implications (i)⇒(ii), (ii)+(iii)⇒(i), and (i)+(iv)⇒(iii) hold true.

Note that for injective $s\colon X \to \bar{X}$ the condition in (ii) equals the measurability of the mapping $s^{-1}\colon s(X) \to X$, where $s(X)$ is equipped with the trace σ-algebra $\bar{\mathcal{B}}_{s(X)}$.

Let us mention some special cases in which this lemma provides $\sigma(s) = \mathcal{B}$: If $s\colon X \to \bar{X}$ is bijective with measurable inverse $s^{-1}\colon \bar{X} \to X$ then Point (ii) and (iii) are satisfied and hence this lemma yields $\sigma(s) = \mathcal{B}$.

Next, if the measurable space X satisfies Point (iv) then $\sigma(s) = \mathcal{B}$ is equivalent to Point (ii) *and* (iii). Note that Point (iv) is satisfied for every metric space equipped with the Borel σ-algebra. But there are more spaces satisfying Point (iv) e.g. if $(X, \mathcal{B})$ is countably separated, i.e. there is a countable subfamily $\mathcal{B}_0 \subseteq \mathcal{B}$ separating points, see e.g. [11, Theorem 6.5.7] for details. A class of countably separated spaces are Souslin spaces, see e.g. [11, Corollary 6.7.5] or [22, Lemma 8.6.12] for details.

Note that for particular distributions P and loss functions L the condition $\sigma(s) = \mathcal{B}$ is not necessary to ensure $\mathcal{R}^*_{L,\bar{P}} = \mathcal{R}^*_{L,P}$, see e.g. Example 2.1.5 and Example 2.1.7 below.

Proof. (i)⇒(ii) Let $A \in \mathcal{B}$ be fixed. Since we assume $\mathcal{B} = \sigma(s)$, there is a subset $\bar{A} \in \bar{\mathcal{B}}$ with $A = s^{-1}(\bar{A})$. Consequently, we have $s(A) = s\big(s^{-1}(\bar{A})\big) = \bar{A} \cap s(X) \in \bar{\mathcal{B}}|_{s(X)}$, where the transformation in the second step is true for all functions and subsets.

(ii)+(iii)⇒(i) Since $s\colon X \to \bar{X}$ is measurable, we have $\sigma(s) \subseteq \mathcal{B}$. For the converse inclusion, let $A \in \mathcal{B}$ be fixed. The injectivity of s ensures $A = s^{-1} \circ s(A)$ and the assumption $s(A) \in \bar{\mathcal{B}}|_{s(X)}$ gives us a subset $\bar{A} \in \bar{\mathcal{B}}$ with $s(A) = \bar{A} \cap s(X)$. Together, we get

$$A = s^{-1} \circ s(A) = s^{-1}\big(\bar{A} \cap s(X)\big) = s^{-1}(\bar{A}) \cap s^{-1}\big(s(X)\big) = s^{-1}(\bar{A}) \in \sigma(s) \ .$$

This proves $\sigma(s) = \mathcal{B}$.

(i)+(iv)⇒(iii) Let $x, x' \in \mathcal{B}$ with $s(x) = s(x')$. Since we have $\{x\} \in \mathcal{B} = \sigma(s)$, there is some $\bar{A} \in \bar{\mathcal{B}}$ with $\{x\} = s^{-1}(\bar{A})$. This gives us $s(x') = s(x) \in \bar{A}$ and hence $x' \in s^{-1}(\bar{A}) = \{x\}$ holds true. As a result, we find $x' = x$ and s is injective. □

Next, we provide some direct yet useful consequences for the transformed learning scenario using the LS loss.

2.1.4 Lemma (LS Regression) *Let P be a probability distribution on $X \times Y$, $s\colon X \to \bar{X}$ be a measurable function, and $\bar{P} = P \circ (s, \mathrm{id}_Y)^{-1}$ be the distribution of the transformed learning scenario. If $|P|_2 < \infty$ holds true then the LS Bayes functions satisfy ν-almost surely*

$$f^*_{\mathrm{LS},\bar{P}} \circ s = \mathbb{E}_\nu(f^*_{\mathrm{LS},P}|s) \enspace .$$

As a direct consequence, for distributions P with $|P|_2 < \infty$ and $s\colon X \to \bar{X}$, we can quantify the discrepancy of the LS Bayes risks $\mathcal{R}^*_{\mathrm{LS},\bar{P}}$ and $\mathcal{R}^*_{\mathrm{LS},P}$, namely from (1.5) and Point (iv) of Lemma 2.1.1 we get

$$\begin{aligned} \mathcal{R}^*_{\mathrm{LS},\bar{P}} - \mathcal{R}^*_{\mathrm{LS},P} &= \mathcal{R}_{\mathrm{LS},\bar{P}}(f^*_{\mathrm{LS},\bar{P}}) - \mathcal{R}^*_{\mathrm{LS},P} \\ &= \mathcal{R}_{\mathrm{LS},P}(f^*_{\mathrm{LS},\bar{P}} \circ s) - \mathcal{R}^*_{\mathrm{LS},P} \\ &= \left\| \mathbb{E}_\nu(f^*_{\mathrm{LS},P}|s) - f^*_{\mathrm{LS},P} \right\|^2_{L_2(\nu)} \enspace . \end{aligned} \tag{2.3}$$

Proof. Recall that $\pi_X\colon X \times Y \to X$ and $\pi_{\bar{X}}\colon \bar{X} \times Y \to \bar{X}$ denote the projections onto X and $\bar{X}$, respectively. Moreover, if there is no risk of confusion π_Y denotes the projection $\pi_Y\colon X \times Y \to Y$ and $\pi_Y\colon \bar{X} \times Y \to Y$. Using Point (iii) of Lemma 2.1.1 we find $|\bar{P}|_2 = |P|_2 < \infty$ and hence according to (1.4) the LS Bayes functions are given by $f^*_{\mathrm{LS},P}(x) = \mathbb{E}_P(\pi_Y|\pi_X = x)$ and $f^*_{\mathrm{LS},\bar{P}}(\bar{x}) = \mathbb{E}_{\bar{P}}(\pi_Y|\pi_{\bar{X}} = \bar{x})$, respectively.

Since $f^*_{\mathrm{LS},\bar{P}} \circ s$ is measurable with respect to the initial σ-algebra $\sigma(s)$ generated by s, it remains to prove, for $\bar{A} \in \bar{\mathcal{B}}$,

$$\int_{s^{-1}(\bar{A})} f^*_{\mathrm{LS},\bar{P}} \circ s \,\mathrm{d}\nu = \int_{s^{-1}(\bar{A})} f^*_{\mathrm{LS},P} \,\mathrm{d}\nu \enspace . \tag{2.4}$$

Using $\bar{\nu} = \nu \circ s^{-1}$ from Point (i) of Lemma 2.1.1, the change-of-variables formula, and Lemma 1.1.1 (for the transformed learning scenario) we find

$$\int_{s^{-1}(\bar{A})} f^*_{\mathrm{LS},\bar{P}} \circ s \,\mathrm{d}\nu = \int_{\bar{A}} f_{\mathrm{LS},\bar{P}} \,\mathrm{d}\bar{\nu} = \int_{\pi_{\bar{X}}^{-1}(\bar{A})} y \,\mathrm{d}\bar{P}(\bar{x}, y) \enspace .$$

Continuing this identity with the help of the change-of-variables formula for $\bar{P} = P \circ (s, \mathrm{id}_Y)^{-1}$ and $\pi_{\bar{X}} \circ (s, \mathrm{id}_Y) = s \circ \pi_X$ we get

$$\begin{aligned}\int_{s^{-1}(\bar{A})} f^*_{\mathrm{LS},\bar{P}} \circ s \,\mathrm{d}\nu &= \int_{(s,\mathrm{id}_Y)^{-1} \circ \pi_{\bar{X}}^{-1}(\bar{A})} y \,\mathrm{d}P(x,y) \\ &= \int_{\pi_X^{-1}(s^{-1}(\bar{A}))} y \,\mathrm{d}P(x,y) \ .\end{aligned}$$

Finally, an application of Lemma 1.1.1 with $A = s^{-1}(\bar{A})$ gives (2.4) and hence the assertion is proven. □

Before we continue, let us consider the following LS regression example.

2.1.5 Example (LS Regression) Let the original regression problem P on $[-\pi, \pi] \times \mathbb{R}$ be given by the marginal distribution $\nu = \mathrm{unif}\big([-\pi, \pi]\big)$ and the conditional distribution $P(\,\cdot\,|x) = \mathrm{unif}\big([\cos(x) - 1/2, \cos(x) + 1/2]\big)$. Furthermore, let $s\colon [-\pi, \pi] \to [0, \pi]$ be given by $s(x) := |x|$. Note that all the measurable spaces are with respect to the Borel σ-algebra. Then $f^*_{\mathrm{LS},P}(x) = \cos(x)$ is a LS Bayes function. Since $f^*_{\mathrm{LS},P} = f^*_{\mathrm{LS},P} \circ s$ holds true, the Bayes function $f^*_{\mathrm{LS},P}$ is even $\sigma(s)$-measurable. Using Point (ii) of Lemma 2.1.2 we find $\mathcal{R}^*_{\mathrm{LS},\bar{P}} = \mathcal{R}^*_{\mathrm{LS},P}$.

For the visualization of a data set $D \sim P^n$ and the corresponding transformed data set $s(D)$ see Figure 2.1. This example shows that if we know some symmetry properties, specified by the transformation s, of the Bayes function in advance then we do not lose any information, in terms of Bayes risk, if we use the transformed data set $s(D)$. In addition, Figure 2.1 shows that in this case—roughly speaking—$s(D)$ is twice as dense in $\bar{X} = [0, \pi]$ as D in $X = [-\pi, \pi]$. Consequently, using $s(D)$ instead of D can possibly even improve the performance of a learning algorithm.

Next, we transfer the results for the LS loss to the classification loss.

2.1.6 Lemma (Classification) *Let P be a probability distribution on $X \times Y$ with $Y = \{\pm 1\}$, $s\colon X \to \bar{X}$ be a measurable function, and $\bar{P} = P \circ (s, \mathrm{id}_Y)^{-1}$ be the distribution of the transformed learning scenario. Then the following statements are satisfied:*

(i) $\bar{p}_\pm = p_\pm$,

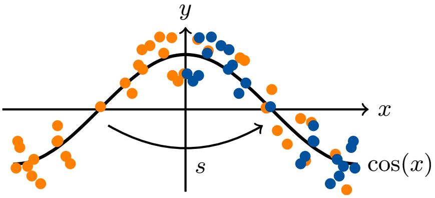


Figure 2.1: Illustration of the LS regression problem in Example 2.1.5. The data points of a data set $D \sim P^n$ are in orange, the data points of $s(D) \sim \bar{P}^n$, which are effected by s, are in blue, and the Bayes function is $f^*_{\mathrm{LS},P}(x) = \cos(x)$.

(ii) $\bar{\nu}_\pm = \nu_\pm \circ s^{-1}$, *and*

(iii) $\bar{\eta} \circ s = \mathbb{E}_\nu(\eta|s)$ ν*-almost surely.*

Point (iii) shows that the original learning scenario is less noisy than the transformed learning scenario.

Proof. (i) Since $\bar{P} = P \circ (s, \mathrm{id}_Y)^{-1}$ is given, we find $\bar{p}_\pm = \bar{P}(\bar{X} \times \{\pm 1\}) = P(X \times \{\pm 1\}) = p_\pm$.

(ii) From Point (i) and $\bar{P} = P \circ (s, \mathrm{id}_Y)^{-1}$ we get, for $\bar{A} \in \bar{\mathcal{B}}$,

$$\bar{\nu}_\pm(\bar{A}) = \frac{\bar{P}\big(\bar{A} \times \{\pm 1\}\big)}{\bar{p}_\pm} = \frac{P\big(s^{-1}(\bar{A}) \times \{\pm 1\}\big)}{p_\pm} = \nu_\pm \circ s^{-1}(\bar{A}) \quad .$$

(iii) Due to Lemma 1.2.1 and Lemma 2.1.4 we get

$$\bar{\eta} \circ s = \frac{f^*_{\mathrm{LS},\bar{P}} \circ s + 1}{2} = \frac{\mathbb{E}_\nu(f^*_{\mathrm{LS},P}|s) + 1}{2} = \mathbb{E}_\nu(\eta|s)$$

and hence the assertion is proven. □

Finally, we give an example of a transformed classification problem.

2.1.7 Example (Classification) Let the original classification problem P on $\mathbb{R}^2 \times \{\pm 1\}$ be given by the uniform distributions $\nu_+ = \mathrm{unif}\big([0,1] \times [-1/2, 1/2]\big)$, $\nu_- = \mathrm{unif}\big([-1,0] \times [-1/2, 1/2]\big)$, and $p_+ = p_- = 1/2$. Note that all measurable spaces in this example are with respect to the Borel σ-algebra. Since the measures ν_+ and ν_- are singular $\nu_+ \perp \nu_-$, the Bayes risk equals the minimal possible value, namely $\mathcal{R}^*_{\mathrm{Class},P} = 0$, see Lemma 1.2.2 for details. Now, we define two different transformed learning scenarios:

(i) Let $s := \pi_1 \colon \mathbb{R}^2 \to \mathbb{R}$ be the projection onto the first coordinate. The corresponding transformed learning scenario $\bar{P}$ is given by $\bar{\nu}_+ = \mathrm{unif}\big([0,1]\big)$ and $\bar{\nu}_- = \mathrm{unif}\big([-1,0]\big)$ (as well as $\bar{p}_+ = \bar{p}_- = 1/2$). Again, the measures $\bar{\nu}_+$ and $\bar{\nu}_-$ are singular $\bar{\nu}_+ \perp \bar{\nu}_-$ and hence the Bayes risk vanishes. Especially, $\mathcal{R}^*_{\mathrm{Class},\bar{P}} = \mathcal{R}^*_{\mathrm{Class},P} = 0$ is satisfied. Moreover, $\sigma(s) = \sigma(\pi_1) = \{A \times \mathbb{R} : \ A \in \mathcal{B}(\mathbb{R})\}$ is a proper subset of $\mathcal{B}(\mathbb{R}^2)$.

(ii) Let $s := \pi_2 \colon \mathbb{R}^2 \to \mathbb{R}$ be the projection onto the second coordinate. The corresponding transformed learning scenario $\bar{P}$ is given by $\bar{\nu}_+ = \bar{\nu}_- = \mathrm{unif}\big([-1/2, 1/2]\big)$ (as well as $\bar{p}_+ = \bar{p}_- = 1/2$). Consequently, we have $\bar{\eta} = 1/2$ and hence the Bayes risk equals the maximal possible value, namely $\mathcal{R}^*_{\mathrm{Class},\bar{P}} = 1/2$. Especially, we have the strict inequality $\mathcal{R}^*_{\mathrm{Class},\bar{P}} > \mathcal{R}^*_{\mathrm{Class},P}$ and $\sigma(s) = \sigma(\pi_2) = \{\mathbb{R} \times A : \ A \in \mathcal{B}(\mathbb{R})\}$ is again a proper subset of $\mathcal{B}(\mathbb{R}^2)$.

Point (i) of this example shows, that there are distributions P and transformations s such that the transformed scenario $\bar{P}$ satisfies $\mathcal{R}^*_{L,\bar{P}} = \mathcal{R}^*_{L,P}$ even if the generated σ-algebra is a proper subset $\sigma(s) \subsetneq \mathcal{B}$. Point (ii) shows, that there are distributions P and non-trivial transformations s such that no information is preserved in the transformed scenario $\bar{P}$. This means that in the transformed scenario, no learning method can do better than guessing. For the visualization of a data set $D \sim P^n$ and the corresponding transformed data sets $\pi_1(D)$ and $\pi_2(D)$ see Figure 2.2. Further interesting examples of transformed classification problems can be found in [25, Chapter 32].

2.2 Sequence of Transformed Scenarios

In this section we consider a whole sequence $(s_d)_{d \geq 1}$ of measurable functions $s_d \colon X \to X_d$ with measurable spaces $(X_d, \mathcal{B}_d)$ for $d \geq 1$. As a result, we

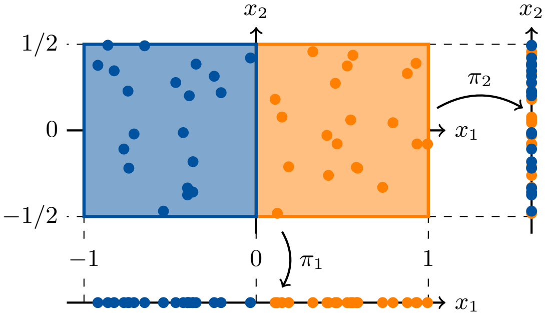


Figure 2.2: Illustration of the classification problem in Example 2.1.7. In the middle a data set $D \sim P^n$ is plotted, where positive and negative labeled data points are in orange and in blue, respectively. Below, the data set $\pi_1(D)$, where π_1 is the projection on the first coordinate, is plotted. On the right, the data set $\pi_2(D)$, where π_2 is the projection on the second coordinate, is plotted. This plot shows that using $\pi_1(D)$ it is an easy task to classify the data points correctly, but using $\pi_2(D)$ there is no method classifying the data points better than guessing.

have a sequence of transformed learning scenarios $P_d = P \circ (s_d, \mathrm{id}_Y)^{-1}$ on $X_d \times Y$. To prevent misunderstandings we use the index d for every quantity in the d-th transformed scenario instead of the bar-notation.

In this context a *learning method* $\bar{\mathcal{L}}$ on $(X_d \times Y)_{d \geq 1}$ is a sequence

$$\bar{\mathcal{L}} = (\mathcal{L}_{d,n})_{d,n \geq 1} \tag{2.5}$$

with two indexes such that $(\mathcal{L}_{d,n})_{n \geq 1}$ is, for every fixed $d \geq 1$, a learning method on $X_d \times Y$ in the sense of (1.1). Then we define the *pull-back learning method* $\mathcal{L}^{(d)} = (\mathcal{L}_n^{(d)})_{n \geq 1}$ on $X \times Y$ by

$$\mathcal{L}_n^{(d)}(D) \coloneqq \mathcal{L}_{d,n}\big(s_d(D)\big) \circ s_d \tag{2.6}$$

with the hyper parameter $d \geq 1$. The quality of $\bar{\mathcal{L}}$ is measured using risks of the pull-back learning method $\mathcal{L}^{(d)}$ in the original scenario, namely

$$\mathcal{R}_{L,P}\big(\mathcal{L}_n^{(d)}(D)\big) - \mathcal{R}^*_{L,P} \ .$$

We call the learning method $\bar{\mathcal{L}}$ *(potentially) L-risk consistent* for P and $(s_d)_{d\geq 1}$ if the corresponding pull-back learning method, with hyper parameter $d \geq 1$, is (potentially) L-risk consistent for P in the sense of (1.2). Analogously, we define the notion of learning rates for $\bar{\mathcal{L}}$. The following lemma provides sufficient conditions that ensure potential L-risk consistency.

2.2.1 Lemma (Potential Consistency) *Let P be a distribution on $X \times Y$, L be a supervised loss function, $(s_d)_{d\geq 1}$ be a sequence of measurable functions $s_d \colon X \to X_d$, and P_d be the corresponding transformed distribution on $X_d \times Y$ for $d \geq 1$. Furthermore, let $\bar{\mathcal{L}} = (\mathcal{L}_{d,n})_{d,n\geq 1}$ be a learning method on $(X_d \times Y)_{d\geq 1}$. If the following statements are true:*

(i) For fixed $d \geq 1$, the learning method $(\mathcal{L}_{d,n})_{n\geq 1}$ on $X_d \times Y$ is L-risk consistent for P_d.

(ii) $\inf_{d\geq 1} \mathcal{R}^*_{L,P_d} = \mathcal{R}^*_{L,P}$.

Then $\bar{\mathcal{L}}$ is potentially L-risk consistent for P and $(s_d)_{d\geq 1}$.

In order to prove consistency we need more information on the learning method $(\mathcal{L}_{d,n})_{d,n\geq 1}$, see e.g. Lemma 3.2.4 below for consistency results of histograms.

Proof. Using the abbreviations

$$\begin{aligned} E_1(d,D) &:= \mathcal{R}_{L,P}\big(\mathcal{L}_n^{(d)}(D)\big) - \mathcal{R}^*_{L,P_d} \qquad \text{and} \\ E_2(d) &:= \mathcal{R}^*_{L,P_d} - \mathcal{R}^*_{L,P} \end{aligned}$$

we split the excess risk into two parts

$$\mathcal{R}_{L,P}\big(\mathcal{L}_n^{(d)}(D)\big) - \mathcal{R}^*_{L,P} = E_1(d,D) + E_2(d) \ .$$

According to Point (iv) of Lemma 2.1.1 and the definition of $\mathcal{L}_n^{(d)}$ the first part $E_1(d,D)$ equals the excess risk of the learning method $(\mathcal{L}_{d,n})_{n\geq 1}$ on

$X_d \times Y$ using the data set $s_d(D)$, namely

$$E_1(d, D) = \mathcal{R}_{L,P_d}\big(\mathcal{L}_{d,n}(s_d(D))\big) - \mathcal{R}^*_{L,P_d} \ .$$

Consequently, $E_1(d, D)$ is a random variable, but $E_2(d)$ depends on d only and describes how well the d-th scenario approximates the original one.

According to our Assumption (ii), for every $\varepsilon > 0$, there is a $d(\varepsilon) \geq 1$ with $E_2(d(\varepsilon)) < \varepsilon/2$. Next, we define the functions

$$\begin{aligned} G(\varepsilon, d, n) &:= P^n\big(D \in (X \times Y)^n : \ E_1(d, D) + E_2(d) > \varepsilon\big) \qquad \text{and} \\ F(\varepsilon, n) &:= P^n\big(D \in (X \times Y)^n : \ E_1(d(\varepsilon), D) > \varepsilon/2\big) \ , \end{aligned}$$

for $n, d \geq 1$, and $\varepsilon > 0$. Note that $F(\varepsilon, n)$ depends on our choice $d(\varepsilon) \geq 1$. Using these functions we have to show that there is a sequence $(d_n)_{n\geq1}$ in $\mathbb{N}$ such that for every $\varepsilon > 0$ we have $G(\varepsilon, d_n, n) \to 0$ for $n \to \infty$. For a data set $D \in (X \times Y)^n$ with $E_1(d(\varepsilon), D) + E_2(d(\varepsilon)) > \varepsilon$ our choice of $d(\varepsilon)$ yields

$$\varepsilon < E_1\big(d(\varepsilon), D\big) + E_2\big(d(\varepsilon)\big) \leq E_1\big(d(\varepsilon), D\big) + \varepsilon/2 \ .$$

and hence $E_1(d(\varepsilon), D) > \varepsilon/2$. This proves, for all $n \geq 1$ and $\varepsilon > 0$,

$$G(\varepsilon, d(\varepsilon), n) \leq F(\varepsilon, n) \ .$$

Since $(\mathcal{L}_{d,n})_{n\geq1}$ is, for every fixed $d \geq 1$, an L-risk consistent learning method for P_d and

$$F(\varepsilon, n) = P_d^n\Big(\bar{D} \in (X_d \times Y)^n : \ \mathcal{R}_{L,P_d}(\mathcal{L}_{d,n}(\bar{D})) - \mathcal{R}^*_{L,P_d} > \varepsilon/2\Big) \ ,$$

we have $\lim_{n\to\infty} F(\varepsilon, n) = 0$ for all $\varepsilon > 0$. Consequently, we can apply [76, Lemma A.1.4] that gives us a sequence $(\varepsilon_n)_{n\geq1}$ with $\varepsilon_n \searrow 0$ and $F(\varepsilon_n, n) \to 0$ for $n \to \infty$.

Finally, we show that the sequence $d_n := d(\varepsilon_n)$ satisfies our requirements. To this end, let $\varepsilon, \delta > 0$ be fixed and choose $n_0 \geq 1$ such that $F(\varepsilon_n, n) < \delta$

and $\varepsilon_n < \varepsilon$ for all $n \geq n_0$. Since $G(\varepsilon, d, n)$ is non-increasing in ε, we find

$$G(\varepsilon, d_n, n) \leq G\big(\varepsilon_n, d(\varepsilon_n), n\big) \leq F(\varepsilon_n, n) < \delta$$

for all $n \geq n_0$ and hence the assertion is proven. □

The following lemma provides a condition that ensures Assumption (ii) of Lemma 2.2.1 in the case of the LS loss.

2.2.2 Lemma (LS Regression) *Let P be a distribution on $X \times Y$ with $|P|_2 < \infty$, $(s_d)_{d\geq 1}$ be a sequence of measurable functions $s_d \colon X \to X_d$, and P_d be the corresponding transformed distribution on $X_d \times Y$ for $d \geq 1$. If*

- *(i) $\sigma(s_d) \subseteq \sigma(s_{d+1})$ for all $d \geq 1$ and*
- *(ii) there is a $\sigma(s_d : \ d \geq 1)$-measurable Bayes function $f^*_{\mathrm{LS},P}$*

then the LS Bayes risks converge $\mathcal{R}^*_{\mathrm{LS},P_d} \searrow \mathcal{R}^*_{\mathrm{LS},P}$ for $d \to \infty$.

The interpretation of Assumption (i) is as follows: For increasing d the amount of information preserved by the transformation s_d is non-decreasing. Moreover, the Assumption (ii) is satisfied if e.g. $\mathcal{B} = \sigma(s_d : \ d \geq 1)$ holds.

Proof. The monotonicity is a direct consequence of Point (i) and the representation $\mathcal{R}^*_{\mathrm{LS},P_d} = \mathcal{R}^*_{\mathrm{LS},P,\mathcal{L}_0(X,\sigma(s_d))}$ from Lemma 2.1.2. Consequently, the limit for $d \to \infty$ exists and it remains to prove that the limit equals $\mathcal{R}^*_{\mathrm{LS},P}$. Since we assume $|P|_2 < \infty$, Lemma 2.1.4 yields

$$f^*_{\mathrm{LS},P_d} \circ s_d = \mathbb{E}_\nu\big(f^*_{\mathrm{LS},P} | \sigma(s_d)\big)$$

for all $d \geq 1$. Point (i) and $f^*_{\mathrm{LS},P} \in \mathcal{L}_2(\nu)$ ensures that $(f^*_{\mathrm{LS},P_d} \circ s_d)_{d\geq 1}$ is a martingale and hence a version of the martingale convergence theorem, see e.g. [11, Theorem 10.2.1], yields

$$f^*_{\mathrm{LS},P_d} \circ s_d \to \mathbb{E}_\nu\big(f^*_{\mathrm{LS},P} | \sigma(s_d : \ d \geq 1)\big)$$

in $L_2(\nu)$ for $d \to \infty$. Moreover, Assumption (ii) ensures that the limit equals $f^*_{\mathrm{LS},P}$ ν-almost surely. Finally, together with (2.3) we get

$$\mathcal{R}^*_{\mathrm{LS},P_d} - \mathcal{R}^*_{\mathrm{LS},P} = \left\| f^*_{\mathrm{LS},P_d} \circ s_d - f^*_{\mathrm{LS},P} \right\|^2_{L_2(\nu)} \to 0$$

for $d \to \infty$ and this finishes the proof. □

2.2.3 Corollary (Classification) *Under the assumptions of Lemma 2.2.2 the same statements hold true for $Y = \{\pm 1\}$ and the classification loss, i.e. $\mathcal{R}^*_{\mathrm{Class},P_d} \searrow \mathcal{R}^*_{\mathrm{Class},P}$ for $d \to \infty$.*

Note that the assumption $|P|_2 < \infty$ of Lemma 2.2.2 is automatically satisfied for classification problems since $Y = \{\pm 1\}$ is bounded. In [25, Theorem 32.3] another sufficient condition for the convergence $\mathcal{R}_{\mathrm{Class},P_d} \to \mathcal{R}^*_{\mathrm{Class},P}$ can be found.

Proof. This can be proved analogously to Lemma 2.2.2 using the calibration inequality

$$\mathcal{R}_{\mathrm{Class},P}(f) - \mathcal{R}^*_{\mathrm{Class},P} \leq \left(\mathcal{R}_{\mathrm{LS},P}(f) - \mathcal{R}^*_{\mathrm{LS},P} \right)^{1/2} ,$$

which holds true for every measurable function $f\colon X \to \mathbb{R}$, see e.g. [76, Example 3.23] for details. □

Finally, we present a situation in which the assumptions of Lemma 2.2.2 and hence also of Corollary 2.2.3 are satisfied. To this end, we define for a topological space T the set $C_b(T)$ of continuous and bounded functions $x\colon T \to \mathbb{R}$. Moreover, we equip $C_b(T)$ with the uniform norm $\|\cdot\|_{C_b(T)}$ and the corresponding Borel σ-algebra $\mathcal{B}(C_b(T))$.

2.2.4 Lemma (Point Evaluations) *Let T be a compact metric space and $X = C_b(T)$ with $\mathcal{B} = \mathcal{B}(X)$. Furthermore, let $(t_i)_{i \geq 1} \subseteq T$ be a countable dense subset of T and $s_d\colon X \to \mathbb{R}^d$ be given by*

$$s_d(x) := \big(x(t_1), \ldots, x(t_d) \big) .$$

Then $\sigma(s_d) \subseteq \sigma(s_{d+1})$ for $d \geq 1$ and $\mathcal{B} = \sigma(s_d :\ d \geq 1)$ hold true.

This lemma implies that Point (i) and (ii) of Lemma 2.2.2 are satisfied.

Since T is assumed to be a compact metric space, it is especially separable. As a result, a countable dense subset $(t_i)_{i\geq 1}$ always exists.

The assumption that T is a compact metric space seems very strict. But note that this assumption is equivalent to the separability of $C_b(T)$ for sufficient rich spaces $C_b(T)$. To be more precise, this equivalence is satisfied for completely regular topological spaces T, see e.g. [23, Theorem V.6.6] for details. Since the separability of $C_b(T)$ is a main ingredient of our proof, it is natural to assume that T is a compact metric space.

Proof. The monotonicity of $(\sigma(s_d))_{d\geq 1}$ is a direct consequence of the definition of s_d.

"$\supseteq$" Since the function s_d is continuous, we have $\sigma(s_d) \subseteq \mathcal{B}(X)$ for every $d \geq 1$ and hence the inclusion "$\supseteq$" follows.

"$\subseteq$" First, we show that the closed ball $B_X(x, r) \subseteq X$ with center $x \in X$ and radius $r > 0$ can be represented by

$$B_X(x,r) = \bigcap_{i\geq 1} \pi_i^{-1}\big([x(t_i) - r, x(t_i) + r]\big) =: A \ , \tag{2.7}$$

where $\pi_i \colon X \to \mathbb{R}$ denotes the point evaluation at t_i. For $x' \in B_X(x, r)$ we have $|x'(t_i) - x(t_i)| \leq \|x' - x\|_{C_b(T)} \leq r$ and hence $x' \in A$. Now, let us fix some $x' \in A$ and $t \in T$. Using the denseness of $(t_i)_{i\geq 1}$ there is a sequence $(i_j)_{j\geq 1}$ in $\mathbb{N}$ with $t_{i_j} \to t$ for $j \to \infty$. Since x and x' are continuous functions, we find

$$|x'(t) - x(t)| = \lim_{j\to\infty} \big|x'(t_{i_j}) - x(t_{i_j})\big| \leq r$$

and hence $\|x' - x\|_{C_b(T)} \leq r$. This shows $x' \in B_X(x, r)$ and (2.7) is proven.

Since the representation in (2.7) is a countable intersection of $\sigma(\pi_i)$-measurable sets, we find $B_X(x, r) \in \sigma(\pi_i : \ i \geq 1) = \sigma(s_d : \ d \geq 1)$. According to [23, Theorem V.6.6] the space $X = C_b(T)$ is separable and hence every open set $O \subseteq X$ is the countable union of closed balls. This shows $O \in \sigma(\pi_i : \ i \geq 1)$ for every open set $O \subseteq X$ and hence the inclusion $\mathcal{B}(X) \subseteq \sigma(\pi_i : \ i \geq 1)$ is proven. □

Finally, we present a corollary which summarizes Lemma 2.2.1–2.2.4.

2.2.5 Corollary (Potential Consistency of Point Evaluations) *Let T be a compact metric space, $X = C_b(T)$ with $\mathcal{B} = \mathcal{B}(X)$, P be a distribution on $X \times Y$ with $|P|_2 < \infty$, L be the LS loss or the classification loss (with $Y = \{\pm 1\}$). Furthermore, let $(t_i)_{i\geq 1} \subseteq T$ be a countable dense subset of T, $s_d \colon X \to \mathbb{R}^d$ be given by*

$$s_d(x) \coloneqq \big(x(t_1), \ldots, x(t_d)\big) \ ,$$

and $\bar{\mathcal{L}} = (\mathcal{L}_{d,n})_{d,n\geq 1}$ be a learning method on $(\mathbb{R}^d \times Y)_{d\geq 1}$. If, for all $d \geq 1$, the learning method $(\mathcal{L}_{d,n})_{n\geq 1}$ on $\mathbb{R}^d \times Y$ is L-risk consistent for $P_d \coloneqq P \circ (s_d, \mathrm{id}_Y)^{-1}$ then the learning method $\bar{\mathcal{L}}$ on $(X_d \times Y)_{d\geq 1}$ is potentially L-risk consistent for P and $(s_d)_{d\geq 1}$.

Proof. According to Lemma 2.2.4 we know that $\sigma(s_d) \subseteq \sigma(s_{d+1})$ for all $d \geq 1$ and $\mathcal{B} = \sigma(s_d : \ d \geq 1)$ is satisfied. Since L is the LS loss or the classification loss, Lemma 2.2.2 and Corollary 2.2.3, respectively, gives us $\inf_{d\geq 1} \mathcal{R}^*_{L,P_d} = \mathcal{R}^*_{L,P}$. Together with the assumption that $(\mathcal{L}_{d,n})_{n\geq 1}$ is L-risk consistent for P_d and every fixed $d \geq 1$, Lemma 2.2.1 yields the assertion. □

In Chapter 3 below we present another learning scenario satisfying the assumptions of Lemma 2.2.2 in detail.

2.3 Histograms

In this section we investigate histograms in the transformed learning scenario. To this end, let $s \colon X \to \bar{X}$ be a function. Then for a partition $\bar{\mathcal{A}} = (\bar{A}_k)_{k\in K}$ of $\bar{X}$ we define the *pull-back partition* $s^{-1}(\bar{\mathcal{A}}) \coloneqq (A_k)_{k\in K}$ of X by

$$A_k \coloneqq s^{-1}(\bar{A}_k) \ . \tag{2.8}$$

If $\bar{\mathcal{A}}$ and s are measurable then the pull-back partition $s^{-1}(\bar{\mathcal{A}})$ is measurable.

The next lemma relates the number of relevant cells of a partition and the corresponding pull-back partition.

2.3.1 Lemma (Relevant Cells) *Let $\bar{\mathcal{A}}$ be a partition of $\bar{X}$, $s\colon X \to \bar{X}$, and $\mathcal{A} := s^{-1}(\bar{\mathcal{A}})$ be the corresponding pull-back partition of X defined in* (2.8). *Then the following statements are true:*

(i) Let $\bar{\mathcal{A}}$ and s be measurable, ν be a distribution on X and $\bar{\nu} := \nu \circ s^{-1}$ be the push-forward measure on $\bar{X}$. Then $\mathcal{A}_\nu = \bar{\mathcal{A}}_{\bar{\nu}}$ is satisfied.

(ii) If $\bar{M} \subseteq \bar{X}$ and $M := s^{-1}(\bar{M}) \subseteq X$ then $\mathcal{A}_M \subseteq \bar{\mathcal{A}}_{\bar{M}}$ is satisfied. Moreover, if $\bar{M}$ is a subset $\bar{M} \subseteq s(X)$ of the image of s then even the equality holds true.

(iii) If $M \subseteq X$ then $\mathcal{A}_M = \mathcal{A}_{s^{-1}(s(M))} = \bar{\mathcal{A}}_{s(M)}$.

(iv) Let X and $\bar{X}$ be Polish spaces equipped with their Borel σ-algebras, s be continuous, ν be a distribution on X, and $\bar{\nu} := \nu \circ s^{-1}$ be the push-forward measure on $\bar{X}$. Then $\mathcal{A}_{\operatorname{supp}\nu} \subseteq \bar{\mathcal{A}}_{\operatorname{supp}\bar{\nu}}$ is satisfied. If, in addition, $s(\operatorname{supp}\nu) \subseteq \bar{X}$ is closed then even the equality holds true.

In Point (iv) the set $s(\operatorname{supp}\nu)$ is closed if e.g. $\operatorname{supp}\nu \subseteq X$ is compact.

Proof. We denote the cells of the considered partitions by $\bar{\mathcal{A}} = (\bar{A}_k)_{k\in K}$ and $\mathcal{A} = (A_k)_{k\in K}$ with $A_k = s^{-1}(\bar{A}_k)$ for $k \in K$.

(i) The definitions of the push-forward measure and the pull-back partition give us the assertion, namely $\bar{\nu}(\bar{A}_k) = \nu\big(s^{-1}(\bar{A}_k)\big) = \nu(A_k)$ implies $k \in \bar{\mathcal{A}}_{\bar{\nu}}$ if and only if $k \in \mathcal{A}_\nu$.

(ii) Note that we have

$$A_k \cap M = s^{-1}(\bar{A}_k) \cap s^{-1}(\bar{M}) = s^{-1}(\bar{A}_k \cap \bar{M}) \tag{2.9}$$

for all $k \in K$. Let $k \in \mathcal{A}_M$ be fixed. Then we have $A_k \cap M \neq \emptyset$ and hence (2.9) implies $\bar{A}_k \cap \bar{M} \neq 0$. This shows $k \in \bar{\mathcal{A}}_{\bar{M}}$ and hence the inclusion "$\subseteq$" is proven. For the converse inclusion we assume $\bar{M} \subseteq s(X)$ additionally. Let $k \in \bar{\mathcal{A}}_{\bar{M}}$ be fixed. Then we have $\bar{A}_k \cap \bar{M} \neq \emptyset$. Since $\bar{A}_k \cap \bar{M} \subseteq s(X)$ is a subset of the image of s, (2.9) implies $\emptyset \neq s^{-1}(\bar{A}_k \cap \bar{M}) = A_k \cap M$. This shows $k \in \mathcal{A}_M$ and hence the inclusion "$\supseteq$" is proven.

(iii) The second equality is a consequence of Point (ii) with $\bar{M} = s(M) \subseteq s(X)$ and it remains to prove the first equality. The inclusion "$\subseteq$" is a direct consequence of $M \subseteq s^{-1}(s(M))$. For the converse inclusion "$\supseteq$"

we have to use the special structure of the pull-back partition $\mathcal{A}$. Let $k \in \mathcal{A}_{s^{-1}(s(M))}$ and $x \in s^{-1}(s(M)) \cap A_k$. Then we have $s(x) \in \bar{A}_k$ and $s(x) \in s(M)$. Especially, there is some $y \in M$ with $s(y) = s(x) \in \bar{A}_k$. This proves $y \in M \cap A_k$ and hence $k \in \mathcal{A}_M$.

(iv) Using Point (iii) for $M = \operatorname{supp} \nu$ gives us $\mathcal{A}_{\operatorname{supp} \nu} = \bar{\mathcal{A}}_{s(\operatorname{supp} \nu)}$. Since $s(\operatorname{supp} \nu) \subseteq \overline{s(\operatorname{supp} \nu)} = \operatorname{supp} \bar{\nu}$ is satisfied according to Lemma B.2, the inclusion is proven. If, in addition, $s(\operatorname{supp} \nu)$ is closed then we even have $s(\operatorname{supp} \nu) = \operatorname{supp} \bar{\nu}$ and hence the equality is proven. □

The following lemma relates the histogram on X and on $\bar{X}$.

2.3.2 Lemma (Histograms) *Let P be a distribution on $X \times Y$, $s\colon X \to \bar{X}$ be a measurable function, and $\bar{P}$ be the corresponding transformed distribution on $\bar{X} \times Y$. Furthermore, let $\bar{\mathcal{A}}$ be a measurable and countable partition of $\bar{X}$ and $\mathcal{A} := s^{-1}(\bar{\mathcal{A}})$ be the corresponding pull-back partition defined in* (2.8)*. Then the histograms $h_{P,\mathcal{A}}$ and $h_{\bar{P},\bar{\mathcal{A}}}$ on X and $\bar{X}$, respectively, satisfy*

$$h_{\bar{P},\bar{\mathcal{A}}} \circ s = h_{P,\mathcal{A}} \quad .$$

According to Point (ii) of Lemma 2.1.1 the statement remains true for empirical histograms. To be more precise, for $D \in (X \times Y)^n$ and $\bar{D} := s(D) \in (\bar{X} \times Y)^n$ we have $h_{\bar{D},\bar{\mathcal{A}}} \circ s = h_{D,\mathcal{A}}$.

If $\bar{X} = X$ and $s\colon X \to \bar{X}$ is a bijective transformation with measurable inverse then Lemma 2.3.2 shows that transforming the data set D with s is equivalent to a transformation of the partition used by the histogram.

This lemma implies that the error in (2.2) for the histogram using the partition $\bar{\mathcal{A}}$ in the transformed scenario equals the excess risk of the histogram using the pull-back partition $\mathcal{A}$ in the original scenario, that is

$$\mathcal{R}_{L,P}(h_{\bar{D},\bar{\mathcal{A}}} \circ s) - \mathcal{R}^*_{L,P} = \mathcal{R}_{L,P}(h_{D,\mathcal{A}}) - \mathcal{R}^*_{L,P} \quad .$$

As a result, we can bound the error in (2.2) using oracle inequalities in the original scenario. We use this approach in Chapter 3 to prove consistency results and in Chapter 4 to prove learning rates for histograms.

Proof. We denote the cells of the considered partitions by $\bar{\mathcal{A}} = (\bar{A}_k)_{k\in K}$ and $\mathcal{A} = (A_k)_{k\in K}$ with $A_k = s^{-1}(\bar{A}_k)$ for $k \in K$. This gives $\mathbb{1}_{\bar{A}_k} \circ s = \mathbb{1}_{A_k}$ and $\bar{\nu}(\bar{A}_k) = \nu(A_k)$ for all $k \in K$. Moreover, the change-of-variables formula for $\bar{\nu} = \nu \circ s^{-1}$ together with Lemma 2.1.4 yields

$$\begin{aligned}\int_{\bar{A}_k} f^*_{\mathrm{LS},\bar{P}} \,\mathrm{d}\bar{\nu} &= \int_{s^{-1}(\bar{A}_k)} f^*_{\mathrm{LS},\bar{P}} \circ s \,\mathrm{d}\nu \\ &= \int_{s^{-1}(\bar{A}_k)} \mathbb{E}_\nu(f^*_{\mathrm{LS},P}|s) \,\mathrm{d}\nu \\ &= \int_{A_k} f^*_{\mathrm{LS},P} \,\mathrm{d}\nu \enspace .\end{aligned}$$

Since $\bar{\mathcal{A}}_{\bar{\nu}} = \mathcal{A}_\nu$ is satisfied according to Lemma 2.3.1, we get

$$\begin{aligned}h_{\bar{P},\bar{\mathcal{A}}} \circ s &= \sum_{k\in\bar{\mathcal{A}}_{\bar{\nu}}} \mathbb{1}_{\bar{A}_k} \circ s \,\frac{1}{\bar{\nu}(\bar{A}_k)} \int_{\bar{A}_k} f^*_{\mathrm{LS},\bar{P}} \,\mathrm{d}\bar{\nu} \\ &= \sum_{k\in\mathcal{A}_\nu} \mathbb{1}_{A_k} \frac{1}{\nu(A_k)} \int_{A_k} f^*_{\mathrm{LS},P} \,\mathrm{d}\nu \\ &= h_{P,\mathcal{A}}\end{aligned}$$

and hence the assertion is proven. □

As final part of this section we consider classification problems in the transformed learning scenario. For classification problems we can additionally investigate the margin-noise function. To this end, let us assume that $(\bar{X}, \bar{d})$ is a pseudo-metric space and $s\colon X \to \bar{X}$ is a measurable function. Then we define the *pull-back pseudo-metric* on X by

$$d(x, x') := \bar{d}\big(s(x), s(x')\big) \enspace . \tag{2.10}$$

Note that d is in general only a pseudo-metric even if $\bar{d}$ is a metric. Moreover, if X is already a topological space the pull-back pseudo-metric d on X defines in general a different topology on X. If $s\colon X \to (\bar{X}, \bar{d})$ is continuous then $\mathrm{id}\colon X \to (X, d)$ is continuous, i.e. the original topology on X is finer than the topology given by the pull-back pseudo-metric d. The next lemma provides bounds on the diameter of the pull-back partition as well as a representation of the distance to the decision boundary both with respect to the pull-back pseudo-metric.

2.3.3 Lemma (Pull-Back Pseudo-Metric) *Let $(\bar{X}, \bar{d})$ be a pseudo-metric space, $s\colon X \to \bar{X}$ be a (not necessarily measurable) function, and X be equipped with the corresponding pull-back pseudo-metric d defined in* (2.10)*. Then the following statements are true:*

(i) For a partition $\bar{\mathcal{A}}$ of $\bar{X}$ and the corresponding pull-back partition $\mathcal{A} := s^{-1}(\bar{\mathcal{A}})$ of X defined in (2.8) *the diameters satisfy* $\operatorname{diam}(\mathcal{A}) \leq \operatorname{diam}(\bar{\mathcal{A}})$*. Moreover, if s is surjective then equality holds true.*

(ii) For a distribution P on $X \times \{\pm 1\}$ the distance to the decision boundary defined in (1.13) *with respect to d is given by*

$$\Delta_d(x) = \begin{cases} \operatorname{dist}\big(s(x), s(X_+)\big), & x \in X_- \\ \operatorname{dist}\big(s(x), s(X_-)\big), & x \in X_+ \\ 0, & \textit{else.} \end{cases} \tag{2.11}$$

If the σ-algebra $\bar{\mathcal{B}}$ of $\bar{X}$ contains the Borel σ-algebra $\mathcal{B}(\bar{X}, \bar{d})$, that is $\bar{\mathcal{B}} \supseteq \mathcal{B}(\bar{X}, \bar{d})$, and $s\colon X \to \bar{X}$ is measurable then (2.11) gives the measurability of $\Delta_d\colon X \to \mathbb{R}$. This condition is slightly stronger than the condition $\mathcal{B} \supseteq \mathcal{B}(X, d)$ provided in (1.14) to ensure the measurability of Δ_d. To show this, we first prove

$$s^{-1}\big(\mathcal{B}(\bar{X}, \bar{d})\big) = \mathcal{B}(X, d) \ .$$

The inclusion "$\subseteq$" is a consequence of the continuity of $s\colon (X, d) \to (\bar{X}, \bar{d})$. The converse inclusion "$\supseteq$" follows from the fact that the pull-back pseudo-metric d induces the initial topology on X under $s\colon X \to (\bar{X}, \bar{d})$. Using this identity together with the measurability of s and $\bar{\mathcal{B}} \supseteq \mathcal{B}(\bar{X}, \bar{d})$ we recover the condition in (1.14), namely

$$\mathcal{B} \supseteq s^{-1}(\bar{\mathcal{B}}) \supseteq s^{-1}\big(\mathcal{B}(\bar{X}, \bar{d})\big) = \mathcal{B}(X, d) \ .$$

Proof. (i) We denote the cells of the considered partitions by $\bar{\mathcal{A}} = (\bar{A}_k)_{k\in K}$ and $\mathcal{A} = (A_k)_{k\in K}$ with $A_k = s^{-1}(\bar{A}_k)$ for $k \in K$. Let $k \in K$ be fixed. Then for all $x, x' \in A_k$ we have $s(x), s(x') \in \bar{A}_k$ and hence the definition of the pull-back pseudo-metric d gives us

$$\operatorname{diam}(A_k) = \sup_{x,x'\in A_k} \bar{d}\bigl(s(x), s(x')\bigr) \le \sup_{\bar{x},\bar{x}'\in \bar{A}_k} \bar{d}(\bar{x}, \bar{x}') = \operatorname{diam}(\bar{A}_k) \ .$$

This proves $\operatorname{diam}(\mathcal{A}) \le \operatorname{diam}(\bar{\mathcal{A}})$. Moreover, if s is surjective for all points $\bar{x}, \bar{x}' \in \bar{A}_k$ there are points $x, x' \in A_k$ with $s(x) = \bar{x}$ and $s(x') = \bar{x}'$ and hence in the above inequality we even get equality. This proves the second assertion.

(ii) In order to prevent misunderstandings we add the considered pseudo-metric as a subscript to the dist-function in the following proof. The definition of dist and the definition of d yield

$$\operatorname{dist}_d(x, A) = \inf_{x'\in A} \bar{d}\bigl(s(x), s(x')\bigr) = \operatorname{dist}_{\bar{d}}\bigl(s(x), s(A)\bigr)$$

for all subsets $A \subseteq X$. Together with the definition of Δ_d the assertion follows. □

The next lemma is a main ingredient for our learning rates proven in Chapter 4 below.

2.3.4 Lemma (Margin-Noise Function) *Let X and $\bar{X}$ be Polish spaces equipped with their Borel σ-algebras, $s\colon X \to \bar{X}$ be a continuous function, P be a probability distribution on $X \times \{\pm 1\}$, and $\bar{P} = P \circ (s, \mathrm{id}_Y)^{-1}$ be the distribution of the transformed learning scenario. Furthermore, let $\bar{d}$ be a metric on $\bar{X}$, which induces the topology on $\bar{X}$, and d be the corresponding pull-back pseudo-metric on X given by* (2.10)*, which does not necessarily induce the topology on X. Then the following statements are true:*

(i) If there is no noise in the original classification problem P then, for every version of η, the margin-noise function with respect to the

pull-back pseudo-metric d satisfies, for $r \geq 0$,

$$\begin{aligned} MN_d(r) = M_d(r) \geq p_+\bar{\nu}_+\big(\operatorname{dist}(\,\cdot\,,\operatorname{supp}\bar{\nu}_-) \leq 2r\big) \\ + p_-\bar{\nu}_-\big(\operatorname{dist}(\,\cdot\,,\operatorname{supp}\bar{\nu}_+) \leq 2r\big) \ . \end{aligned}$$

(ii) If $\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-$ *is a ν_+- or ν_--zero set then there is a version of η such that the equality holds true in (i).*

Note that the dist-function on the right hand side is with respect to the metric $\bar{d}$ on $\bar{X}$. Moreover, this result is especially applicable for $\bar{X} = X$ and $s = \mathrm{id}_X$.

Since $\operatorname{supp}\bar{\nu}_+ \cap \operatorname{supp}\bar{\nu}_-$ is a subset of $\{\operatorname{dist}(\,\cdot\,,\operatorname{supp}\bar{\nu}_\pm) \leq 2r\}$ for all $r \geq 0$, we get

$$MN_d(r) = M_d(r) \geq \bar{\nu}(\operatorname{supp}\bar{\nu}_+ \cap \operatorname{supp}\bar{\nu}_-) \ .$$

In addition, using $\operatorname{supp}\nu_\pm \subseteq s^{-1}(\operatorname{supp}\bar{\nu}_\pm)$ from (B.1) we find $s^{-1}(\operatorname{supp}\bar{\nu}_+ \cap \operatorname{supp}\bar{\nu}_-) = s^{-1}(\operatorname{supp}\bar{\nu}_+) \cap s^{-1}(\operatorname{supp}\bar{\nu}_-) \supseteq \operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-$ and hence we get the lower bound

$$MN_d(r) = M_d(r) \geq \nu(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-)$$

for all $r \geq 0$. Note that the right hand side depends only on the original learning problem.

Proof. First note that in this case the distance to the decision boundary $\Delta_d \colon X \to \mathbb{R}$ is a measurable function according to Lemma 2.3.3 and the remark after that lemma. The first equality $MN_d(r) = M_d(r)$, for $r \geq 0$, is a direct consequence of our assumption that there is no noise. In the following let $r \geq 0$ be fixed. Moreover, since $M_d(r) = p_+\nu_+(\Delta_d \leq 2r) + p_-\nu_-(\Delta_d \leq 2r)$ it is enough to consider $\nu_+(\Delta_d \leq 2r)$ for symmetry reasons.

(i) According to Lemma 1.2.2 the sets $X_\pm$ are sets of full measure with respect to $\nu_\pm$, respectively. Together with the representation in (2.11) and

$\bar{\nu}_+ = \nu_+ \circ s^{-1}$ we get

$$\begin{aligned} \nu_+(\Delta_d \leq 2r) &= \nu_+\big(X_+ \cap \{\Delta_d \leq 2r\}\big) \\ &= \nu_+\big(\mathrm{dist}(s(\,\cdot\,), s(X_-)) \leq 2r\big) \\ &= \bar{\nu}_+\big(\mathrm{dist}(\,\cdot\,, s(X_-)) \leq 2r\big) \quad . \end{aligned}$$

Since we have $\nu_-(X_-) = 1$, the support of ν_- is contained in the closure $\operatorname{supp}\nu_- \subseteq \overline{X_-}$. Moreover, the continuity of s implies $s\big(\overline{X_-}\big) \subseteq \overline{s(X_-)}$, see e.g. [32, Propositoin 1.4.1], and hence $\overline{s(\overline{X_-})} = \overline{s(X_-)}$. Together with Lemma B.2 we get

$$\operatorname{supp}\bar{\nu}_- = \overline{s(\operatorname{supp}\nu_-)} \subseteq \overline{s(\overline{X_-})} = \overline{s(X_-)} \quad .$$

As a result, for all $\bar{x} \in \bar{X}$ the distance can be bounded by

$$\mathrm{dist}\big(\bar{x}, s(X_-)\big) = \mathrm{dist}\big(\bar{x}, \overline{s(X_-)}\big) \leq \mathrm{dist}\big(\bar{x}, \operatorname{supp}\bar{\nu}_-\big) \quad .$$

Together, we get $\nu_+(\Delta_d \leq 2r) \geq \bar{\nu}_+\big(\mathrm{dist}(\,\cdot\,, \operatorname{supp}\bar{\nu}_-) \leq 2r\big)$ which proves the assertion.

(ii) According to Lemma 1.2.2 our assumption implies that there is no noise and hence Point (i) applies. Moreover, in Point (i) the inequality comes from the inclusion $\operatorname{supp}\nu_\pm \subseteq \overline{X_\pm}$ which we used in the proof. Consequently, we only need to find a version of η such that the equality $\operatorname{supp}\nu_\pm = \overline{X_\pm}$ is satisfied. To this end, we assume without loss of generality $\nu_-(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = 0$ and define

$$\eta(x) \coloneqq \begin{cases} 1, & x \in \operatorname{supp}\nu_+ \\ 0, & x \in \operatorname{supp}\nu_- \backslash (\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) \\ 1/2, & \text{else} \end{cases}$$

for $x \in X$. As a result, we have $X_+ = \operatorname{supp}\nu_+$ and

$$X_- = \operatorname{supp}\nu_- \backslash (\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) \subseteq \operatorname{supp}\nu_- \quad .$$

Consequently, we find $\operatorname{supp}\nu_\pm = \overline{X_\pm}$ and it remains to prove $\eta = p_+ \mathrm{d}\nu_+/\mathrm{d}\nu$ ν-almost surely. For $A \in \mathcal{B}$ we get

$$\begin{aligned}\int_A \eta \; \mathrm{d}\nu &= \nu(A \cap \operatorname{supp}\nu_+) \\ &= p_+\nu_+(A \cap \operatorname{supp}\nu_+) + p_-\nu_-\big(A \cap \operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-\big) \\ &= p_+\nu_+(A) \ ,\end{aligned}$$

where we used $\nu_-(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = 0$ in the last step. This shows $\eta = p_+ \mathrm{d}\nu_+/\mathrm{d}\nu$ ν-almost surely and hence the assertion is proven. □

Chapter 3

Learning in Sequence Spaces

In this chapter we consider learning problems using a sequence space as input space. For such a sequence space we can consider various projections onto finitely many coordinates (also called features). This defines a family of transformed learning scenarios, which describe the circumstance that we can only access finite-dimensional projections of an underlying infinite-dimensional learning problem.

3.1 Definitions and Basic Properties

In this chapter we consider learning scenarios on a *sequence space*

$$X := \prod_{i \geq 1} X_i \quad , \tag{3.1}$$

where $(X_i)_{i\geq 1}$ is a sequence of Polish spaces, i.e. complete separable metric spaces. For $i \geq 1$, we denote a corresponding complete metric on X_i by d_i. Moreover, we equip X with the product topology, i.e. convergence in X is equivalent to pointwise convergence.

Note that the separability of X_i ensures that it has a second-countable topology and hence the product σ-algebra and the Borel σ-algebra of the product topology on X coincide

$$\mathcal{B} := \bigotimes_{i \geq 1} \mathcal{B}(X_i) = \mathcal{B}(X)$$

see e.g. [11, Lemma 6.4.2] for details. Moreover, as a countable product of Polish spaces the space X itself is a Polish space, see e.g. [22, Proposition 8.1.4]. Recall from the discussion around (1.11) that for every probability measure ν on X the support $\operatorname{supp}\nu$ is a set of full measure.

For an index set $I \subseteq \mathbb{N}$ we denote the projection onto $X_I \coloneqq \prod_{i\in I} X_i$ by $\pi_I \colon X \to X_I$ where $\pi_I((x_i)_{i\geq 1}) \coloneqq (x_i)_{i\in I}$. Moreover, we define $\mathcal{F}(\mathbb{N})$ as the set of all finite subsets of $\mathbb{N}$ and $[d] \coloneqq \{1, 2, \dots, d\}$ as the set of the first $d \geq 1$ positive integers.

For every index set $I \in \mathcal{F}(\mathbb{N})$ we have a corresponding transformed learning scenario on $X_I \times Y$ as defined in (2.1) with $s = \pi_I$. To indicate on which coordinates we are projecting we use the index set I as subscript instead of the bar-notation, i.e. the distribution on $X_I \times Y$ is denoted by $P_I = P \circ (\pi_I, \mathrm{id}_Y)^{-1}$, its marginal distribution on X_I by $\nu_I = \nu \circ \pi_I^{-1}$, and the data set by $D_I = \pi_I(D) \in (X_I \times Y)^n$.

We interpret each coordinate x_i of a data point $x = (x_i)_{i\geq 1} \in X$ as a *feature* of x. In practice each feature is usually a measurement point. In this sense I specifies the features of x that are available in the transformed learning scenario P_I and hence I is called *feature set.* Consequently, the transformed learning scenario with a finite $I \subseteq \mathbb{N}$, i.e. $I \in \mathcal{F}(\mathbb{N})$, exactly describes the circumstance that we only have access to finitely many features of an infinite-dimensional data point.

In this context a *learning method* $\bar{\mathcal{L}}$ on $(X_I \times Y)_{I\in\mathcal{F}(\mathbb{N})}$ is a family

$$\bar{\mathcal{L}} = (\mathcal{L}_{I,n})_{I\in\mathcal{F}(\mathbb{N}),n\geq 1}$$

with two indexes such that $(\mathcal{L}_{I,n})_{n\geq 1}$ is, for every fixed $I \in \mathcal{F}(\mathbb{N})$, a learning method on $X_I \times Y$ in the sense of (1.1). Then we define the *pull-back learning method* $\mathcal{L}^{(I)} = (\mathcal{L}_n^{(I)})_{n\geq 1}$ on $X \times Y$ by

$$\mathcal{L}_n^{(I)}(D) \coloneqq \mathcal{L}_{I,n}\big(\pi_I(D)\big) \circ \pi_I \tag{3.2}$$

with the hyper parameter $I \in \mathcal{F}(\mathbb{N})$. In contrast to the pull-back learning method introduced in (2.6) the hyper parameter space is $\mathcal{F}(\mathbb{N})$ instead of $\mathbb{N}$. The quality of such a learning method is measured using risks of the pull-back learning method in the original scenario, namely

$$\mathcal{R}_{L,P}\big(\mathcal{L}_n^{(I)}(D)\big) - \mathcal{R}_{L,P}^* \ .$$

We call the learning method $\bar{\mathcal{L}}$ *(potentially) L-risk consistent* for P and $(\pi_I)_{I\in\mathcal{F}(\mathbb{N})}$ if the corresponding pull-back learning method, with hyper parameter $I \in \mathcal{F}(\mathbb{N})$, is (potentially) L-risk consistent for P in the sense of (1.2). Analogously, we define the notion of learning rates for $\bar{\mathcal{L}}$. Recall, potentially consistency means that there is some feature set sequence $(I_n)_{n\geq 1}$ such that the learning method $\mathcal{L}^{(I)}$ using $(I_n)_{n\geq 1}$ is consistent for P, i.e. $(\mathcal{L}_n^{(I_n)})_{n\geq 1}$ is consistent. In contrast, for consistency and learning rates we have to specify the used feature set sequence $(I_n)_{n\geq 1}$ exactly.

We interpret the sequence $(I_n)_{n\geq 1}$ as the amount of information, or in practice measurements, per data point that we need to guarantee consistency or learning rates, respectively. Here it is important to mention that if we have more information available, we can omit some features to reduce the amount of information. Consequently, the sequence $(I_n)_{n\geq 1}$ only describes a lower bound on the needed amount of information. To be more precise, if our application at hand provides us the features in $I'_n \in \mathcal{F}(\mathbb{N})$ with $I'_n \supseteq I_n$ for $n \geq 1$ then we—as a user—can decide to use only the features in I_n for learning. Formally, if we have some performance guaranties for $\bar{\mathcal{L}} = (\mathcal{L}_{I,n})_{I\in\mathcal{F}(\mathbb{N}),n\geq 1}$ using the feature set sequence $(I_n)_{n\geq 1}$ then we can consider the learning method $\bar{\mathcal{M}} = (\mathcal{M}_{I,n})_{I\in\mathcal{F}(\mathbb{N}),n\geq 1}$ on $(X_I \times Y)_{I\in\mathcal{F}(\mathbb{N})}$ given by

$$\mathcal{M}_{I,n}(D) \coloneqq \mathcal{L}_{I\cap I_n,n}\big(\pi_{I\cap I_n}(D)\big) \circ \pi_{I\cap I_n} \tag{3.3}$$

with the projection $\pi_{I\cap I_n} : X_I \to X_{I\cap I_n}$. In this case $\pi_{I\cap I_n}$ acts as a data independent feature selection method and $\bar{\mathcal{M}}$ is nothing more than the learning method $\bar{\mathcal{L}}$ combined with this specific feature selection method. Moreover, the corresponding pull-back learning method $\mathcal{M}^{(I)} = (\mathcal{M}_n^{(I)})_{n\geq 1}$ satisfies

$$\mathcal{M}_n^{(I)} = \mathcal{L}_n^{(I\cap I_n)}$$

for all $n \geq 1$. Since $I'_n \supseteq I_n$ this means that the learning method $\bar{\mathcal{M}}$ using $(I'_n)_{n\geq 1}$ equals $\bar{\mathcal{L}}$ using $(I_n)_{n\geq 1}$. As a result, every performance guaranty for a learning method $\bar{\mathcal{L}}$ using $(I_n)_{n\geq 1}$ also applies for any feature set sequence $(I'_n)_{n\geq 1}$ satisfying the lower bound $I'_n \supseteq I_n$ for all $n \geq 1$ if we allow the selection of a suitable subset of features for the actual learning. This is an advantage of our infinite-dimensional modeling of high-dimensional learning problems.

Being convinced that every condition on the feature set sequence is actually only a lower bound we briefly consider the important special case where $X_i = \mathbb{R}$ for all $i \geq 1$. In this case the size $d_n := |I_n|$ of the feature set $I_n \in \mathcal{F}(\mathbb{N})$ equals the dimension of the input space $X_{I_n} = \mathbb{R}^{d_n}$ in the corresponding transformed learning scenario and hence we can investigate the *curse of dimensionality*. This expression was coined by Bellman [6] in the field of dynamic programming. Nowadays, the term is popular in many fields, e.g. combinatorics, numerics, machine learning, etc., and—roughly speaking—refers to the general phenomena when the number of required data points depend exponentially on the dimension of the problem to ensure non-trivial performance guaranties. In our situation the curse of dimensionality can be specified as follows: A performance guaranty, e.g. a consistency result or a learning rate, for the learning method $\bar{\mathcal{L}} = (\mathcal{L}_{I,n})_{I\in\mathcal{F}(\mathbb{N}),n\geq 1}$ using feature sets sequences $(I_n)_{n\geq 1}$ suffers from the curse of dimensionality if there are some constants $c, \alpha > 0$ with

$$n \geq c \exp\bigl(\alpha |I_n|\bigr) \tag{3.4}$$

for all $n \geq 1$. Note that (3.4) is an upper growth bound on the sequence $(|I_n|)_{n\geq 1}$.

Assume for a moment, that we have a performance guaranty for $\bar{\mathcal{L}}$ using $(I_n)_{n\geq 1}$ that suffers from the curse of dimensionality. As in (3.3), we can combine $\bar{\mathcal{L}}$ with a feature selection method such that the same performance guaranty is satisfied using any feature set sequence $(I'_n)_{n\geq 1}$ with $I'_n \supseteq I_n$. This means, that $(|I'_n|)_{n\geq 1}$ only has to satisfy a lower growth bound. Especially, there is a feature set sequence $(I'_n)_{n\geq 1}$ that does not satisfy the upper growth bound in (3.4), but for which our performance guaranty applies. Consequently, in our setting the curse of dimensionality is not present if we allow the usage of a feature selection method. However, we do not claim that selecting the *right* features in practice is an easy task. Since the study of feature selection methods is its own research field, we focus on conditions for $(I_n)_{n\geq 1}$ that allows for consistency or learning rates and keep in mind that these performance guaranties do not suffer from the curse of dimensionality if we combine our learning method with a suitable feature selection method. We continue this discussion in Section 4.3 below in which we present polynomial learning rates.

Furthermore, note that from a practical point of view the usage of less information, i.e. smaller feature sets I_n, are also beneficial in terms of memory and time consumption. Before we go to the next section, the following lemma translates the potential consistency results from Section 2.2 into the situation of this chapter.

3.1.1 Corollary (Potential Consistency) *Let $X = \prod_{i\geq 1} X_i$ be given by (3.1), P be a distribution on $X \times Y$ with $|P|_2 < \infty$, L be the LS loss or the classification loss (with $Y = \{\pm 1\}$), and $\bar{\mathcal{L}} = (\mathcal{L}_{I,n})_{I\in\mathcal{F}(\mathbb{N}),n\geq 1}$ be a learning method on $(X_I \times Y)_{I\in\mathcal{F}(\mathbb{N})}$. If, for every fixed $I \in \mathcal{F}(\mathbb{N})$, the learning method $(\mathcal{L}_{I,n})_{n\geq 1}$ on $X_I \times Y$ is L-risk consistent for P_I then the learning method $\bar{\mathcal{L}}$ on $(X_I \times Y)_{I\in\mathcal{F}(\mathbb{N})}$ is potentially L-risk consistent for P and $(\pi_I)_{I\in\mathcal{F}(\mathbb{N})}$.*

Proof. Let $(I_d)_{d\geq 1}$ be an arbitrary sequence with $I_d \subseteq \mathbb{N}$, $I_d \subseteq I_{d+1}$, and $\bigcup_{d\geq 1} I_d = \mathbb{N}$, e.g. we can choose $I_d = [d]$. As a result, $\bar{\mathcal{M}} \coloneqq (\mathcal{M}_{d,n})_{d,n\geq 1}$ given by $\mathcal{M}_{d,n} \coloneqq \mathcal{L}_{I_d,n}$ defines a learning method in the sense of (2.5) on $(X_{I_d} \times Y)_{d\geq 1}$ with the transformations $s_d \coloneqq \pi_{I_d} \colon X \to X_{I_d}$. Since we assume $I_d \subseteq I_{d+1}$ for $d \geq 1$, the σ-algebras satisfy $\sigma(s_d) \subseteq \sigma(s_{d+1})$ as well. Moreover, the product σ-algebra is per definition the initial σ-algebra with respect to the projections $\pi_i \colon X \to X_i$ and hence

$$\mathcal{B} = \bigotimes_{i\geq 1} \mathcal{B}(X_i) = \sigma(\pi_i :\ i \geq 1) = \sigma(s_d :\ d \geq 1) \ .$$

As a consequence, $\bar{\mathcal{M}}$ satisfies the assumptions of Lemma 2.2.2 and Corollary 2.2.3, respectively. This ensures that the assumptions of Lemma 2.2.1 are satisfied which gives us a sequence $(d_n)_{n\geq 1}$ such that the pull-back learning method $(\mathcal{M}^{(d_n)})_{n\geq 1} = (\mathcal{L}^{(I_{d_n})})_{n\geq 1}$ is an L-risk consistent learning method for P. □

3.2 Histograms

Let $X = \prod_{i \geq 1} X_i$ be a sequence space as defined in (3.1). In order to apply histograms to a learning problem on $(X_I \times Y)_{I \in \mathcal{F}(\mathbb{N})}$ we need a partition of X_I for every $I \in \mathcal{F}(\mathbb{N})$. To this end, we introduce product partitions. We call a sequence $\mathcal{A} = (\mathcal{A}_i)_{i \geq 1}$ a *partition* of $(X_i)_{i \geq 1}$ if $\mathcal{A}_i = (A_{i,k})_{k \in K_i}$ is a partition of X_i for all $i \geq 1$. Then for every index set $I \in \mathcal{F}(\mathbb{N})$ the *product partition* $\mathcal{A}_I = (A_k)_{k \in K_I}$ of X_I is given by $K_I := \prod_{i \in I} K_i$ and

$$A_k := \prod_{i \in I} A_{i,k_i} \subseteq X_I \tag{3.5}$$

for $k = (k_i)_{i \in I} \in K_I$. It is easy to see that $\mathcal{A}_I$ is actually a partition of X_I.

We say a partition $\mathcal{A} = (\mathcal{A}_i)_{i \geq 1}$ of $(X_i)_{i \geq 1}$ is *measurable* or *countable* if $\mathcal{A}_i$ is a measurable or countable partition of X_i for all $i \geq 1$. Since $I \in \mathcal{F}(\mathbb{N})$ is finite, the product partition $\mathcal{A}_I$ is countable if $\mathcal{A}$ is countable. Analogously, since X is equipped with the product σ-algebra, the product partition $\mathcal{A}_I$ is measurable if $\mathcal{A}$ is measurable. Since X_i is a metric space for every $i \geq 1$, the diameter $\operatorname{diam}(\mathcal{A}_i)$ of $\mathcal{A}_i$ is defined in Section 1.3 and hence we define the *diameter* of $\mathcal{A}$ as $\operatorname{diam}(\mathcal{A}) := \sup_{i \geq 1} \operatorname{diam}(\mathcal{A}_i)$. But note that there can be multiple metrics on X_i inducing the topology of X_i and hence we have to specify a metric on X_i explicitly, for all $i \geq 1$, if we talk about the diameter $\operatorname{diam}(\mathcal{A})$. Moreover, for a distribution ν on X and $I \in \mathcal{F}(\mathbb{N})$ we use the abbreviation

$$\mathcal{A}_{I,\nu} := (\mathcal{A}_I)_{\nu_I}$$

for the indexes of the relevant cells of the product partition $\mathcal{A}_I$ with respect to the measure $\nu_I = \nu \circ \pi_I^{-1}$.

3.2.1 Lemma (Relevant Cells) *Let $X = \prod_{i \geq 1} X_i$ be a sequence space as defined in* (3.1)*, $\mathcal{A} = (\mathcal{A}_i)_{i \geq 1}$ be a measurable partition of $(X_i)_{i \geq 1}$, and ν be a distribution on X. Then, for $I \in \mathcal{F}(\mathbb{N})$, the relevant cells satisfy the inclusion*

$$\mathcal{A}_{I,\nu} \subseteq \prod_{i \in I} (\mathcal{A}_i)_{\nu_i} \quad ,$$

where $\nu_i := \nu \circ \pi_i^{-1}$ is the marginal distribution of ν on X_i for $i \geq 1$.

Proof. Let $I \in \mathcal{F}(\mathbb{N})$ and $k = (k_i)_{i \in I} \in \mathcal{A}_{I,\nu}$. With the notation from (3.5) this means

$$0 < \nu_I\Big(\prod_{i \in I} A_{i,k_i}\Big) \leq \nu_I\Big(A_{j,k_j} \times \prod_{i \neq j} X_i\Big) = \nu_j(A_{j,k_j})$$

for all $j \in I$. Consequently, $k_j \in (\mathcal{A}_j)_{\nu_j}$ for all $j \in I$ or equivalently $k \in \prod_{i \in I}(\mathcal{A}_i)_{\nu_i}$. □

For a measurable and countable partition $\mathcal{A}$ of $(X_i)_{i \geq 1}$ we consider the histogram learning method $\bar{\mathcal{L}}^{(\mathcal{A})} = (\mathcal{L}_{I,n}^{(\mathcal{A})})_{I \in \mathcal{F}(\mathbb{N}), n \geq 1}$ given by $\mathcal{L}_{I,n}^{(\mathcal{A})}(D_I) := h_{D_I,\mathcal{A}_I}$ for $D_I \in (X_I \times Y)^n$. Recall that the histogram $h_{D_I,\mathcal{A}_I} : X_I \to \mathbb{R}$ is defined in (1.17) and $\mathcal{A}_I$ denotes the product partition from (3.5). As an abbreviation of the corresponding pull-back learning method, defined in (3.2), we introduce the notation

$$h_{D,\mathcal{A},I} := h_{D_I,\mathcal{A}_I} \circ \pi_I : X \to \mathbb{R} \ .$$

Note that this pull-back learning method depends on two hyper parameters, the feature set I and the partition $\mathcal{A}$ of $(X_i)_{i \geq 1}$. Moreover, we use the same notation for the infinite-sample version $h_{P,\mathcal{A},I} := h_{P_I,\mathcal{A}_I} \circ \pi_I$.

Next, we present some properties of the approximation error. To this end, we need the following preparatory lemma.

3.2.2 Lemma (Lipschitz Continuous Functions) *Let X be a metric space, ν be a regular measure on the Borel σ-algebra $\mathcal{B}(X)$, and $f \in \mathcal{L}_2(\nu)$. Then, for every $\varepsilon > 0$, there is some Lipschitz continuous and bounded function $h : X \to \mathbb{R}$ with $\nu(h \neq 0) < \infty$ and*

$$\|f - h\|_{L_2(\nu)} \leq \varepsilon \ .$$

Note that for locally compact spaces X the set $C_c(X)$ of compactly supported continuous functions is dense in $\mathcal{L}_p(\nu)$, see e.g. [5, Theorem 29.14] or [22, Proposition 7.4.3] for details. This well-known approximation property of $C_c(X)$ is in many cases sufficient to investigate the approximation error. However, the assumption that X is locally compact excludes many non-compact infinite-dimensional spaces X and hence Lemma 3.2.2 is more suited for our purpose.

Proof. First, we approximate an indicator function $\mathbb{1}_M$ for some $M \in \mathcal{B}(X)$ with $\nu(M) < \infty$. To this end, let $\varepsilon > 0$ be fixed. Since ν is regular, there is a compact set $K \subseteq M$ and an open set $M \subseteq U$ with $\nu(U \backslash K) < \varepsilon^2$. Since K and U^c are disjoint and closed sets, we have $\operatorname{dist}(K, U^c) > 0$. Consequently, we can define

$$\varphi(x) := \min\left\{\frac{\operatorname{dist}(x, U^c)}{\operatorname{dist}(K, U^c)}, 1\right\}$$

for $x \in X$. The Lipschitz continuity of $x \mapsto \operatorname{dist}(x, U^c)$ implies the Lipschitz continuity of φ. Moreover, we directly get $0 \leq \varphi \leq 1$, $\varphi = 1$ on K, and $\varphi = 0$ on U^c. This implies $\nu(\varphi \neq 0) \leq \nu(U) \leq \nu(M) + \nu(U \backslash K) < \infty$ and

$$\|\mathbb{1}_M - \varphi\|_{L_2(\nu)}^2 = \int_U |\mathbb{1}_M - \varphi|^2 \, \mathrm{d}\nu \leq \nu(U \backslash K) < \varepsilon^2 \ .$$

As a result, the assertion is proven for indicator functions.

Now, let $f \in \mathcal{L}_2(\nu)$ and $\varepsilon > 0$ be fixed. According to [22, Proposition 3.4.2] there is a function $g = \sum_{i=1}^n a_i \mathbb{1}_{M_i}$ with $a_i \in \mathbb{R}$, $M_i \in \mathcal{B}(X)$, and $\nu(M_i) < \infty$ for all $i \geq 1$ such that $\|f - g\|_{L_2(\nu)} < \varepsilon/2$. Using the assertion for each function $\mathbb{1}_{M_i}$ gives us a Lipschitz continuous and bounded function φ_i with $\nu(\varphi_i \neq 0) < \infty$ and $\|\mathbb{1}_{M_i} - \varphi_i\|_{L_2(\nu)} < \varepsilon/(2a)$, where $a := |a_1| + \ldots + |a_n|$. Then the function $h := \sum_{i=1}^n a_i \varphi_i$ is Lipschitz continuous, bounded, and satisfies $\nu(h \neq 0) < \infty$. Moreover, we have

$$\|g - h\|_{L_2(\nu)} \leq \sum_{i=1}^n |a_i| \cdot \|\mathbb{1}_{M_i} - \varphi_i\|_{L_2(\nu)} < \varepsilon/2 \ .$$

All together we find $\|f - h\|_{L_2(\nu)} \leq \|f - g\|_{L_2(\nu)} + \|g - h\|_{L_2(\nu)} < \varepsilon$ and hence h has the desired properties. □

Now, we can prove a basic property of the LS approximation error.

3.2.3 Lemma (LS Approximation Error) *Let $X = \prod_{i\geq 1} X_i$ be a sequence space as defined in* (3.1) *and P be a probability distribution on $X \times Y$ with $|P|_2 < \infty$. Then, for every $\varepsilon > 0$, there is some $r_0 > 0$ and $d_0 > 0$ such that for all partitions $\mathcal{A}$ of $(X_i)_{i\geq 1}$ with* $\operatorname{diam}(\mathcal{A}) \leq 2r_0$ *and $d \geq d_0$ the histogram $h_{P,\mathcal{A},I}$ using the feature set $I = [d]$ satisfies the following inequality*

$$\mathcal{R}_{\mathrm{LS},P}(h_{P,\mathcal{A},I}) - \mathcal{R}^*_{\mathrm{LS},P} = \left\|h_{P,\mathcal{A},I} - f^*_{\mathrm{LS},P}\right\|^2_{L_2(\nu)} \leq \varepsilon$$

The following proof is an adaption of [66, Lemma E.4] to the infinite-dimensional setting.

Proof. Let $\varepsilon > 0$ be fixed, $I \in \mathcal{F}(\mathbb{N})$, $r > 0$, and $\mathcal{A}$ be a partition of $(X_i)_{i\geq 1}$ with $\operatorname{diam}(\mathcal{A}) \leq 2r$. Since the product partition $\mathcal{A}_I$ is a partition of X_I, we can consider the corresponding pull-back partition $\mathfrak{A} := \pi_I^{-1}(\mathcal{A}_I)$ of X given by (2.8). Together with Lemma 2.3.2 we find $h_{P,\mathcal{A},I} = h_{P_I,\mathcal{A}_I} \circ \pi_I = h_{P,\mathfrak{A}}$. According to (1.5) we have

$$\mathcal{R}_{\mathrm{LS},P}(h_{P,\mathcal{A},I}) - \mathcal{R}^*_{\mathrm{LS},P} = \left\|h_{P,\mathfrak{A}} - f^*_{\mathrm{LS},P}\right\|^2_{L_2(\nu)} \ . \tag{3.6}$$

Note that X is a Polish space and hence the marginal distribution ν is regular, see e.g. [31, Satz VIII.1.16]. As a result, Lemma 3.2.2 is applicable and gives us a Lipschitz continuous and bounded function $f\colon X \to \mathbb{R}$ with $\nu(f \neq 0) < \infty$ and

$$\left\|f - f^*_{\mathrm{LS},P}\right\|_{L_2(\nu)} \leq \sqrt{\varepsilon}/3 \ .$$

However, the notion of Lipschitz continuity depends on the metric and there are multiple (possibly non-equivalent) metrics on X inducing the product topology. We use Lemma 3.2.2 with the metric

$$d(x, x') := \sum_{i\geq 1} 2^{-i} \frac{d_i(x_i, x_i')}{1 + d_i(x_i, x_i')} \tag{3.7}$$

for $x = (x_i)_{i\geq 1}, x' = (x_i')_{i\geq 1} \in X$ which is well-known to induce the product topology on X. Now, we define a probability measure Q on $X \times Y$ by defining the marginal distribution $Q \circ \pi_X^{-1} := \nu$ and the conditional distribution

$Q(\,\cdot\,|x) := \delta_{f(x)}$ for $x \in X$. For Q the LS Bayes function equals $f^*_{\mathrm{LS},Q} = f$. Using $h_{P,\mathfrak{A}} = \mathbb{E}_\nu(f^*_{\mathrm{LS},P}|\sigma(\mathfrak{A}))$ from (1.18), $Q \circ \pi_X^{-1} = P \circ \pi_X^{-1} = \nu$, and the projection property of conditional expectations we find

$$\left\|h_{Q,\mathfrak{A}} - h_{P,\mathfrak{A}}\right\|_{L_2(\nu)} = \left\|\mathbb{E}_\nu(f^*_{\mathrm{LS},Q} - f^*_{\mathrm{LS},P}|\sigma(\mathfrak{A}))\right\|_{L_2(\nu)} \leq \left\|f - f^*_{\mathrm{LS},P}\right\|_{L_2(\nu)} \ .$$

This results in

$$\begin{aligned} \left\|h_{P,\mathfrak{A}} - f^*_{\mathrm{LS},P}\right\|_{L_2(\nu)} &\leq \left\|h_{P,\mathfrak{A}} - h_{Q,\mathfrak{A}}\right\|_{L_2(\nu)} \\ &\quad + \left\|h_{Q,\mathfrak{A}} - f\right\|_{L_2(\nu)} \\ &\quad + \left\|f - f^*_{\mathrm{LS},P}\right\|_{L_2(\nu)} \\ &\leq \ 2\sqrt{\varepsilon}/3 + \left\|h_{Q,\mathfrak{A}} - f\right\|_{L_2(\nu)} \end{aligned} \tag{3.8}$$

and hence it remains to bound $\|h_{Q,\mathfrak{A}} - f\|_{L_2(\nu)}$. To this end, let A be a cell of $\mathfrak{A}$ with $\nu(A) > 0$ and $x \in A$. Using (1.18) and the Lipschitz continuity of f we find

$$\left|h_{Q,\mathfrak{A}}(x) - f(x)\right| \leq \frac{1}{\nu(A)} \int_A |f(x') - f(x)| \ \mathrm{d}\nu(x') \leq L \cdot \mathrm{diam}(A) \ ,$$

where L denotes the Lipschitz constant and $\mathrm{diam}(A)$ the diameter with respect to the metric d in (3.7). Since the cells A of $\mathfrak{A}$ with $\nu(A) = 0$ are ignored by the $L_\infty(\nu)$-norm, we find

$$\left\|h_{Q,\mathfrak{A}} - f\right\|^2_{L_2(\nu)} \leq \left\|h_{Q,\mathfrak{A}} - f\right\|^2_{L_\infty(\nu)} \leq L \cdot \mathrm{diam}(\mathfrak{A}) \tag{3.9}$$

and we need to bound the diameter of the pull-back partition $\mathfrak{A}$. Note that the diameter of $\mathfrak{A}$ with respect to the pull-back metric, defined in (2.10), is already given by Lemma 2.3.3 but this metric does not induce the product topology on X and hence this result cannot be used for this proof. Again, let A be a cell of $\mathfrak{A}$ and $x = (x_i)_{i\geq 1}, x' = (x'_i)_{i\geq 1} \in A$. Since $\mathrm{diam}(\mathcal{A}) \leq 2r$ and $\mathfrak{A} = \pi_I^{-1}(\mathcal{A}_I)$ hold true, we have $d_i(x_i, x'_i) \leq 2r$ for $i \in I$. Together with $I = [d]$ we find

$$d(x, x') \le \frac{2r}{1+2r} \sum_{i=1}^{d} 2^{-i} + \sum_{i>d} 2^{-i} \le \frac{2r}{1+2r} + 2^{-d} \le 2r + 2^{-d} \quad .$$

As a result, the diameter with respect to d is bounded by $\operatorname{diam}(\mathfrak{A}) \le 2r + 2^{-d}$. Together with (3.9) we get $\|h_{Q,\mathfrak{A}} - f\|_{L_2(\nu)} \le L(2r + 2^{-d})$.

Finally, we choose $r_0 := \sqrt{\varepsilon}/(12L)$ and $d_0 := \lceil \log(6L/\sqrt{\varepsilon})/\log(2) \rceil$. Then for $0 < r \le r_0$ and $d \ge d_0$ we have $L(2r + 2^{-d}) \le \sqrt{\varepsilon}/3$ and together with (3.6) and (3.8) we get the assertion. □

For us, the most important case is $X_i = (\mathbb{R}^{p_i}, \|\cdot\|_{\ell_\infty^{p_i}})$ with $p_i \ge 1$, where

$$\|x\|_{\ell_\infty^{p_i}} := \sup_{j=1,\dots,p_i} |x_j|$$

for $x = (x_j)_{j=1}^{p_i} \in \mathbb{R}^{p_i}$. In this case we call a partition $\mathcal{A} = (\mathcal{A}_i)_{i\ge1}$ of $(X_i)_{i\ge1}$ *cubic* with radius $r > 0$ if $\mathcal{A}_i$ is a cubic partition, as defined in Section 1.3, with radius r for all $i \ge 1$. For $I \in \mathcal{F}(\mathbb{N})$, the product partition $\mathcal{A}_I$ of a cubic partition is again a cubic partition of $X_I \cong \mathbb{R}^p$ with the same radius and $p = \sum_{i\in I} p_i$. Moreover, we introduce the abbreviation

$$h_{D,r,I} := h_{D,\mathcal{A},I} \quad ,$$

where $\mathcal{A}$ is some predefined cubic partition of $(X_i)_{i\ge1}$ with radius $r > 0$ and $I \in \mathcal{F}(\mathbb{N})$. In addition, we use an analogous notation for the population version $h_{P,r,I} := h_{P,\mathcal{A},I}$.

The following lemma establishes consistency for histograms using cubic partitions and finitely many features.

3.2.4 Lemma (LS-Risk Consistency) *Let $X = \mathbb{R}^{\mathbb{N}}$ be a sequence space as defined in* (3.1) *with $X_i = \mathbb{R}$ for all $i \ge 1$ and P be a probability distribution on $X \times Y$ with $Y = [-M, M]$ for some $M > 0$ and $P\big([-b, b]^{\mathbb{N}} \times Y\big) = 1$ for some $b > 0$. Furthermore, let $(r_n)_{n\ge1}$ and $(d_n)_{n\ge1}$ be sequences in $(0, b]$*

and $\mathbb{N}$*, respectively, with* $r_n \to 0$*,* $d_n \to \infty$*, and*

$$\frac{(3b)^{d_n}}{nr_n^{d_n}} \to 0 \tag{3.10}$$

for $n \to \infty$*. Then the histogram learning method* $D \mapsto h_{D,r_n,I_n}$ *using the feature set* $I_n := [d_n]$ *is LS-risk consistent for* P*.*

This is a generalization of the well-known consistency of cubic histograms from finite- to infinite-dimensional spaces, cf. [43, Theorem 4.2].

Note that scaling the data influences the condition in (3.10). To be more precise, if we scale the data by a factor of $a > 0$, the data is concentrated on $[-ab, ab]^{\mathbb{N}} \times Y$ and we need to scale the sequence $(r_n)_{n\geq 1}$ by the same factor to ensure that (3.10) is still satisfied.

Since $r_n \leq b$ holds true, the condition in (3.10) implies $n \geq c \cdot 3^{d_n}$ for all $n \geq 1$ with some constant $c > 0$. At first sight, it seems that this consistency result suffers from the curse of dimensionality defined in (3.4). However, recall from the discussion around (3.4) that if we allow the combination with a feature selection method the curse of dimensionality is no longer present.

Proof. Let $n \geq 1$ and $\mathcal{A}$ be the predefined cubic partition of $(X_i)_{i\geq 1}$ with radius r_n. Moreover, we define the sequence

$$a_n := \frac{(3b)^{d_n}}{nr_n^{d_n}}$$

which converges to zero according to our assumption. Since $\mathcal{A}_I$ is a partition of $\mathbb{R}^I$, we can consider the corresponding pull-back partition $\mathfrak{A} := \pi_I^{-1}(\mathcal{A}_I)$ of $X = \mathbb{R}^{\mathbb{N}}$ given by (2.8). Using Lemma 2.3.2 the excess LS-risk equals

$$\mathcal{R}_{\mathrm{LS},P}(h_{D,r_n,I_n}) - \mathcal{R}^*_{\mathrm{LS},P} = \mathcal{R}_{\mathrm{LS},P}(h_{D,\mathfrak{A}}) - \mathcal{R}^*_{\mathrm{LS},P} \ .$$

If we apply Lemma 1.3.1 to the right hand side for $\tau = \tau_n := a_n^{-1/3}$ and

$\varepsilon = \varepsilon_n \coloneqq 3M \exp(-a_n^{-1/3})$, we get

$$\begin{aligned}\mathcal{R}_{\mathrm{LS},P}(h_{D,r_n,I_n}) - \mathcal{R}^*_{\mathrm{LS},P} \leq 4\big\|h_{P,\mathfrak{A}} - f^*_{\mathrm{LS},P}\big\|^2_{L_2(\nu)}& \\ + 20M\varepsilon_n& \\ + 1536M^2 \log(3M/\varepsilon_n) \cdot \frac{\tau_n|\mathfrak{A}_\nu|}{n}&\end{aligned}$$

with probability P^n not less than $1 - e^{-\tau_n}$. Note that our choice $\tau_n = a_n^{-1/3}$ ensures that $1 - e^{-\tau_n} \to 1$ for $n \to \infty$. Consequently, it remains to prove that all the terms on the right hand side vanish for $n \to \infty$.

The assumption $a_n \to 0$ for $n \to \infty$ implies $40M\varepsilon_n \to 0$ for $n \to \infty$. Since $r_n \to 0$ and $d_n \to \infty$ for $n \to \infty$, Lemma 3.2.3 yields, for $n \to \infty$,

$$\big\|h_{P,\mathfrak{A}} - f^*_{\mathrm{LS},P}\big\|^2_{L_2(\nu)} = \big\|h_{P,r_n,I_n} - f^*_{\mathrm{LS},P}\big\|^2_{L_2(\nu)} \to 0 \ \ .$$

Finally, we consider the number of relevant cells. Using Point (i) of Lemma 2.3.1, Lemma 3.2.1, and Lemma 1.3.4 with $M = [-b, b]$ we find

$$|\mathfrak{A}_\nu| = \big|(\mathcal{A}_I)_{\nu_I}\big| \leq \prod_{i\in I}\big|(\mathcal{A}_i)_{\nu_i}\big| \leq \prod_{i\in I}\big|(\mathcal{A}_i)_{[-b,b]}\big| \ \ .$$

Consequently, it remains to bound $|(\mathcal{A}_i)_{[-b,b]}|$. To this end, let A_k be the cell that contains the left corner of the interval $[-b, b]$. Since all cells are aligned, the remaining interval $[-b, b]\backslash A_k$ of length $L \leq 2b$ is covered by $\lceil L/(2r)\rceil$ cells. Together with $r_n \leq b$ this yields

$$\big|(\mathcal{A}_i)_{[-b,b]}\big| \leq 1 + \lceil b/r_n\rceil \leq 2 + b/r_n \leq 3b/r_n \ \ .$$

Putting both together and using $\log(3M/\varepsilon_n) = a_n^{-1/3}$ we get

$$1536M^2 \log(3M/\varepsilon) \cdot \frac{\tau|\mathfrak{A}_\nu|}{n} \preccurlyeq a_n^{-1/3}\frac{\tau_n(3b/r_n)^{d_n}}{n} = a_n^{1/3} \to 0$$

for $n \to \infty$. As a result, the consistency of $D \mapsto h_{D,r_n,I_n}$ is proven. □

The next corollary transfers the LS-risk consistency of the previous lemma to classification-risk consistency.

3.2.5 Corollary (Classification-Risk Consistency) *Let the assumptions of Lemma 3.2.4 with $Y = \{\pm 1\}$ be satisfied. Then the histogram learning method $D \mapsto h_{D,r_n,I_n}$ is classification-risk consistent for P.*

This is a generalization of the well-known consistency for cubic histograms from finite- to infinite-dimensional spaces, cf. [25, Theorem 6.2].

Proof. This is a direct consequence of Lemma 3.2.4 and the calibration inequality

$$\mathcal{R}_{\mathrm{Class},P}(f) - \mathcal{R}^*_{\mathrm{Class},P} \leq \bigl(\mathcal{R}_{\mathrm{LS},P}(f) - \mathcal{R}^*_{\mathrm{LS},P}\bigr)^{1/2} \ ,$$

which holds true for every measurable function $f\colon X \to \mathbb{R}$, see e.g. [76, Example 3.23] for details. □

As final part of this section we consider a classification problem P on $X \times \{\pm 1\}$, where $X = \prod_{i \geq 1} X_i$ is a sequence space. Since (X_i, d_i) is a metric space for all $i \geq 1$, we can turn X_I, for $I \in \mathcal{F}(\mathbb{N})$, into a metric space using the metric

$$d_I\bigl((x_i)_{i\in I}, (x'_i)_{i\in I}\bigr) \coloneqq \sup_{i \in I} d_i(x_i, x'_i) \ . \tag{3.11}$$

For a partition $\mathcal{A} = (\mathcal{A}_i)_{i\geq 1}$ of $(X_i)_{i \geq 1}$, the diameter of the product partition $\mathcal{A}_I$ with respect of d_I satisfies $\mathrm{diam}(\mathcal{A}_I) = \sup_{i\in I} \mathrm{diam}(\mathcal{A}_i) \leq \mathrm{diam}(\mathcal{A})$. Moreover, for the corresponding pull-back pseudo-metric

$$d(x, x') \coloneqq d_I\bigl(\pi_I(x), \pi_I(x')\bigr) \tag{3.12}$$

on X, introduced in (2.10), we denote the distance to the decision boundary by $\Delta_I \coloneqq \Delta_d$, see (1.13) for the definition of Δ_d. In this case the representation in (2.11) reads

$$\Delta_I(x) = \begin{cases} \mathrm{dist}\bigl(\pi_I(x), \pi_I(X_+)\bigr), & x \in X_- \\ \mathrm{dist}\bigl(\pi_I(x), \pi_I(X_-)\bigr), & x \in X_+ \\ 0, & \text{else,} \end{cases} \tag{3.13}$$

where $X_\pm$ are defined in (1.8). Moreover, we write MN_I for the margin-noise function and M_I for the margin function with respect to Δ_I, see (1.15) for their definitions. The following lemma provides some monotonicity properties of Δ_I, MN_I, and M_I in I.

3.2.6 Lemma (Monotonicity) *Let $X = \prod_{i\geq 1} X_i$ be a sequence space as defined in* (3.1) *and P be a distribution on $X \times \{\pm 1\}$. Then the following statements are true, for finite $I \subseteq J \subseteq \mathbb{N}$:*

(i) $\Delta_I(x) \leq \Delta_J(x)$ *for all* $x \in X$*,*

(ii) $M_I(r) \geq M_J(r)$ *for all* $r \geq 0$*, and*

(iii) $MN_I(r) \geq MN_J(r)$ *for all* $r \geq 0$*.*

Since the margin-noise function bounds the approximation error, see (A.4) in the proof of Lemma 1.3.2, Point (iii) underpins our intuition: The larger the index set I, the more information of the original scenario is preserved in the I-th transformed scenario, and the smaller the approximation error bound.

Proof. Point (ii) and (iii) are direct consequences of Point (i). To prove Point (i) we fix some $x = (x_i)_{i\geq 1} \in X$. If $x \in X_+$ then for an arbitrary $x_- = (x_{i,-})_{i\geq 1} \in X_-$ we have

$$\Delta_I(x) \leq d_I\big(\pi_I(x), \pi_I(x_-)\big) = \sup_{i\in I} d_i(x_i, x_{i,-}) \leq d_J\big(\pi_J(x), \pi_J(x_-)\big) \ .$$

Taking the infimum over $x_- \in X_-$ yields $\Delta_I(x) \leq \Delta_J(x)$. For $x \in X_-$ an analogous argument yields $\Delta_I(x) \leq \Delta_J(x)$. Finally, we have $\Delta_I(x) = 0 = \Delta_J(x)$ in all remaining cases. □

Finally, we apply the oracle inequality of Lemma 1.3.2 in the case of a sequence space X. Since this is the basis of all the investigations of Chapter 4, we formulate it as a corollary.

3.2.7 Corollary (Oracle Inequality for Histograms on Sequence Spaces) *Let $X = \prod_{i\geq 1} X_i$ be a sequence space as defined in* (3.1)*, P be a distribution on $X \times \{\pm 1\}$ with noise exponent $0 \leq q \leq \infty$ in the sense of* (1.12)*, $r > 0$,* *and $\mathcal{A}$ be a measurable partition of $(X_i)_{i\geq 1}$ with* $\operatorname{diam}(\mathcal{A}) \leq 2r$*. Then the histogram using $\mathcal{A}$ and the feature set $I \in \mathcal{F}(\mathbb{N})$ satisfies, for $\tau \geq 1$ and $n \geq 1$,*

$$\mathcal{R}_{\text{Class},P}(h_{D,\mathcal{A},I}) - \mathcal{R}^*_{\text{Class},P} \leq 6MN_I(r) + C\left(\frac{\tau|\mathcal{A}_{I,\nu}|}{n}\right)^{\frac{q+1}{q+2}}$$

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

Proof. Let $I \in \mathcal{F}(\mathbb{N})$, $n \geq 1$, $\tau \geq 1$, and $D \in (X \times Y)^n$. Moreover, we denote the product partition defined in (3.5) by $\mathcal{A}_I$ and the corresponding pull-back partition of X defined in (2.8) by $\mathfrak{A} := \pi_I^{-1}(\mathcal{A}_I)$. From Lemma 2.3.2 we get

$$\mathcal{R}_{\text{Class},P}(h_{D,\mathcal{A},I}) = \mathcal{R}_{\text{Class},P}\big(h_{D_I,\mathcal{A}_I} \circ \pi_I\big) = \mathcal{R}_{\text{Class},P}(h_{D,\mathfrak{A}})$$

According to Point (i) of Lemma 2.3.3 the diameter with respect to the pull-back pseudo-metric, given in (3.12), is bounded by $\operatorname{diam}(\mathfrak{A}) = \operatorname{diam}(\mathcal{A}_I) \leq \operatorname{diam}(\mathcal{A}) \leq 2r$. Consequently, using Lemma 1.3.2 for $\mathfrak{A}$ and the pull-back pseudo-metric gives us

$$\mathcal{R}_{\text{Class},P}(h_{D,\mathcal{A},I}) - \mathcal{R}^*_{\text{Class},P} < 6MN_I(r) + C\left(\frac{\tau|\mathfrak{A}_{\nu}|}{n}\right)^{\frac{q+1}{q+2}}$$

with probability P^n not less than $1 - e^{-\tau}$. Using Lemma 2.3.1 we get $|\mathfrak{A}_\nu| = |(\mathcal{A}_I)_{\nu_I}| = |\mathcal{A}_{I,\nu}|$ and hence the assertion is proven. □

3.3 Product Distributions

In this section we consider the margin-noise function and the number of relevant cells for the following type of classification problems on sequence spaces.

3.3.1 Assumption Let $X = \prod_{i\geq 1} X_i$ be a sequence space as defined in (3.1) and P be a distribution on $X \times \{\pm 1\}$, where the marginals $\nu_\pm$, defined in

(1.6), are of the form

$$\nu_+ = \bigotimes_{i\geq 1} \nu_{+,i} \qquad \text{and} \qquad \nu_- = \bigotimes_{i\geq 1} \nu_{-,i}$$

with probability measures $\nu_{+,i}$ and $\nu_{-,i}$ on X_i for $i \geq 1$.

Recall from (1.7) that $\nu = p_+\nu_+ + p_-\nu_-$ holds true and hence the marginal distribution ν on X is generally not a product measure. The assumption that ν_+ and ν_- are product measures allows us to transfer properties of the corresponding one-dimensional distributions $\nu_{\pm,i}$ to $\nu_\pm$. The first lemma specifies the representation of the margin-noise in Lemma 2.3.4 under Assumption 3.3.1.

3.3.2 Lemma (Margin-Noise Function) *Let Assumption 3.3.1 be satisfied. If* $\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-$ *is a* ν_+*- or a* ν_-*-zero set then there is a version of* η *such that the following equality is satisfied, for all* $I \in \mathcal{F}(\mathbb{N})$ *and* $r \geq 0$,

$$\begin{aligned} MN_I(r) = M_I(r) = p_+ \prod_{i\in I} \nu_{+,i}\bigl(\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{-,i}) \leq 2r\bigr) \\ + p_- \prod_{i\in I} \nu_{-,i}\bigl(\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{+,i}) \leq 2r\bigr) \ . \end{aligned}$$

Moreover, for any other version of η *we have the inequality "*$\geq$*".*

To prove this we need the following auxiliary result.

3.3.3 Lemma (Auxiliary Result) *Let* $X = \prod_{i\geq 1} X_i$ *be a sequence space as defined in* (3.1), $I \in \mathcal{F}(\mathbb{N})$, *and* $M_I = \prod_{i\in I} M_i \subseteq X_I$ *be a cylinder in* X_I. *Then the following equality is satisfied, for* $x = (x_i)_{i\in I} \in X_I$,

$$\operatorname{dist}(x, M_I) = \max_{i\in I} \operatorname{dist}(x_i, M_i) \ ,$$

where X_I *is equipped with the metric* d_I *defined in* (3.11).

As the proof shows it is essential that the set I is finite.

Proof. The definitions of dist and d_I give us

$$\operatorname{dist}(x, M_I) = \inf_{(y_i)_{i\in I}\in M_I} \max_{i\in I} d_i(x_i, y_i) \quad .$$

For $(y_i)_{i\in I} \in M_I$ the fact that M_I is a cylinder implies $y_i \in M_i$ and hence $d_i(x_i, y_i) \geq \operatorname{dist}(x_i, M_i)$ holds for all $i \in I$. This proves the inequality "$\geq$". Conversely, for every $i \in I$ there is a sequence $\bigl(y_i^{(k)}\bigr)_{k\geq 1}$ in M_i with

$$\lim_{k\to\infty} \operatorname{dist}\bigl(x_i, y_i^{(k)}\bigr) = \operatorname{dist}(x_i, M_i) \quad .$$

Since M_I is a cylinder, we have $y^{(k)} \coloneqq \bigl(y_i^{(k)}\bigr)_{i\in I} \in M_I$ and hence

$$\operatorname{dist}(x, M_I) \leq d_I(x, y^{(k)}) = \max_{i\in I} d_i\bigl(x_i, y_i^{(k)}\bigr) \quad .$$

Taking the limit $k \to \infty$ gives the desired inequality "$\leq$" since the limit interchanges with the maximum over the finite set I. As a result, the assertion is proven. □

Now, we are ready to prove Lemma 3.3.2.

Proof of Lemma 3.3.2. Let $I \in \mathcal{F}(\mathbb{N})$ and $r \geq 0$ be fixed. After an application of Lemma 2.3.4 for $s = \pi_I$ it remains to prove

$$\nu_{\pm,I}\bigl(\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{\mp,I}) \leq 2r\bigr) = \prod_{i\in I} \nu_{\pm,i}\bigl(\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{\mp,i}) \leq 2r\bigr) \quad .$$

For symmetry reasons, it enough to consider $\nu_{+,I}$. Using Lemma B.2 and Lemma B.3 we get

$$\operatorname{supp}\nu_{-,I} = \overline{\pi_I(\operatorname{supp}\nu_{-})} = \overline{\prod_{i\in I} \operatorname{supp}\nu_{-,i}} = \prod_{i\in I} \operatorname{supp}\nu_{-,i} \quad ,$$

where we used that products of closed sets are closed, see e.g. [32, Corollary 2.3.4], in the last step. This shows that $\operatorname{supp}\nu_{-,I} \subseteq X_I$ is a cylinder in X_I. Using Lemma 3.3.3 with $M_I = \prod_{i\in I} \operatorname{supp}\nu_{-,i}$ and $\nu_{+,I} = \bigotimes_{i\in I} \nu_{+,i}$

we find

$$\begin{aligned}
&\nu_{+,I}\big(\operatorname{dist}(\,\cdot\,,\operatorname{supp}\nu_{-,I})\leq 2r\big)\\
&=\nu_{+,I}\Big((x_i)_{i\in I}\in X_I:\ \max_{i\in I}\operatorname{dist}(x_i,\operatorname{supp}\nu_{-,i})\leq 2r\Big)\\
&=\prod_{i\in I}\nu_{+,i}\big(\operatorname{dist}(\,\cdot\,,\operatorname{supp}\nu_{-,i})\leq 2r\big)
\end{aligned}$$

and hence the assertion is proven. □

The final lemma of this section provides some bounds on the number of relevant cells of a product partition under Assumption 3.3.1.

3.3.4 Lemma (Relevant Cells) *Let Assumption 3.3.1 be satisfied and $\mathcal{A}=(\mathcal{A}_i)_{i\geq 1}$ be a measurable partition of $(X_i)_{i\geq 1}$. Then, for every $I\in\mathcal{F}(\mathbb{N})$, the number of relevant cells satisfies*

$$\max\Big\{\prod_{i\in I}|\mathcal{A}_{i,\nu_{+,i}}|,\prod_{i\in I}|\mathcal{A}_{i,\nu_{-,i}}|\Big\}\leq|\mathcal{A}_{I,\nu}|\leq\prod_{i\in I}|\mathcal{A}_{i,\nu_{+,i}}|+\prod_{i\in I}|\mathcal{A}_{i,\nu_{-,i}}|\ .$$

Note that the upper bound is already contained in Lemma 3.2.1 even for general distributions ν. Consequently, this lemma states that Lemma 3.2.1 is almost optimal under Assumption 3.3.1. In other words, for fixed one-dimensional distributions $\nu_{\pm,i}$ on X_i, the number of relevant cells has the worst possible behavior if Assumption 3.3.1 is satisfied.

Proof. Since we have $\nu_I=p_+\nu_{+,I}+p_-\nu_{-,I}$ from (1.7) with some $p_\pm>0$, Lemma 1.3.3 gives us $\mathcal{A}_{I,\nu}=\mathcal{A}_{I,\nu_+}\cup\mathcal{A}_{I,\nu_-}$. Consequently, it is enough to determine $\mathcal{A}_{I,\nu_\pm}$. To this end, let $\mathcal{A}_i=(A_{i,k_i})_{k_i\in K_i}$ for $i\geq 1$ and $k=(k_i)_{i\in I}\in K_I$. Then $k\in\mathcal{A}_{I,\nu_\pm}$ if and only if

$$\nu_{\pm,I}(A_k)=\prod_{i\in I}\nu_{\pm,i}(A_{i,k_i})>0$$

where $A_k=\prod_{i\in I}A_{i,k_i}$. For a finite product of non-negative numbers, being positive is equivalent to the positivity of every factor, i.e. $\nu_{\pm,i}(A_{i,k_i})>0$ for all $i\in I$. The latter is equivalent to $k_i\in\mathcal{A}_{i,\nu_{\pm,i}}$ for all $i\in I$ or in other words $k \in \prod_{i \in I} \mathcal{A}_{i,\nu_{\pm,i}}$. All together this shows

$$\left|\mathcal{A}_{I,\nu_{\pm,I}}\right| = \left|\prod_{i \in I} \mathcal{A}_{i,\nu_{\pm,i}}\right| = \prod_{i \in I} \left|\mathcal{A}_{i,\nu_{\pm,i}}\right|$$

and hence the assertion is proven. □

Chapter 4

A Prototypical Example

In this chapter we investigate a prototypical infinite-dimensional classification problem. We can explicitly calculate the margin-noise function and give sharp bounds on the number of relevant cells for this learning problem. Depending on some parameters there are situations in which we get polynomial learning rates from Corollary 3.2.7 and other situations in which we can show that Corollary 3.2.7 does not provide polynomial learning rates. To get an over overview of the obtained learning rates see Table 4.3 in Section 4.3 below. Note that this classification problem is the basis for various generalizations presented in Chapter 5.

4.1 Definition and Basic Properties

For easy referencing we formulate our classification problem of interest as an assumption.

4.1.1 Assumption (Prototypical Example) Let $(\sigma_i)_{i\geq 1}$, $(\kappa_i)_{i\geq 1}$, $(q_i)_{i\geq 1}$ be sequences with $\sigma_i > 0$, $0 \leq \kappa_i \leq 1/2$, and $q_i > 0$ for all $i \geq 1$, and define the function

$$f_i(t) := q_i t^{q_i - 1} \mathbb{1}_{[0,1]}(t)$$

for $t \in \mathbb{R}$ and $i \geq 1$. Furthermore, let Assumption 3.3.1 be satisfied with $X_i = \mathbb{R}$, $\nu_{\pm,i} \ll \lambda$, where λ is the Lebesgue measure, and

$$\frac{\mathrm{d}\nu_{\pm,i}}{d\lambda}(t) = \frac{f_i(\kappa_i \pm t/\sigma_i)}{\sigma_i}$$

for λ-almost all $t \in \mathbb{R}$ and all $i \geq 1$.

First, we want to give some intuition for the influence of the parameters σ_i, κ_i, and q_i on the distributions $\nu_{\pm,i}$. To this end, assume that x_i is a random variable with $x_i \sim f_i \,\mathrm{d}\lambda$ then the transformed random variables $x_{\pm,i} := \pm(x_i - \kappa_i)\sigma_i$ satisfy $x_{\pm,i} \sim \nu_{\pm,i}$. In other words, the parameter κ_i shifts and the parameter σ_i scales the random variables $x_{\pm,i}$. Moreover, since $x_{+,i} = -x_{-,i}$ the distribution $\nu_{-,i}$ equals $\nu_{+,i}$ reflected at 0. As a result, the supports equal

$$\begin{aligned} \operatorname{supp}\nu_{+,i} &= \bigl[-\kappa_i\sigma_i, (1-\kappa_i)\sigma_i\bigr] \qquad \text{and} \\ \operatorname{supp}\nu_{-,i} &= \bigl[-(1-\kappa_i)\sigma_i, \kappa_i\sigma_i\bigr] \ , \end{aligned} \tag{4.1}$$

respectively. The effects of σ_i, κ_i, and q_i are visualized in Figure 4.1.

Now, we discuss the effect of κ_i in more detail. For $\kappa_i < 0$ there is a gap between the supports $\operatorname{supp}\nu_{+,i}$ and $\operatorname{supp}\nu_{-,i}$. Since this is an easy classification problem, this case is excluded by Assumption 4.1.1. For $\kappa_i = 0$ the supports touch but do not overlap. For $0 < \kappa_i < 1/2$ the supports overlap but there are still regions where only one class can be observed. For $\kappa_i = 1/2$ the supports $\operatorname{supp}\nu_{\pm,i} = [-\sigma_i/2, \sigma_i/2]$ are the same.

The parameter q_i describes where the probability masses of $\nu_{+,i}$ and $\nu_{-,i}$ are located within their support. The larger q_i, the less probability mass is in the overlapping region and vice versa. Moreover, for $q_i = 1$, the distributions $\nu_{+,i}$ and $\nu_{-,i}$ are uniform distributions.

The first lemma provides basic properties of the one-dimensional distributions $\nu_{\pm,i}$.

4.1.2 **Lemma (One-Dimensional Distributions)** *Let Assumption 4.1.1 be satisfied. Then the following statements are true, for $i \geq 1$:*

(i) $\nu_{\pm,i}\bigl(\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{\mp,i}) \leq 2r\bigr) = \min\bigl\{(2\kappa_i + 2r/\sigma_i)^{q_i}, 1\bigr\}$ *for $r \geq 0$.*

(ii) For any cubic partition $\mathcal{A}_i$ of $\mathbb{R}$ with radius $r > 0$ the number of relevant cells is bounded by

$$\lceil \sigma_i/(2r) \rceil \leq |\mathcal{A}_{i,\nu_\pm}| \leq 1 + \lceil \sigma_i/(2r) \rceil \ .$$

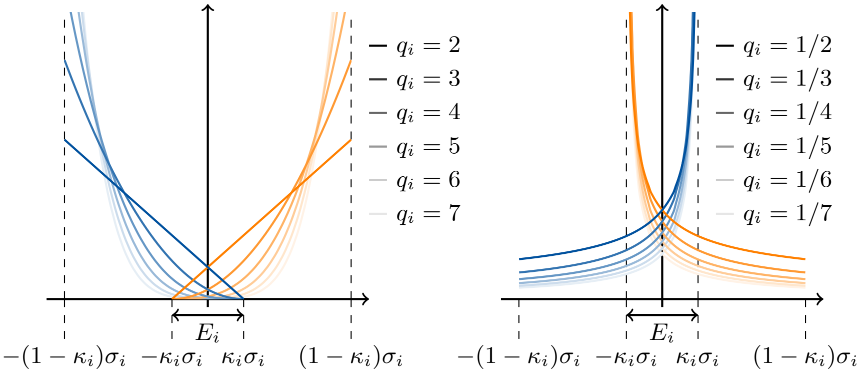


Figure 4.1: Plots of the (Lebesgue) density $\mathrm{d}\nu_{+,i}/\mathrm{d}\lambda$ in orange and the density $\mathrm{d}\nu_{-,i}/\mathrm{d}\lambda$ in blue for different values of q_i and fixed $\kappa_i = 1/5$, $\sigma_i = 1$. The left plot shows that for increasing $q_i \nearrow \infty$ the probability mass moves away from the overlapping region $E_i := [-\kappa_i\sigma_i, \kappa_i\sigma_i]$. The right plot shows that for decreasing $q_i \searrow 0$ the probability mass is more and more concentrated in the overlapping region E_i.

Proof. We prove both statements for $\nu_{+,i}$ and note that for symmetry reasons the same is true for $\nu_{-,i}$.

(i) Since $\operatorname{supp}\nu_{-,i} = [-(1-\kappa_i)\sigma_i, \kappa_i\sigma_i]$ holds true, we find

$$\big\{\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{-,i}) \le 2r\big\} = \big[-(1-\kappa_i)\sigma_i - 2r,\; \kappa_i\sigma_i + 2r\big]$$

and together with $\operatorname{supp}\nu_{+,i} = [-\kappa_i\sigma_i, (1-\kappa_i)\sigma_i]$ we get

$$\begin{aligned}&\operatorname{supp}\nu_{+,i} \cap \big\{\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{-,i}) \le 2r\big\}\\ &= \begin{cases}[-\kappa_i\sigma_i, \kappa_i\sigma_i + 2r], & 2\kappa_i + 2r/\sigma_i \le 1\\ \operatorname{supp}\nu_{+,i}, & 2\kappa_i + 2r/\sigma_i > 1\end{cases}.\end{aligned}$$

In the case $2\kappa_i + 2r/\sigma_i > 1$ we directly get $\nu_{+,i}\big(\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{-,i}) \le 2r\big) = 1$. In the case $2\kappa_i + 2r/\sigma_i \le 1$ the definition of $\nu_{+,i}$ together with some integral transformations yields

$$\begin{aligned}\nu_{+,i}\big(\operatorname{dist}(\,\cdot\,,\operatorname{supp}\nu_{-,i})\le 2r\big) &= \int_{-\kappa_i\sigma_i}^{\kappa_i\sigma_i+2r}\frac{f_i(\kappa_i+t/\sigma_i)}{\sigma_i}\,\mathrm{d}t\\ &= \int_0^{2\kappa_i+2r/\sigma_i} f_i(t)\,\mathrm{d}t\\ &= (2\kappa_i+2r/\sigma_i)^{q_i}\le 1\ .\end{aligned}$$

(ii) Since every cell of the cubic partition $\mathcal{A}_i$ is contained in a closed ball of radius r, Corollary 1.3.6 gives

$$|\mathcal{A}_{i,\nu_+}|\ge\mathcal{N}\big([-\kappa_i\sigma_i,(1-\kappa_i)\sigma_i],r\big)=\left\lceil\frac{\sigma_i}{2r}\right\rceil\ ,$$

where we used in the last step that $\operatorname{supp}\nu_{+,i}=[-\kappa_i\sigma_i,(1-\kappa_i)\sigma_i]$ is an interval of length σ_i. For the upper bound, Lemma 1.3.4 gives us $|\mathcal{A}_{i,\nu_+}|\le|(\mathcal{A}_i)_{\operatorname{supp}\nu_{+,i}}|$. Now, let A_k be the cell that contains the left corner of the interval $\operatorname{supp}\nu_{+,i}$, i.e. $-\kappa_i\sigma_i\in A_k$. Since all cells are aligned, the remaining interval $\operatorname{supp}\nu_{+,i}\backslash A_k$ of length $L\le\sigma_i$ is covered by $\lceil L/(2r)\rceil$ cells. Together this yields

$$|\mathcal{A}_{i,\nu_+}|\le\big|(\mathcal{A}_i)_{\operatorname{supp}\nu_{+,i}}\big|\le 1+\big\lceil L/(2r)\big\rceil\le 1+\big\lceil\sigma_i/(2r)\big\rceil$$

and hence the assertion is proven. □

Before we go to the multi-dimensional case let us have a look at Figure 4.2 that visualizes the effect of κ_i on the placement of the supports of ν_+ and ν_- in the two-dimensional case.

The next result combines Lemma 4.1.2 with the results of Section 3.3.

4.1.3 Corollary (Finite-Dimensional Distributions) *Let Assumption 4.1.1 be satisfied. Then the following statements are true:*

(i) If $\operatorname{supp}\nu_+\cap\operatorname{supp}\nu_-$ *is a* ν_+*- or a* ν_-*-zero set then there is no noise*

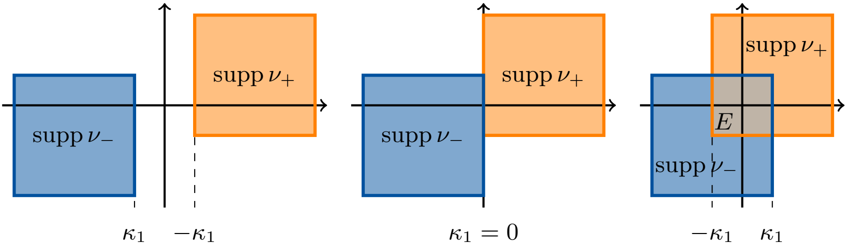


Figure 4.2: Plots of the supports of $\nu_+ := \nu_{+,1} \otimes \nu_{+,2}$ in orange and $\nu_- := \nu_{-,1} \otimes \nu_{-,2}$ in blue for different values of κ_i and $\sigma_1 = \sigma_2 = 1$. For all three plots we have $\kappa_2 = 1/4$. The left plot ($\kappa_1 = -1/4$) shows that there is a gap between the supports if one κ_i is negative. Since this is an easy classification problem, this case is excluded by Assumption 4.1.1. The middle plot ($\kappa_1 = 0$) shows that the supports touch each other if one κ_i (here κ_1) is zero. In this case they touch at a line, e.g. a one-dimensional set, because there is one other positive κ_j (here κ_2). Imagine if κ_2 would be zero too then the supports would only touch at a single point. The right plot ($\kappa_1 = 1/4$) shows that there is a overlapping region $E := \operatorname{supp} \nu_+ \cap \operatorname{supp} \nu_-$ if all κ_i are positive. This overlapping region E increases with increasing κ_i.

and there is a version of η with

$$MN_I(r) = \prod_{i \in I} \min\{(2\kappa_i + 2r/\sigma_i)^{q_i}, 1\}$$

for every $I \in \mathcal{F}(\mathbb{N})$ and $r \geq 0$. Moreover, for any other version of η we have the inequality "$\geq$".

(ii) For $I \in \mathcal{F}(\mathbb{N})$ and a cubic partition $\mathcal{A} = (\mathcal{A}_i)_{i \geq 1}$ of radius $r > 0$ the number of relevant cells satisfies

$$\prod_{i \in I} \lceil \sigma_i/(2r) \rceil \leq |\mathcal{A}_{I,\nu}| \leq 2 \cdot \prod_{i \in I} \bigl(1 + \lceil \sigma_i/(2r) \rceil\bigr) \ .$$

Note that the bound on the number of relevant cells is independent of

$(\kappa_i)_{i\geq 1}$ and $(q_i)_{i\geq 1}$. In contrast, the margin-noise function is non-decreasing in κ_i and non-increasing in q_i for every individual $i \geq 1$. As a result, the smaller κ_i and the larger q_i, the better our bound on the excess risk given by Corollary 3.2.7. In this sense, the sequences $(\kappa_i)_{i\geq 1}$ and $(q_i)_{i\geq 1}$ have the *most favorable behavior* if $\kappa_i \to 0$ and $q_i \to \infty$ for $i \to \infty$, respectively. Conversely, they have the *least favorable behavior* if $\kappa_i \to 1/2$ and $q_i \to 0$ for $i \to \infty$, respectively.

In a finite-dimensional setting, decreasing the radius r is typically the only way to decrease the margin-noise function. In contrast, in our infinite-dimensional setting if we add a feature $i \geq 1$ with $2\kappa_i + 2r/\sigma_i < 1$ to our feature set I then the margin-noise function decreases, too.

Recall from the discussion after Lemma 3.3.4 that the number of relevant cells $|\mathcal{A}_{I,\nu}|$ behaves unfavorably for an increasing feature set I. In order to get polynomial learning rates we have to compensate this bad behavior by ensuring that the margin-noise function $MN_I(r)$ decreases fast enough for increasing I. In other words, we have to ensure that we gain *enough* information for each feature that we use for learning. In Section 4.3 below we provide conditions on the sequences $(\kappa_i)_{i\geq 1}$ and $(q_i)_{i\geq 1}$ that allow feature selections such that $MN_I(r)$ behaves benignly in I.

Proof. Point (i) is a direct consequence of Lemma 3.3.2 and Point (i) of Lemma 4.1.2. Analogously, Point (ii) follows from Lemma 3.3.4 and Point (ii) of Lemma 4.1.2. □

An essential assumption of Lemma 4.1.2 it that $\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-$ is a ν_+- or a ν_--zero set. To check this it is useful to note that Lemma B.3 and (4.1) imply

$$\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_- = \prod_{i\geq 1}\big(\operatorname{supp}\nu_{+,i} \cap \operatorname{supp}\nu_{-,i}\big) = \prod_{i\geq 1}[-\kappa_i\sigma_i, \kappa_i\sigma_i] \ .$$

Plugging in the definitions of ν_+ and ν_-, respectively, yields

$$\nu_\pm(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = \prod_{i\geq 1}(2\kappa_i)^{q_i} \ . \tag{4.2}$$

As a result, $\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-$ is a ν_+-zero set if and only if it is a ν_--zero. Moreover, the scale sequence $(\sigma_i)_{i\geq 1}$ has no influence on $\nu(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-)$. The final lemma of this section provides conditions on $(\kappa_i)_{i\geq 1}$ and $(q_i)_{i\geq 1}$ that ensure $\nu(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = 0$.

4.1.4 Lemma (No Noise) *Let Assumption 4.1.1 be satisfied. Then* $\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-$ *is a* ν*-zero set for all the following situations:*

(i) There is some $i_0 \geq 1$ *with* $\kappa_{i_0} = 0$.

(ii) There are $q > 0$ *and* $\kappa < 1/2$ *with* $q_i \geq q$ *and* $\kappa_i \leq \kappa$ *for all* $i \geq 1$.

(iii) There is some $\kappa < 1/2$ *with* $\kappa_i \leq \kappa$ *for all* $i \geq 1$ *and* $q_i \to 0$ *with*

$$\sum_{i\geq 1} q_i = \infty \quad .$$

(iv) There is some $q > 0$ *with* $q_i \geq q$ *for all* $i \geq 1$ *and* $\kappa_i \to 1/2$ *with*

$$\sum_{i\geq 1}(1 - 2\kappa_i) = \infty \quad .$$

It is important to mention that, according to Lemma 1.2.2, all the conditions in Point (i)–(iv) imply that there is no noise and hence $\mathcal{R}^*_{\text{Class},P} = 0$ is satisfied. The condition in Point (i) additionally implies $\mathcal{R}^*_{\text{Class},P_I} = 0$ for all $I \in \mathcal{F}(\mathbb{N})$ with $i_0 \in I$. However, the conditions in Point (ii)–(iv) can also be satisfied if $\kappa_i > 0$ for all $i \geq 1$. In this case we have $\mathcal{R}^*_{\text{Class},P_I} > 0$ for all $I \in \mathcal{F}(\mathbb{N})$ and hence *no* (pull-back) learning method using a fixed feature set $I \in \mathcal{F}(\mathbb{N})$ for all $n \geq 1$ can be (potentially) consistent.

Proof. Due to (4.2) it is enough to show $\prod_{i\geq 1}(2\kappa_i)^{q_i} = 0$ in all cases. (i) Since there is a zero factor, the product is zero. (ii) Since $(2\kappa)^q < 1$ is satisfied, we find

$$\prod_{i\geq 1}(2\kappa_i)^{q_i} \leq \prod_{i\geq 1}(2\kappa)^q = 0 \quad .$$

(iii) Since $\log(2\kappa) < 0$ and $\sum_{i\geq 1} q_i = \infty$ are satisfied, we find

$$\prod_{i\geq 1}(2\kappa_i)^{q_i} \leq \prod_{i\geq 1}(2\kappa)^{q_i} = \exp\Bigl(\log(2\kappa)\sum_{i\geq 1} q_i\Bigr) = 0 \ .$$

(iv) Using $\log(x) \leq x - 1$ for $x = 2\kappa_i$ we find

$$\prod_{i\geq 1}(2\kappa_i)^{q_i} \leq \prod_{i\geq 1}(2\kappa_i)^{q} = \exp\Bigl(q\sum_{i\geq 1}\log(2\kappa_i)\Bigr) \leq \exp\Bigl(-q\sum_{i\geq 1}(1-2\kappa_i)\Bigr) \ .$$

Together with our assumption $\sum_{i\geq 1}(1-2\kappa_i) = \infty$ we get the assertion. □

4.2 Auxiliary Results: Polynomial Learning Rates

In order to establish learning rates for a learning algorithm which depends on a hyper parameter, say m, one typically has to balance an expression like

$$a_m + \left(\frac{\tau b_m}{n}\right)^v \tag{4.3}$$

with $a_m, b_m > 0$ and $v > 0$ for all data set sizes $n \geq 1$. Since in our case $m \geq 1$ is often integer-valued, we cannot use derivatives to explicitly minimize (4.3) for $m \geq 1$.

In this section we present conditions that allow a hyper parameter selection $m^*(n)$ giving a polynomial decay of (4.3) for $n \to \infty$ as well as conditions on a parameter selection $m^*(n)$ that prevent a polynomial decay of (4.3) with $m = m^*(n)$. To this end, we introduce the following notation. For a function $f\colon [x_0, \infty) \to [0, \infty)$ with $x_0 > 0$ we say that f *increases sub-polynomially* for $x \to \infty$ if

$$\lim_{x\to\infty} f(x)x^{-\varepsilon} = 0$$

for every $\varepsilon > 0$.

First, we present the positive result which ensures a polynomial hyper parameter selection.

4.2.1 Lemma (Polynomial Parameter Selection) *Let $v > 0$, $(a_m)_{m\geq 1}$ and $(b_m)_{m\geq 1}$ be sequences with $a_m, b_m > 0$ for $m \geq 1$ and $a_m \to 0$ for $m \to \infty$. Furthermore, define*

$$\rho_m := \sup_{k\geq m} \frac{\log(b_k)}{-\log(a_{k-1})}$$

for $m \geq 2$ and let

$$\rho := \lim_{m\to\infty} \rho_m = \limsup_{m\to\infty} \frac{\log(b_m)}{-\log(a_{m-1})} < \infty$$

be finite. Then the following statements are true:

(i) There is a constant $x_0 \geq 1$ such that for every $x \geq x_0$ there is a $m \geq 2$ with

$$a_m \leq x^{-\frac{v}{1+v\rho}} < a_{m-1} \quad . \tag{4.4}$$

The minimal integer m with this property is denoted by $m^*(x)$. This function satisfies $m^*(x) \to \infty$ for $x \to \infty$.

(ii) The function $f\colon [x_0, \infty) \to (0, \infty)$ given by

$$f(x) := 1 + \exp\left(\frac{v^2}{1+v\rho} \cdot \log(x) \cdot (\rho_{m^*(x)} - \rho)\right)$$

increases sub-polynomially and is even bounded if $\log(x) \cdot (\rho_{m^*(x)} - \rho)$ is bounded.

(iii) For the choice $m = m^*(n/\tau)$ the following bound is satisfied

$$a_m + \left(\frac{\tau b_m}{n}\right)^v \leq f(n/\tau) \cdot \left(\frac{\tau}{n}\right)^{\frac{v}{1+v\rho}} \quad .$$

Let us briefly demonstrate the typical usage of that lemma. For $m \geq 1$ and some $\mathcal{I}_m \subseteq \mathcal{F}(\mathbb{N})$ and $R_m \subseteq (0, \infty)$, we prove bounds on the margin-noise function $MN_I(r) \leq c \cdot a_m$ and the number of relevant cells $|\mathcal{A}_{I,\nu}| \leq c \cdot b_m$ for all $I \in \mathcal{I}_m$, $r \in R_m$, and all cubic partitions $\mathcal{A}$ of radius $r \in R_m$. If the sequences $(a_m)_{m\geq 1}$ and $(b_m)_{m\geq 1}$ satisfy the assumptions of Lemma 4.2.1

then Corollary 3.2.7 and Lemma 4.2.1 for $v = \frac{q+1}{q+2}$ yield

$$\begin{aligned}\mathcal{R}_{\text{Class},P}(h_{D,\mathcal{A},I}) - \mathcal{R}^*_{\text{Class},P} &\leq 6MN_I(r) + C\left(\frac{\tau|\mathcal{A}_{I,\nu}|}{n}\right)^{\frac{q+1}{q+2}} \\ &\leq \max\{6c, Cc\} \cdot f(n/\tau) \cdot \left(\frac{\tau}{n}\right)^{\frac{q+1}{(q+2)+(q+1)\rho}}\end{aligned}$$

for all $\tau \geq 1$, $n \geq x_0\tau$, $I \in \mathcal{I}_{m^*(n/\tau)}$, $r \in R_{m^*(n/\tau)}$ with high probability. This proves polynomial learning rates for the histogram.

Proof. (i) Let $x_0 > 0$ be a number with $x_0^{-\frac{v}{1+v\rho}} < \sup_{m\geq 1} a_m$. Since $a_m \to 0$ for $m \to \infty$, for every $x \geq x_0$ there is an integer m satisfying (4.4) . Moreover, since $a_m > 0$ for all $m \geq 1$ and $x^{-\frac{v}{1+v\rho}} \to 0$ for $x \to \infty$ the function $m^*(x)$ converges to ∞ for $x \to \infty$.

(ii) Since $m^*(x) \to \infty$ for $x \to \infty$, we have $\rho_{m^*(x)} - \rho \to 0$ for $x \to \infty$. Consequently, for every $\varepsilon > 0$, the following function converges to 0 for $x \to \infty$

$$f(x)x^{-\varepsilon} = x^{-\varepsilon} + \exp\left(\log(x) \cdot \left(\frac{v^2}{1+v\rho} \cdot (\rho_{m^*(x)} - \rho) - \varepsilon\right)\right) \quad .$$

The statement about the boundedness of f follows immediately by the definition of f.

(iii) The definition of ρ_m ensures

$$b_m \leq \left(\frac{1}{a_{m-1}}\right)^{\rho_m}$$

for $m \geq 2$ and the choice $m = m^*(n/\tau)$ gives $(\tau/n)^{\frac{v}{1+v\rho}} < a_{m-1}$. Together we find

$$\left(\frac{\tau b_m}{n}\right)^v \leq \left(\frac{\tau}{n} \cdot (n/\tau)^{\frac{v\rho_m}{1+v\rho}}\right)^v = \left(\frac{\tau}{n}\right)^{(1-\frac{v\rho_m}{1+v\rho})v}$$

Since the exponent equals $\left(1 - \frac{v\rho_m}{1+v\rho}\right)v = \frac{v}{1+v\rho} + \frac{v^2}{1+v\rho}(\rho_m - \rho)$ a combination with $a_m \leq (\tau/n)^{\frac{v}{1+v\rho}}$ gives the assertion. □

Finally, we present the negative result.

4.2.2 Lemma (No Polynomial Parameter Selection) *Let $v > 0$, $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ be sequences with $a_n, b_n > 0$ for $n \geq 1$ and $a_n \to 0$ for $n \to \infty$. If*

$$\limsup_{n\to\infty} \frac{\log(b_n)}{-\log(a_n)} = \infty \tag{4.5}$$

then for all $\gamma > 0$

$$\limsup_{n\to\infty} \left(a_n + \left(\frac{b_n}{n}\right)^v \right) n^\gamma = \infty \ .$$

Let us briefly demonstrate the typical usage of that lemma. We take arbitrary sequences $(r_n)_{n\geq 1}$ and $(I_n)_{n\geq 1}$ in $(0,\infty)$ and $\mathcal{F}(\mathbb{N})$, respectively. Then we prove lower bounds on the margin-noise function $MN_{I_n}(r_n) \geq c \cdot a_n$ and the number of relevant cells $|\mathcal{A}_{I_n,\nu}| \geq c \cdot b_n$ for all cubic partitions $\mathcal{A}$ of radius r_n with some $c > 0$. If the sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ satisfy the assumption of Lemma 4.2.2 then the bound of Corollary 3.2.7 satisfies, for every $\gamma > 0$,

$$\begin{aligned} &\limsup_{n\to\infty} \left(6MN_{I_n}(r_n) + C\left(\frac{\tau|\mathcal{A}_{I_n,\nu}|}{n}\right)^{\frac{q+1}{q+2}} \right) n^\gamma \\ \geq\ & \min\{6c, Cc\} \cdot \limsup_{n\to\infty} \left(a_n + \left(\frac{b_n}{n}\right)^v \right) n^\gamma = \infty \ . \end{aligned}$$

Since we considered arbitrary sequences $(r_n)_{n\geq 1}$ and $(I_n)_{n\geq 1}$, Corollary 3.2.7 does not provide polynomial learning rates in this situation.

Proof. We assume that there are $C, \gamma > 0$ with

$$a_n + \left(\frac{b_n}{n}\right)^v \leq Cn^{-\gamma}$$

and show that this assumption leads to a contradiction. Since $a_n \to 0$ for $n \to \infty$, the assumption in (4.5) implies $b_n \to \infty$ for $n \to \infty$. Our assumed inequality implies two further inequalities, namely

$$a_n \leq Cn^{-\gamma} \qquad \text{and} \qquad b_n \leq (Cn^{v-\gamma})^{1/v} \ .$$

Since $b_n \to \infty$ for $n \to \infty$, this implies $v > \gamma$. Moreover, we get

$$\frac{\log(b_n)}{-\log(a_n)} \leq \frac{1}{v} \cdot \frac{\log(C) + (v-\gamma)\log(n)}{\log(1/C) + \gamma\log(n)} \to \frac{v-\gamma}{v\gamma} < \infty$$

for $n \to \infty$. This is a contradiction to our assumption in (4.5) and hence the assertion is proven. □

4.3 Polynomial Learning Rates

In this section we provide learning rates for histograms applied to the prototypical classification problem from Assumption 4.1.1. For this purpose, we combine the results of Section 4.1 and Section 4.2 with the oracle inequality for histograms in Corollary 3.2.7. For some cases this establishes polynomial learning rates. For other cases this shows that Corollary 3.2.7 does not provide polynomial learning rates. Note that the learning rates of this section are the basis for various generalizations presented in Chapter 5. To get an overview of the obtained learning rates see Table 4.3.

Before we start with the learning rates let us comment on the scaling sequence $(\sigma_i)_{i\geq 1}$. For the histogram, as for most learning algorithms, the scaling σ_i of a feature relative to the other features determines its influence on the decision function. Larger values of σ_i correspond to a higher influence. Since we do not have any additional information about the relevance of a feature, we assume a fixed scale factor $\sigma_i := \sigma$ for all features $i \geq 1$ in this section. Also in practice, scaling all features approximately to the same order of magnitude is typically the starting point of every data analysis. However, there are situations in which all features share a common *measurement unit*, see e.g. Section 5.2 and Section 5.3 below for situations in which the scaling naturally decreases and increases, respectively, for $i \to \infty$.

This section is organized as follows: We start with a result which provides polynomial learning rates for the histogram under Assumption 4.1.1 and some additional assumptions on the sequences $(\kappa_i)_{i\geq 1}$ and $(q_i)_{i\geq 1}$. Afterwards, we show that if we slightly strengthen these assumptions, we can

	learning problem		hyper parameters		learning
Lem.	$(\kappa_i)_{i\geq 1}$	$(q_i)_{i\geq 1}$	radius r_n	feature set I_n	rate $n^{-\rho}$
4.3.1	$\kappa_i \ll 1/2$	$q_i \gg 0$	$r_n = r$	$\lvert I_n \rvert \asymp \log(n)$	$\frac{q\alpha}{q\alpha+\beta}$
4.3.2	$0 \ll \kappa_i \ll 1/2$	$0 \ll q_i \ll \infty$	$r_n \to 0$	x	x
4.3.3	$\kappa_i \ll 1/2$	$\boldsymbol{q_i \nearrow \infty}$	$r_n = r$	$\lvert I_n \rvert \to \infty$ or $I_n = \{i_n\}$ with $i_n \to \infty$	1
4.3.4	$\boldsymbol{\kappa_i \searrow 0}$	$q_i \gg 0$	$r_n \to 0$	——"——	$\frac{q}{q+1}$
4.3.5	$0 \ll \kappa_i \ll 1/2$	$\boldsymbol{q_i \searrow 0}$	x	x	x
4.3.6	$\boldsymbol{\kappa_i \nearrow 1/2}$	$0 \ll q_i \ll \infty$	x	x	x

Table 4.3: Simplified versions of the learning rates presented in Section 4.3 for the prototypical example from Assumption 4.1.1. In all situations we use a constant scale sequence $\sigma_i = \sigma > 0$ for $i \geq 1$ and ignore sub-polynomial terms in the learning rate. The symbols "$\ll$" and "$\gg$" mean *bounded away from* by κ and q, respectively. Recall that the margin-noise function is *small* if κ_i is close to 0 or if q_i is close to ∞ and vice versa it is *large* if κ_i is close to $1/2$ or if q_i is close to 0. In this sense, the color and fond-weight indicate that the respective parameter sequence has ***the least favorable behavior***, *both, the most and least favorable behavior is excluded*, *only the least favorable behavior is excluded*, and ***has the most favorable behavior*** for $i \to \infty$. The symbol "x" means that there is no restriction or that Corollary 3.2.7 does not provide polynomial learning rates.

improve the polynomial order of the learning rate under weaker restrictions on the feature set. In some cases we even get the optimal learning rate n^{-1}. Finally, we show that if we slightly weaken these assumptions, Corollary 3.2.7 does not provide polynomial learning rates. In this sense, the next theorem, which is the main result of this chapter, is on a thin line between the optimal (polynomial) rate n^{-1} and no polynomial learning rate.

4.3.1 Theorem ($\boldsymbol{\kappa_i \leq \kappa}$ and $\boldsymbol{q_i \geq q}$) *Let Assumption 4.1.1 be satisfied for $\sigma_i = \sigma$, $\kappa_i \leq \kappa$, and $q_i \geq q$ for all $i \geq 1$ with some $\sigma, q > 0$, and $0 \leq \kappa < 1/2$. Furthermore, let $0 < r < (1/2 - \kappa)\sigma$ be fixed and $\alpha := -\log(2\kappa + 2r/\sigma)$,*

$\beta := \log\bigl(1 + \lceil \sigma/(2r) \rceil\bigr)$. Then there are constants $x_0, C \geq 1$ with the following property:
For $\tau \geq 1$, $n \geq x_0\tau$, and every feature set $I_n \in \mathcal{F}(\mathbb{N})$ with

$$|I_n| = \left\lceil \frac{\log(n/\tau)}{q\alpha + \beta} \right\rceil - 1$$

the histogram using a cubic partition with radius r and the feature set I_n satisfies

$$\mathcal{R}_{\mathrm{Class},P}(h_{D,r,I_n}) - \mathcal{R}^*_{\mathrm{Class},P} < C \cdot \left(\frac{\tau}{n}\right)^{\frac{q\alpha}{q\alpha+\beta}} \tag{4.6}$$

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

This lemma provides polynomial learning rates for every fixed radius $0 < r < (1/2 - \kappa)\sigma$ if κ_i is bounded away from $1/2$ and q_i is bounded away from 0. In other words, we get polynomial learning rates if the least favorable behavior of $(\kappa_i)_{i\geq 1}$ and $(q_i)_{i\geq 1}$ for $i \to \infty$ is excluded. Moreover, the polynomial order of the learning rate depends on the fixed radius r. In Section 4.4 below we present *good* choices for the radius r.

Note that there is only a restriction on the number of features $|I|$ and hence it does not matter which features are exactly used for learning. Here the easy feature selection method which chooses the indicated number of features randomly does the job. More precisely, if the application at hand provides us the features $I'_n = [d'_n] = \{1, 2, \ldots, d'_n\}$ with

$$d'_n \geq \left\lceil \frac{\log(n/\tau)}{q\alpha + \beta} \right\rceil - 1 \ ,$$

then histograms combined with this easy feature selection strategy learn with the rate in (4.6). Note that for stronger restrictions on $(\kappa_i)_{i\geq 1}$ or $(q_i)_{i\geq 1}$, i.e. for a smaller upper bound κ or a larger lower bound q, the learning rate in (4.6) is faster and the restriction on d'_n is weaker, i.e. we need less information per data point.

Proof. The proof is an application of Corollary 3.2.7 and Lemma 4.2.1. To this end, let us fix some feature set $I \in \mathcal{F}(\mathbb{N})$ with $|I| =: m$ and some cubic partition $\mathcal{A}$ with radius r. According to Point (ii) of Lemma 4.1.4 the assumption of Corollary 4.1.3 is satisfied and hence we find

$$\begin{aligned} MN_I(r) &\leq \prod_{i\in I}(2\kappa + 2r/\sigma)^q = e^{-q\alpha m} = e^{q\alpha} a_m \qquad \text{and} \\ |\mathcal{A}_{I,\nu}| &\leq 2\cdot \prod_{i\in I}\bigl(1 + \lceil \sigma/(2r)\rceil\bigr) = 2e^{\beta m} = 2b_m \end{aligned}$$

with $a_m \coloneqq \exp(-q\alpha(m+1))$ and $b_m \coloneqq \exp(\beta m)$. It remains to investigate the properties of the sequences $(a_m)_{m\geq 1}$ and $(b_m)_{m\geq 1}$. Note that the restriction $0 < r < (1/2-\kappa)\sigma$ ensures $\alpha > 0$ and hence $a_m \to 0$ for $m \to \infty$. Moreover, we find that

$$\frac{\log(b_m)}{-\log(a_{m-1})} = \frac{m\beta}{q\alpha m} = \frac{\beta}{q\alpha} =: \rho < \infty$$

is constant. As a result, Lemma 4.2.1 gives $x_0 \geq 1$, $m^*\colon [x_0,\infty) \to \mathbb{N}$, and $f\colon [x_0,\infty) \to (0,\infty)$ such that Corollary 3.2.7 yields

$$\mathcal{R}_{\text{Class},P}(h_{D,\mathcal{A},I}) - \mathcal{R}^*_{\text{Class},P} \leq f(n/\tau)\cdot \left(\frac{\tau}{n}\right)^{\frac{q\alpha}{q\alpha+\beta}}$$

for $\tau \geq 1$, $n \geq x_0\tau$, $|I| = m^*(n/\tau)$ with high probability. Since the above fraction is even constant, the function f is bounded. Consequently, it remains to determine the function $m^*(x)$. To this end, fix $x \geq x_0$ and

$$\begin{aligned} a_m = \exp(-q\alpha(m+1)) \leq x^{-\frac{1}{1+\rho}} &\iff & x^{\frac{1}{1+\rho}} &\leq \exp(q\alpha(m+1)) \\ &\iff & \frac{\log(x)}{q\alpha(1+\rho)} &\leq m+1 \ . \end{aligned}$$

Using $\rho = \beta/(q\alpha)$ and an analogous equivalence for $x^{-\frac{1}{1+\rho}} < a_{m-1}$ we get

$$a_m \leq x^{-\frac{1}{1+\rho}} < a_{m-1} \qquad \iff \qquad m < \frac{\log(x)}{q\alpha+\beta} \leq m+1 \ .$$

This proves the claimed representation of $m^*(x)$ and hence the assertion is proven. □

The choice of a fixed radius r in the previous lemma is unexpected since for finite-dimensional learning problems one typically has to assume that the radius $r_n \to 0$ vanishes for $n \to \infty$ to ensure consistency or learning rates of the histogram learning method, see e.g. [10, Example 2.14] and [66, Proposition E.5]. For this reason, we consider the situation $r_n \to 0$ for $n \to \infty$ in the next lemma.

4.3.2 Lemma ($\kappa' \leq \kappa_i \leq \kappa$ and $q \leq q_i \leq q'$ with $r_n \to 0$) *Let the assumptions of Theorem 4.3.1 be satisfied and additionally $\kappa_i \geq \kappa'$ and $q_i \leq q'$ for all $i \geq 1$ with some $0 < \kappa' < 1/2$ and $0 < q' < \infty$. Furthermore, let $(r_n)_{n\geq 1}$ be a sequence with $r_n > 0$ for $i \geq 1$ and $r_n \to 0$ for $n \to \infty$. Then Corollary 3.2.7 does not provide polynomial learning rates for the histogram h_{D,r_n,I_n} with any feature set sequence $(I_n)_{n\geq 1}$.*

Roughly speaking, the restrictions on κ_i and q_i ensure that the most and the least favorable behavior of the parameter sequences $(\kappa_i)_{i\geq 1}$ and $(q_i)_{i\geq 1}$ for $i \to \infty$ are excluded. Note that the additional requirements cannot be removed without replacement. To be more precise, if $\kappa_i \searrow 0$ then Lemma 4.3.4 below shows that the polynomial learning rate of Theorem 4.3.1 can even be improved when $r_n \to 0$.

The negative result in Lemma 4.3.2 seems to be contradicting to the consistency result in Corollary 3.2.5 in which we assumed $r_n \to 0$ for $n \to \infty$. But even if Corollary 3.2.7 does not provide polynomial learning rates the histogram can still be consistent.

Proof. The proof is an application of Lemma 4.2.2. Let $(I_n)_{n\geq 1}$ be an arbitrary sequence of feature sets $I_n \in \mathcal{F}(\mathbb{N})$ and $\mathcal{A}$ be a cubic partition of $\mathbb{R}^{\mathbb{N}}$ with radius r_n. According to Point (ii) of Lemma 4.1.4 the assumption of Corollary 4.1.3 is satisfied and hence gives us

$$MN_{I_n}(r_n) \geq \prod_{i\in I_n} \min\bigl\{(2\kappa' + 2r_n/\sigma)^{q_i}, 1\bigr\} \geq (2\kappa')^{|I_n|q'} =: a_n \qquad \text{and}$$

$$|\mathcal{A}_{I_n,\nu}| \geq \bigl\lceil \sigma/(2r_n) \bigr\rceil^{|I_n|} \geq \bigl(\sigma/(2r_n)\bigr)^{|I_n|} =: b_n \ .$$

Since $r_n \to 0$ for $n \to \infty$, we find

$$\frac{\log(b_n)}{-\log(a_n)} = \frac{\log\bigl(\sigma/(2r_n)\bigr)}{q' \log\bigl(1/(2\kappa')\bigr)} \to \infty$$

for $n \to \infty$ and hence Lemma 4.2.2 gives the assertion. □

Now, we present two further results that show if one of the parameter sequences $(\kappa_i)_{i\geq 1}$ or $(q_i)_{i\geq 1}$ has the most favorable behavior, i.e. $q_i \nearrow \infty$ or $\kappa_i \searrow 0$ for $i \to \infty$, then we even can improve the polynomial learning rates of Theorem 4.3.1.

4.3.3 Lemma ($\kappa_i \leq \kappa$ and $q_i \nearrow \infty$) *Let Assumption 4.1.1 be satisfied for $\sigma_i = \sigma$, $\kappa_i \leq \kappa$, and $q_i > 0$ for all $i \geq 1$ with some $\sigma > 0$, $0 \leq \kappa < 1/2$ as well as $q_i \nearrow \infty$ for $i \to \infty$. Furthermore, let $0 < r < (1/2 - \kappa)\sigma$ be fixed. Then the following statements are true:*

(i) There is a constant $x_1 \geq 1$, a function $m_1^*\colon [x_0, \infty) \to \mathbb{N}$, and a function $f_1\colon [x_1, \infty) \to (0, \infty)$ increasing sub-polynomially with the following property:
For $\tau \geq 1$, $n \geq x_1\tau$, and every feature set $I_n \in \mathcal{F}(\mathbb{N})$ with $|I_n| = m_1^*(n/\tau)$ the histogram using a cubic partition with radius r and the feature set I_n satisfies

$$\mathcal{R}_{\text{Class},P}(h_{D,r,I_n}) - \mathcal{R}^*_{\text{Class},P} < f_1(n/\tau) \cdot \frac{\tau}{n}$$

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

(ii) There is a constant $x_2 \geq 1$, a function $m_2^*\colon [x_0, \infty) \to \mathbb{N}$, and a constant $C_2 \geq 1$ with the following property:
For $\tau \geq 1$, $n \geq x_0\tau$, and every feature set $I_n = \{i_n\}$ with $i_n \geq m_2^*(n/\tau)$ the histogram using a cubic partition with radius r and the feature set I_n satisfies

$$\mathcal{R}_{\text{Class},P}(h_{D,r,I_n}) - \mathcal{R}^*_{\text{Class},P} < C_2 \cdot \frac{\tau}{n}$$

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

This lemma provides the *optimal* learning rate n^{-1}, up to a sub-polynomial term, for every fixed radius $0 < r < (1/2 - \kappa)\sigma$ in two extreme feature selection regimes. In Point (i) there is only a restriction on the number of features $|I|$ used for learning. Roughly speaking, we do not have to worry about how useful the features are as long as we choose the right number of features. Point (ii) shows that using a single *good* feature for learning is enough to learn with rate n^{-1}. Therefore, at first sight, this learning problem seems to be actually a finite-dimensional one. However, the good feature depends on n and since $m_2^*(x) \to \infty$ for $x \to \infty$ all together we still need infinitely many different features for the whole learning process. Since $(q_i)_{i\geq 1}$ is non-decreasing, in this situation it is easy to find the good feature, namely the feature with the largest index available. But in practice it is typically a challenging problem to find useful features.

If the application at hand provides us the features $I'_n = [d'_n]$ then the feature selection strategies of Point (i) and (ii) can be applied if

$$d'_n \geq m_1^*(n/\tau) \qquad \text{and} \qquad d'_n \geq m_2^*(n/\tau)$$

is satisfied, respectively. Moreover, if the behavior of $(q_i)_{i\geq 1}$ becomes more favorable for $i \to \infty$, i.e. $(q_i)_{i\geq 1}$ increases faster, then a closer look at the proof shows that the functions $m_i^*(x)$, $i = 1, 2$, increase slower for $x \to \infty$. This means that the restriction on d'_n gets weaker, i.e. we need less information per data point.

Although for every fixed $0 < r < (1/2 - \kappa)\sigma$ the function f_1 increases sub-polynomially, these functions do not have the same asymptotic behavior for different values of r. Analogously, the constant C_2 depends on the radius r as well.

Proof. The proof is an application of Corollary 3.2.7 and Lemma 4.2.1. To this end, we use the notation $\alpha := -\log(2\kappa + 2r/\sigma)$ and $\beta := \log\bigl(1 + \lceil\sigma/(2r)\rceil\bigr)$ from Theorem 4.3.1. Recall that the restriction $0 < r < (1/2-\kappa)\sigma$ ensures $\alpha > 0$. Furthermore, we fix some cubic partition $\mathcal{A}$ of $\mathbb{R}^{\mathbb{N}}$ with radius r and a feature set $I \in \mathcal{F}(\mathbb{N})$. According to Point (ii) of Lemma 4.1.4

the assumption of Corollary 4.1.3 is satisfied and hence we find

$$\begin{aligned} MN_I(r) &\le \prod_{i\in I}(2\kappa + 2r/\sigma)^{q_i} \qquad \text{and} \\ |\mathcal{A}_{I,\nu}| &\le 2\cdot \prod_{i\in I}\bigl(1+\lceil \sigma/(2r)\rceil\bigr) = 2e^{\beta |I|} \ . \end{aligned}$$

Now, we consider Point (i) and (ii) separately.

(i) Let I be a feature set of cardinality $m := |I| \ge 1$. Then the monotonicity of $(q_i)_{i\ge 1}$ ensures

$$MN_I(r) \le \prod_{i=1}^{m}(2\kappa + 2r/\sigma)^{q_i} = \exp\Bigl(-\alpha \sum_{i=1}^{m} q_i\Bigr) =: a_m$$

and $|\mathcal{A}_{I,\nu}| \le 2b_m$ with $b_m := e^{\beta m}$. Since the sequence of the arithmetic means $\bigl(\frac{1}{m-1}\sum_{i=1}^{m-1} q_i\bigr)_{m\ge 2}$ has the same limit as the sequence $(q_i)_{i\ge 1}$ itself, we get

$$\frac{\log(b_m)}{-\log(a_{m-1})} = \frac{m\beta}{\alpha\sum_{i=1}^{m-1} q_i} = \frac{m}{m-1}\cdot \frac{\beta}{\frac{1}{m-1}\sum_{i=1}^{m-1} q_i} \to 0 =: \rho < \infty$$

for $m \to \infty$. As a result, Lemma 4.2.1 gives $x_1 \ge 1$, $m_1^*\colon [x_1,\infty) \to \mathbb{N}$, and a function $f_1\colon [x_1,\infty) \to (0,\infty)$ increasing sub-polynomially such that in combination with Corollary 3.2.7 we get the desired bound for feature sets $I \in \mathcal{F}(\mathbb{N})$ with $|I| = m_1^*(n/\tau)$ and cubic partitions with radius r.

(ii) Let $I = \{i\}$ be a singleton with $i \ge m \ge 1$. Then the monotonicity of $(q_i)_{i\ge 1}$ ensures

$$MN_I(r) \le (2\kappa + 2r/\sigma)^{q_i} \le \exp(-\alpha q_m) =: a_m$$

and $|\mathcal{A}_{I,\nu}| \le 2e^{\beta} b_m$ with $b_m := 1$. Since

$$\frac{\log(b_m)}{-\log(a_{m-1})} = \frac{\log(1)}{\alpha q_{m-1}} = 0 =: \rho < \infty$$

holds true for all $m \ge 1$, Lemma 4.2.1 gives $x_2 \ge 1$, $m_2^*\colon [x_2,\infty) \to \mathbb{N}$, and a constant $C_2 \geq 1$ such that in combination with Corollary 3.2.7 we get the desired bound for feature sets $I = \{i\}$ with $i \geq m_2^*(n/\tau)$ and cubic partitions with radius r. □

Next, we consider a further improvement of Theorem 4.3.1 in the case $\kappa_i \searrow 0$ for $i \to \infty$.

4.3.4 Lemma ($\kappa_i \searrow 0$ and $q_i \geq q$) *Let Assumption 4.1.1 be satisfied for $\sigma_i = \sigma$, $\kappa_i > 0$, and $q_i \geq q$ for all $i \geq 1$ with some $\sigma, q > 0$ as well as $\kappa_i \searrow 0$ for $i \to \infty$. Furthermore, define*

$$\bar{\kappa}_m := \frac{1}{m} \sum_{i=1}^{m} \kappa_i$$

for $m \geq 1$. Then the following statements are true:

(i) There is a constant $x_1 \geq 1$, a function $m_1^* \colon [x_1, \infty) \to \mathbb{N}$, and a function $f_1 \colon [x_0, \infty) \to (0, \infty)$ increasing sub-polynomially with the following property:
For $\tau \geq 1$, $n \geq x_1\tau$, and every feature set $I_n \in \mathcal{F}(\mathbb{N})$ with $|I_n| = m_1^*(n/\tau)$ the histogram using a cubic partition with radius $r_n := \bar{\kappa}_{m_1^*(n/\tau)}$ and the feature set I_n satisfies

$$\mathcal{R}_{\mathrm{Class},P}(h_{D,r_n,I_n}) - \mathcal{R}^*_{\mathrm{Class},P} < f_1(n/\tau) \cdot \left(\frac{\tau}{n}\right)^{\frac{q}{q+1}}$$

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

(ii) There is a constant $x_2 \geq 1$, a function $m_2^* \colon [x_2, \infty) \to \mathbb{N}$, and a constant $C_2 \geq 1$ with the following property:
For $\tau \geq 1$, $n \geq x_2\tau$, and every feature set $I_n = \{i_n\}$ with $i_n \geq m_2^*(n/\tau)$ the histogram using a cubic partition with radius $r_n := (\tau/n)^{1/(q+1)}$ and the feature set I_n satisfies

$$\mathcal{R}_{\mathrm{Class},P}(h_{D,r_n,I_n}) - \mathcal{R}^*_{\mathrm{Class},P} < C_2 \cdot \left(\frac{\tau}{n}\right)^{\frac{q}{q+1}}$$

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

In contrast to Theorem 4.3.1 and Lemma 4.3.3, where a fixed radius r is used, here we use a vanishing radius $r_n \searrow 0$ for $n \to \infty$.

If $\kappa_1 < 1/2$ is additionally satisfied then also Theorem 4.3.1 applies in this situation with $\kappa = \kappa_1$. However, in (4.10) below we see that the polynomial rate of order $\frac{q}{q+1}$ is always faster than the polynomial rate of order $\frac{q\alpha}{q\alpha+\beta}$ provided by Theorem 4.3.1.

Analogously to Lemma 4.3.3, we consider two extreme feature selection regimes. In Point (i) only the cardinality of the feature set I_n matters and in Point (ii) only one good feature is used. However, this good feature depends on n such that infinitely many features are used in total. In contrast, if $\kappa_{i_0} = 0$ holds true for some $i_0 \geq 1$ then we can even prove the same learning rate if we consider the fixed feature set $I = \{i_0\}$ for all $n \geq 1$. This case is excluded since it is not really an infinite-dimensional problem.

Again, analogously to Theorem 4.3.1 and Lemma 4.3.3, if the application at hand provides us the features $I'_n = [d'_n]$ then the results of Point (i) and (ii) can be applied if

$$d'_n \geq m_1^*(n/\tau) \qquad \text{and} \qquad d'_n \geq m_2^*(n/\tau)$$

is satisfied, respectively. Moreover, if the behavior of $(\kappa_i)_{i\geq 1}$ becomes more favorable, i.e. $(\kappa_i)_{i\geq 1}$ decreases faster, then the functions $m_i^*(x)$, $i = 1, 2$, increase slower for $x \to \infty$, i.e. we need less information per data point to obtain the desired learning rate.

Proof. Let us fix a cubic partition $\mathcal{A}$ of $\mathbb{R}^{\mathbb{N}}$ with radius $r > 0$ and a feature set $I \in \mathcal{F}(\mathbb{N})$. According to Point (ii) of Lemma 4.1.4 the assumption of Corollary 4.1.3 is satisfied and hence we find

$$MN_I(r) \leq \prod_{i\in I} (2\kappa_i + 2r/\sigma)^q \qquad \text{and} \qquad |\mathcal{A}_{I,\nu}| \leq 2 \cdot \big(1 + \lceil \sigma/(2r) \rceil\big)^{|I|} \ .$$

Now, we consider Point (i) and (ii) separately.

(i) Let I be a feature set of cardinality $m := |I| \geq 1$. Then the monotonicity of $(\kappa_i)_{i\geq 1}$ ensures

$$\begin{aligned} MN_I(\bar{\kappa}_m) &\leq \prod_{i\in I}(2\kappa_i + 2\bar{\kappa}_m/\sigma)^q \\ &\leq \prod_{i=1}^{m}(2\kappa_i + 2\bar{\kappa}_m/\sigma)^q \\ &= (2\bar{\kappa}_m)^{qm}\prod_{i=1}^{m}(1/\sigma + \kappa_i/\bar{\kappa}_m)^q \quad . \end{aligned}$$

Since the geometric mean is bounded by the arithmetic mean, we find

$$\prod_{i=1}^{m}(1/\sigma + \kappa_i/\bar{\kappa}_m)^{1/m} \leq \frac{1}{m}\sum_{i=1}^{m}\Big(1/\sigma + \frac{\kappa_i}{\bar{\kappa}_m}\Big) = 1/\sigma + \frac{\bar{\kappa}_m}{\bar{\kappa}_m} = 1/\sigma + 1 \quad .$$

Moreover, we have

$$\frac{\bar{\kappa}_m}{\bar{\kappa}_{m+1}} = \frac{m+1}{m}\cdot\frac{\sum_{i=1}^{m}\kappa_i}{\kappa_{m+1}+\sum_{i=1}^{m}\kappa_i} \leq \frac{m+1}{m} \leq 2 \quad .$$

Combining all three bounds we get

$$MN_I(\bar{\kappa}_m) \leq (2\bar{\kappa}_m)^{qm}(1/\sigma+1)^{qm} \leq (4\bar{\kappa}_{m+1})^{qm}(1/\sigma+1)^{qm} =: a_m \quad .$$

Next, we consider the number of relevant cells. Since $(\kappa_i)_{i\geq 1}$ is non-increasing, we have $2\bar{\kappa}_m \leq 2\kappa_1 < 1$ and hence, for $r = \bar{\kappa}_m$, we find

$$|\mathcal{A}_{I,\nu}| \leq 2\cdot\big(1+\lceil\sigma/(2\bar{\kappa}_m)\rceil\big)^m \leq 2\cdot\Big(\frac{\sigma+2}{2\bar{\kappa}_m}\Big)^m =: 2\cdot b_m \quad .$$

It remains to investigate the properties of the sequences $(a_m)_{m\geq 1}$ and $(b_m)_{m\geq 1}$. Since $\kappa_i \to 0$ for $i\to\infty$, the arithmetic mean $\bar{\kappa}_m \to 0$ vanishes for $m\to\infty$ as well. As a result, we find $a_m \to 0$ for $m\to\infty$. Again, using $\bar{\kappa}_m \to 0$ for $m\to\infty$ we get

$$\frac{\log(b_m)}{-\log(a_{m-1})} \to 1/q =: \rho < \infty$$

for $m \to \infty$. As a result, Lemma 4.2.1 gives $x_1 \geq 1$, $m_1^*\colon [x_1, \infty) \to \mathbb{N}$, and a function $f_1\colon [x_1, \infty) \to (0, \infty)$ increasing sub-polynomially such that in combination with Corollary 3.2.7 we get the desired bound for feature sets $I \in \mathcal{F}(\mathbb{N})$ with $|I| = m_1^*(n/\tau)$ and cubic partitions with radius $r = \bar{\kappa}_{m_1^*(n/\tau)}$.

(ii) In this case we do not use Lemma 4.2.1. We define $m_2^*(x)$ as the minimal integer $m \geq 1$ with

$$\kappa_m \leq x^{-\frac{1}{q+1}}$$

for $x \geq x_2 := 1$. Now, let $I = \{i\}$ be a singleton with $i \geq m := m_2^*(n/\tau)$ and $r = (\tau/n)^{\frac{1}{q+1}}$. As a direct consequence we have $\kappa_i \leq \kappa_m \leq (\tau/n)^{\frac{1}{q+1}} = r$. Then we find

$$\begin{aligned} MN_I(r) &\leq (2\kappa_i + 2r/\sigma)^q \leq 2^q(1 + 1/\sigma)^q \cdot r^q \qquad \text{and} \\ |\mathcal{A}_{I,\nu}| &\leq (4 + \sigma) \cdot r^{-1} \quad , \end{aligned}$$

where we used $r \leq 1$. Together with Corollary 3.2.7 we get

$$\begin{aligned} \mathcal{R}_{\text{Class},P}(h_{D,r,I}) - \mathcal{R}^*_{\text{Class},P} &\leq 6 \cdot 2^q(1 + 1/\sigma)^q \cdot r^q + C(4 + \sigma)\frac{\tau}{nr} \\ &= C_2 \cdot \left(\frac{\tau}{n}\right)^{\frac{q}{q+1}} \end{aligned}$$

with probability P^n not less than $1 - e^{-\tau}$ for $C_2 := 6 \cdot 2^q(1+1/\sigma)^q + C(4+\sigma)$. As a result, the assertion is proven. □

Before we continue with the negative results let us briefly summarize the results so far. In Theorem 4.3.1 we have seen that if $(\kappa_i)_{i\geq 1}$ is bounded away from 1/2 and $(q_i)_{i\geq 1}$ is bounded away from 0 then we get polynomial learning rates. To this end, we only have to use an appropriate number of features $|I_n|$, depending on the data set size n, for training. Moreover, in Lemma 4.3.3 we have seen that if $q_i \nearrow \infty$ is additionally satisfied, we even reach the optimal rate n^{-1}. Finally, in Lemma 4.3.4 we have seen that if $\kappa_i \searrow 0$ is additionally satisfied, we reach the improved rate $n^{-\frac{q}{q+1}}$. In all these situations, the more favorable the parameter sequences behave, the less information per data point is needed to obtain the desired rate.

Now, we continue with the negative results. Note that these results are always of the form: Under some conditions Corollary 3.2.7 does not provide polynomial learning rates for the histogram. Consequently, we prove—in some sense—lower bounds on Corollary 3.2.7 and not on the learning problems themselves. However, these results are already not a good omen for polynomial learning rates.

4.3.5 Lemma ($\boldsymbol{\kappa' \leq \kappa_i \leq \kappa}$ and $\boldsymbol{q_i \searrow 0}$) *Let Assumption 4.1.1 be satisfied for $\sigma_i = \sigma$, $\kappa' \leq \kappa_i \leq \kappa$ for all $i \geq 1$ with some $\sigma > 0$ and $0 < \kappa' \leq \kappa < 1/2$ as well as $q_i \searrow 0$ for $i \to \infty$ with*

$$\sum_{i\geq 1} q_i = \infty \ .$$

Then Corollary 3.2.7 does not give polynomial learning rates for any choice of cubic partition and feature set sequence.

Note that $\sum_{i\geq 1} q_i = \infty$ is a condition which prevents the sequence $(q_i)_{i\geq 1}$ from decreasing too fast. This condition is essential since it ensures that $\operatorname{supp} \nu_+ \cap \operatorname{supp} \nu_-$ is a ν-zero set and hence the representation of the margin-noise function of Corollary 4.1.3 applies. If $(q_i)_{i\geq 1}$ decreases too fast, i.e. $\sum_{i\geq 1} q_i < \infty$, our intuition says that the situation is even worse and hence we would not expect polynomial learning rates from Corollary 3.2.7.

Proof. The proof is an application of Lemma 4.2.2. Let $(r_n)_{n\geq 1}$ and $(I_n)_{n\geq 1}$ be sequences with $r_n > 0$ and $I_n \in \mathcal{F}(\mathbb{N})$ for all $n \geq 1$ and $MN_{I_n}(r_n) \to 0$ for $n \to \infty$. Moreover, let $\mathcal{A}$ be a cubic partition with radius r_n. According to Point (iii) of Lemma 4.1.4 the assumption of Corollary 4.1.3 is satisfied and hence we find

$$\begin{aligned} MN_{I_n}(r_n) &\geq \min\{2\kappa' + 2r_n/\sigma, 1\}^{\sum_{i=1}^{\ell_n} q_i} \qquad \text{and} \\ |\mathcal{A}_{I_n,\nu}| &\geq (\sigma/(2r_n))^{\ell_n} =: b_n \ , \end{aligned}$$

where we additionally used the monotonicity of $(q_i)_{i\geq 1}$ and the notation $\ell_n := |I_n|$. Since $MN_{I_n}(r_n) \to 0$ for $n \to \infty$, there are infinitely many $n \geq 1$ with $2\kappa' + 2r_n/\sigma \leq 1$. By transitioning to a subsequence we can assume that $2\kappa' + 2r_n/\sigma \leq 1$ holds for all $n \geq 1$. The lower bound on the margin-noise function implies two further lower bounds. First, we find

$$(2\kappa')^{\sum_{i=1}^{\ell_n} q_i} \leq MN_{I_n}(r_n) \to 0 \ ,$$

which yields $\ell_n \to \infty$ for $n \to \infty$. Second, we have

$$MN_{I_n}(r_n) \geq (2r_n/\sigma)^{\sum_{i=1}^{\ell_n} q_i} =: a_n \ ,$$

which implies

$$\frac{\log(b_n)}{-\log(a_n)} = \frac{\ell_n \cdot \log\bigl(\sigma/(2r_n)\bigr)}{\log\bigl(\sigma/(2r_n)\bigr)\sum_{i=1}^{\ell_n} q_i} = \frac{1}{\frac{1}{\ell_n}\sum_{i=1}^{\ell_n} q_i} \ .$$

Since $\ell_n \to \infty$ for $n \to \infty$ and the arithmetic mean $\frac{1}{\ell_n}\sum_{i=1}^{\ell_n} q_i$ has the same limit as the sequence $(q_i)_{i\geq 1}$ itself, the right hand side converges to ∞ for $n \to \infty$. Consequently, we found a subsequence for which the left hand side converges to ∞ and hence the limes superior of the left hand is infinite. Together with Lemma 4.2.2 we get the assertion. □

Finally, we present a second negative result.

4.3.6 Lemma ($\kappa_i \nearrow 1/2$ and $q \leq q_i \leq q'$) *Let Assumption 4.1.1 be satisfied for $\sigma_i = \sigma$, $\kappa_i > 0$, $q \leq q_i \leq q'$ for all $i \geq 1$ with some $\sigma > 0$ and $0 < q' \leq q < \infty$ as well as $\kappa_i \nearrow 1/2$ for $i \to \infty$ with*

$$\sum_{i\geq 1} 1 - 2\kappa_i = \infty \ .$$

Then Corollary 3.2.7 does not give polynomial learning rates for any choice of cubic partition and feature set sequence.

The condition $\sum_{i\geq 1} 1 - 2\kappa_i = \infty$ prevents the sequence $(\kappa_i)_{i\geq 1}$ from increasing too fast. If $(\kappa_i)_{i\geq 1}$ increases too fast, i.e. $\sum_{i\geq 1} 1 - 2\kappa_i < \infty$, analogously to Lemma 4.3.5, our intuition says that the situation is even worse and hence we would not expect polynomial learning rates from Corollary 3.2.7.

Proof. The proof is an application of Lemma 4.2.2. Let $(r_n)_{n\geq 1}$ and $(I_n)_{n\geq 1}$ be sequences with $r_n > 0$ and $I_n \in \mathcal{F}(\mathbb{N})$ for all $n \geq 1$ and $MN_{I_n}(r_n) \to 0$ for $n \to \infty$. Moreover, let $\mathcal{A}$ be a cubic partition with radius r_n. According to Point (iv) of Lemma 4.1.4 the assumption of Corollary 4.1.3 is satisfied and hence we find

$$\begin{aligned} MN_{I_n}(r_n) &\geq \prod_{i=1}^{|I_n|} \min\bigl\{2\kappa_i + 2r_n/\sigma, 1\bigr\}^{q'} \qquad \text{and} \\ |\mathcal{A}_{I_n,\nu}| &\geq \bigl\lceil \sigma/(2r_n)\bigr\rceil^{|I_n|} \quad , \end{aligned}$$

where we additionally used the monotonicity of $(\kappa_i)_{i\geq 1}$ and $q_i \leq q'$. Let $1 \leq \ell_n \leq |I_n|$ be the maximal integer with $2\kappa_{\ell_n} + 2r_n/\sigma < 1$. If there is no such integer, we set $\ell_n := 0$. With this notation we find

$$MN_{I_n}(r_n) \geq \prod_{i=1}^{\ell_n} (2\kappa_i + 2r_n/\sigma)^{q'} \qquad \text{and} \qquad |\mathcal{A}_{I_n,\nu}| \geq \bigl(\sigma/(2r_n)\bigr)^{\ell_n} \quad .$$

The lower bound on the margin-noise function implies

$$MN_{I_n}(r_n) \geq \prod_{i=1}^{\ell_n} (2\kappa_i)^{q'} =: a_n \quad .$$

Since $0 < \kappa_1 \leq \kappa_i$ holds true, this implies $(2\kappa_1)^{q'\ell_n} \leq MN_{I_n}(r_n) \to 0$ and hence $\ell_n \to \infty$ for $n \to \infty$. Consequently, we can assume $\ell_n \geq 1$ without loss of generality. As a result, we have $2\kappa_{\ell_n} + 2r_n/\sigma \leq 1$ and hence $2r_n/\sigma \leq 1 - 2\kappa_1$ for all $n \geq 1$. This implies

$$|\mathcal{A}_{I_n,\nu}| \geq (1 - 2\kappa_1)^{-\ell_n} =: b_n \quad .$$

All together we find

$$\frac{\log(b_n)}{-\log(a_n)} = \frac{\log\bigl(1/(1-2\kappa_1)\bigr)}{q'} \cdot \frac{1}{\frac{1}{\ell_n}\sum_{i=1}^{\ell_n} \log\bigl(1/(2\kappa_i)\bigr)} \quad .$$

Since $\ell_n \to \infty$ for $n \to \infty$ and the arithmetic mean $\frac{1}{\ell_n}\sum_{i=1}^{\ell_n} \log\bigl(1/(2\kappa_i)\bigr)$ has the same limit as the sequence itself, the right hand side converges to infinity for $n \to \infty$. Together with Lemma 4.2.2 we get the assertion. □

4.4 Choosing the Radius r

In this section we investigate choices for the (fixed) radius r in Theorem 4.3.1 that approximately maximize the polynomial order of the learning rate presented in that theorem. To this end, we choose the fixed scaling $\sigma = 1$ throughout this section and recall that Theorem 4.3.1 provides the polynomial learning rate of order

$$\frac{q\alpha}{q\alpha + \beta} = \frac{q}{q + f_\kappa(r)}$$

with

$$f_\kappa(r) := \frac{\beta}{\alpha} = \frac{\log\bigl(1 + \lceil 1/(2r) \rceil\bigr)}{\log\bigl(1/(2\kappa + 2r)\bigr)} \quad , \tag{4.7}$$

which depends on κ, q, and r. In contrast to the parameters κ and q, which are given by the learning scenario, the radius $0 < r < 1/2 - \kappa$ can be chosen by the user. Since the smaller $f_\kappa(r)$ the faster our learning rate, an optimal choice for r minimizes $f_\kappa(r)$ over $0 < r < 1/2 - \kappa$. Note that an optimal choice for r depends only on κ and is independent of q. For the function mapping κ to the optimal value we write $F\colon [0, 1/2) \to (0, \infty)$ and set

$$F(\kappa) := \inf_{0<r<1/2-\kappa} f_\kappa(r) \quad . \tag{4.8}$$

In this section we use Lambert's W-function W_{-1}, which is defined as the inverse of $t \mapsto te^t$ for $t \le -1$. For a brief introduction to Lambert's W-function see e.g. Section 7.2 of Part II.

The goal of this section is to prove that the choices

$$\begin{aligned} r_1(\kappa) &:= -\frac{b_\kappa/2}{W_{-1}(-b_\kappa/e)} \qquad \text{and} \\ r_2(\kappa) &:= -\frac{b_\kappa/2}{(1+2b_\kappa)W_{-1}\big(-\frac{b_\kappa/e}{1+2b_\kappa}\big)+2b_\kappa} \end{aligned} \tag{4.9}$$

for the radius r with $b_\kappa := 2\kappa\log\big(1/(2\kappa)\big)$ are *optimal* in some sense. To be more precise, we show the strong asymptotic equivalence $f_\kappa(r_i(\kappa)) \sim F(\kappa)$ for $\kappa \to 0^+$ as well as $\kappa \to (1/2)^-$, i.e. $f_\kappa(r_i(\kappa))/F(\kappa) \to 1$ in both cases. Moreover, for a visualization of these radii and functions see Figure 4.4. However, let us start with some basic properties of the functions f_κ and F.

4.4.1 Lemma (Basic Properties) *The functions* $f_\kappa\colon (0,1/2-\kappa) \to (0,\infty)$ *for* $0 \le \kappa < 1/2$ *and* $F\colon [0,1/2) \to (0,\infty)$ *defined in* (4.7) *and* (4.8), *respectively, have the following properties:*

(i) The function $F\colon [0,1/2) \to (0,\infty)$ *is non-decreasing and equals*

$$F(\kappa) = \inf_{\substack{m\in\mathbb{N}:\\ 1/(2m)<1/2-\kappa}} f_\kappa\big(1/(2m)\big) \ .$$

(ii) For $0 < \kappa < 1/2$ *there is a minimizer* $0 < r^*(\kappa) < 1/2-\kappa$ *with* $F(\kappa) = f_\kappa\big(r^*(\kappa)\big)$.

(iii) For $\kappa = 0$ *there is no minimizer, i.e.* $F(0) < f_0(r)$ *for all* $0 < r < 1/2$, *and* $F(0) = 1$.

Since F is non-decreasing, the order of the polynomial learning rate of Theorem 4.3.1 (for an optimal choice of the radius $r = r^*(\kappa)$) becomes better for decreasing κ. If we combine Point (i) and (iii), we get the first lower bound, namely for all $0 < \kappa < 1/2$ we have

$$F(\kappa) \ge F(0) = 1 \ . \tag{4.10}$$

Proof. (i) Let $0 \le \kappa < \kappa' < 1/2$ be fixed. Since $1/2-\kappa' < 1/2-\kappa$ holds true, for every $r < 1/2-\kappa'$ where $f_{\kappa'}(r)$ is defined $f_\kappa(r)$ is defined as well.

Moreover, for fixed $r < 1/2 - \kappa'$ we have $f_{\kappa'}(r) \geq f_\kappa(r)$. All together this proves the monotonicity of F, namely

$$F(\kappa') = \inf_{r<1/2-\kappa'} f_{\kappa'}(r) \geq \inf_{r<1/2-\kappa'} f_\kappa(r) \geq \inf_{r<1/2-\kappa} f_\kappa(r) = F(\kappa) \ .$$

Next, let us fix some $0 \leq \kappa < 1/2$ and $m \in \mathbb{N}$ with $1/(2m) < 1/2 - \kappa$. Then, for $1/(2m) \leq r < 1/(2(m-1))$, the numerator in the definition of $f_\kappa(r)$ depends only on m and not on r

$$f_\kappa(r) = \frac{\log(1+m)}{\log\bigl(1/(2\kappa + 2r)\bigr)} \ .$$

Consequently, if we minimize $f_\kappa(r)$ over $1/(2m) \leq r < 1/(2(m-1))$ we find

$$\min_{1/(2m) \leq r < 1/(2(m-1))} f_\kappa(r) = f_\kappa\bigl(1/(2m)\bigr) \ .$$

As a result, it is enough to take the infimum over all integers $m \in \mathbb{N}$ with $1/(2m) < 1/2 - \kappa$.

(ii) Let us fix some $0 < \kappa < 1/2$. We directly see that the denominator in the definition of f_κ converges $\log\bigl(1/(2\kappa + 2r)\bigr) \to \log\bigl(1/(2\kappa)\bigr) \in (0, \infty)$ for $r \to 0$ and hence $f_\kappa(r) \to \infty$ for $r \to 0$. Consequently, for $M \coloneqq 2F(\kappa) \geq 2$ there is some $r_0 < 1/2 - \kappa$ with $f_\kappa(r) > M$ for all $r \leq r_0$ and we can further restrict the infimum in the representation of $F(\kappa)$ in Point (i)

$$F(\kappa) = \inf_{\substack{m \in \mathbb{N}: \\ r_0 < 1/(2m) < 1/2 - \kappa}} f_\kappa\bigl(1/(2m)\bigr) \ .$$

Since there are only finitely many integers $m \in \mathbb{N}$ with $r_0 \leq 1/(2m) < 1/2 - \kappa$, the infimum is actually a minimum and the assertion is proven.

(iii) For $\kappa = 0$, we find, for $r \to 0$,

$$f_0(r) = \frac{\log\bigl(1 + \lceil 1/(2r) \rceil\bigr)}{\log\bigl(1/(2r)\bigr)} \to 1 \ .$$

Since $f_0(r) > 1$ for $r < 1/2$, there is no minimizer and we have $F(0) = 1$. □

The following auxiliary result is the basis for all our bounds.

4.4.2 Lemma (Auxiliary Result) *Let the function $g_{a,b}\colon \big(\max\{1/b, 1\}, \infty\big) \to (0, \infty)$ be defined by*

$$g_{a,b}(x) := \frac{(x+a)\log(x+a)}{xb-1} \quad ,$$

for $a \geq 0$ and $b > 0$. Then the following statements are true:

(i) $g_{a,b}(x) \to \infty$ *for* $x \to \infty$.

(ii) If $b > 1 + ab$ then $g_{a,b}$ is increasing.

(iii) If $b \leq 1 + ab$ then

$$\begin{aligned} x^*(a,b) &:= \exp\Big(-W_{-1}\Big(-\frac{b/e}{1+ab}\Big) - 1\Big) - a \\ &= -\frac{1+ab}{b} W_{-1}\Big(-\frac{b/e}{1+ab}\Big) - a \end{aligned}$$

is well-defined and $x^*(a,b) \geq \max\{1/b, 1\}$. Moreover, $g_{a,b}$ is decreasing for $x < x^*(a,b)$, increasing for $x > x^*(a,b)$, and has a global minimum at $x = x^*(a,b)$ with

$$g_{a,b}(x^*(a,b)) = \frac{x^*(a,b)+a}{1+ab} = -\frac{1}{b} \cdot W_{-1}\Big(-\frac{b/e}{1+ab}\Big) \quad .$$

Proof. First note that $x > 1/b$ ensures that the denominator in the definition of $g_{a,b}(x)$ is positive. Moreover, $x \geq 1$ and $a \geq 0$ ensure that the numerator is positive as well.

(i) The first assertion is a consequence of the fact, that $g_{a,b}(x)$ is the product of a rational function, which converges to $1/b$ for $x \to \infty$, and a logarithm, which converges to ∞ for $x \to \infty$.

Before we continue with the proof of Point (ii) and (iii), we calculate the derivative of $g_{a,b}$

$$\begin{aligned} g'_{a,b}(x) &= \frac{\bigl(\log(x+a)+1\bigr)(xb-1)-(x+a)\log(x+a)b}{(xb-1)^2} \\ &= \frac{-(1+ab)\log(x+a)+(x+a)b-(1+ab)}{(xb-1)^2} \quad . \end{aligned} \tag{4.11}$$

Since the denominator is positive for all $x > 1/b$, we only have to consider the numerator in order to determine the monotonicity properties of $g_{a,b}$. Elementary transformations show

$$\begin{aligned} g'_{a,b}(x) > 0 \quad &\iff \quad (1+ab)\log\bigl(e(x+a)\bigr) < (x+a)b \\ &\iff \quad \frac{1}{e(x+a)}\log\left(\frac{1}{e(x+a)}\right) > -\frac{b}{e(1+ab)} \quad . \end{aligned} \tag{4.12}$$

(ii) Now, we consider the case $b > 1+ab$. In this case we have

$$-\frac{b}{e(1+ab)} < -\frac{1}{e} \quad .$$

The left hand side of (4.12) is the concatenation of the functions $x \mapsto \log\bigl(1/(e(x+a))\bigr)$ and $t \mapsto te^t$, and the latter is always greater or equal to $-1/e$, see Figure 7.1 of Part II for a plot of the function $t \mapsto te^t$. As a result, the condition in (4.12) is satisfied for all $x > 1 = \max\{1/b, 1\}$ and hence $g'_{a,b} > 0$ holds true. This proves the assertion.

(iii) Now, we consider the case $b \leq 1+ab$. Since we assume $x \geq 1$, we have $\log\bigl(1/(e(x+a))\bigr) \leq -1$. This ensures that we can apply the decreasing branch W_{-1} of Lambert's W-function to the inequality in (4.12). Together with some basic transformations we find

$$\begin{aligned} g'_{a,b}(x) > 0 \quad &\iff \quad \log\left(\frac{1}{e(x+a)}\right) < W_{-1}\left(-\frac{b}{e(1+ab)}\right) \\ &\iff \quad x > x^*(a,b) \ . \end{aligned}$$

Analogously, we obtain $g'_{a,b}(x) < 0$ if and only if $x < x^*(a,b)$ as well as $g'_{a,b}(x^*(a,b)) = 0$. This proves the stated monotonicity properties of $g_{a,b}$ and that $g_{a,b}$ has a global minimum at $x = x^*(a,b)$. Next, the second representation of $x^*(a,b)$ is a direct consequence of $\exp \circ W_{-1}(-t) = -t/W_{-1}(-t)$ for all $0 < t \leq 1/e$, see e.g. Lemma 7.2.4 of Part II. Finally, we calculate the minimum. Using $g'_{a,b}(x^*(a,b)) = 0$ in (4.11) yields

$$x^*(a,b)b - 1 = (1+ab)\log\bigl(x^*(a,b)+a\bigr)$$

Plugging this into the denominator of $g_{a,b}$ gives the claimed value of the minimum. □

The next lemma provides upper bounds for the functions $f_\kappa(r_i(\kappa))$ with $i = 1, 2$.

4.4.3 Lemma (Upper Bounds) *For $0 < \kappa < 1/2$, f_κ defined in (4.7), and F defined in (4.8) the following statements are satisfied:*

(i) For the radius $r_1(\kappa)$ defined in (4.9) the following upper bound is satisfied

$$f_\kappa(r_1(\kappa)) \leq -\frac{1}{\log\bigl(1/(2\kappa)\bigr)} W_{-1}(-b_\kappa/e) \cdot \frac{\log\bigl(1 + \lceil 1/(2r_1(\kappa))\rceil\bigr)}{\log\bigl(1/(2r_1(\kappa))\bigr)} \ .$$

(ii) For the radius $r_2(\kappa)$ defined in (4.9) the following upper bound is satisfied

$$f_\kappa(r_2(\kappa)) \leq -\frac{1}{\log\bigl(1/(2\kappa)\bigr)} W_{-1}\left(-\frac{b_\kappa/e}{1+2b_\kappa}\right) \ .$$

Note that the case $\kappa = 0$ is excluded in this lemma and that each statement provide a bound on F via $F(\kappa) \leq f_\kappa(r_i(\kappa))$ for $i = 1, 2$.

Since there are explicit bounds for the function W_{-1} in the literature, see e.g. [20], these bounds can be made more concrete if necessary. Moreover, for a visualization of these bounds see Figure 4.4.

Proof. Our strategy is to linearize the logarithm in the denominator of f_κ at $1/(2\kappa)$ and then determine the minimum explicitly using Lemma 4.4.2.

This approach shows that the choice r_i given by (4.9) essentially coincides with the minimizer given by Lemma 4.4.2.

Since the logarithm is a concave function, we can estimate the denominator of $f_\kappa(r)$ by

$$\begin{aligned}\log\left(\frac{1}{2\kappa+2r}\right) &= -\log(2\kappa+2r)\\ &\geq -\Big(\big((2\kappa+2r)-2\kappa\big)\log'(2\kappa)+\log(2\kappa)\Big)\\ &= \log\big(1/(2\kappa)\big)-2r/(2\kappa)\ .\end{aligned}$$

The right hand side is positive if and only if $r < b_\kappa/2$ is satisfied, where $b_\kappa = 2\kappa\log\big(1/(2\kappa)\big)$. Using $\log(1+t)\leq t$ we find

$$b_\kappa = 2\kappa\log\Big(\frac{1}{2\kappa}\Big) \leq 2\kappa\Big(\frac{1}{2\kappa}-1\Big) = 1-2\kappa$$

and hence $b_\kappa/2 \leq 1/2-\kappa$. Since $f_\kappa(r)$ is defined for $r < 1/2-\kappa$, we can apply the above bound for $r < b_\kappa/2$. This gives

$$f_\kappa(r) \leq 2\kappa\cdot\frac{\log\big(1+\lceil 1/(2r)\rceil\big)}{2rb_\kappa-1} \tag{4.13}$$

for all $r < b_\kappa/2$. Now, we continue this estimate differently for r_1 and r_2.

(i) For r_1 we rewrite the right hand side of (4.13) using the function g_{0,b_κ} defined in Lemma 4.4.2 for $a=0$ and $b=b_\kappa$

$$\begin{aligned}f_\kappa(r) &\leq 2\kappa\cdot\frac{1/(2r)\cdot\log\big(1+\lceil 1/(2r)\rceil\big)}{b_\kappa/(2r)-1}\\ &= 2\kappa\cdot g_{0,b_\kappa}\big(1/(2r)\big)\cdot\frac{\log\big(1+\lceil 1/(2r)\rceil\big)}{\log\big(1/(2r)\big)}\ .\end{aligned}$$

Note that using derivatives we find $0 < b_\kappa \leq 1/e$ for all $0<\kappa<1/2$ and hence the monotonicity properties of g_{0,b_κ} are described in Point (iii) of

Lemma 4.4.2. According to Lemma 4.4.2

$$x^*(0, b_\kappa) = -\frac{W_{-1}(-b_\kappa/e)}{b_\kappa} = \frac{1}{2r_1(\kappa)}$$

is a minimizer of g_{0,b_κ}. Plugging in the minimum provided by Lemma 4.4.2 gives the desired upper bound on $f_\kappa(r_1(\kappa))$.

(ii) For this statement we continue the estimate in (4.13) differently than for Point (i). Since

$$f_\kappa(r) \leq 2\kappa \cdot \frac{(1/(2r) + 2)\log\bigl(1/(2r) + 2\bigr)}{b_\kappa/(2r) - 1} = 2\kappa \cdot g_{2,b_\kappa}\bigl(1/(2r)\bigr)$$

holds true, again Point (iii) of Lemma 4.4.2 describes the monotonicity properties of the right hand side. According to Lemma 4.4.2

$$x^*(2, b_\kappa) = -\frac{1 + 2b_\kappa}{b_\kappa} \cdot W_{-1}\left(-\frac{b_\kappa/e}{1 + 2b_\kappa}\right) - 2 = \frac{1}{2r_2(\kappa)}$$

is a minimizer of g_{2,b_κ}. Plugging in the minimum provided by Lemma 4.4.2 gives the desired upper bound on $f_\kappa(r_2(\kappa))$. □

In order to prove the optimality of our choices $r_i(\kappa)$ for $\kappa \to (1/2)^-$ we use a lower bound on F which is provided by the following lemma.

4.4.4 Lemma (Lower Bound) *For $1/(2e) < \kappa < 1/2$, F defined in (4.8), and b_κ defined after (4.9) the following lower bound is satisfied*

$$F(\kappa) \geq -\frac{1}{\log\bigl(1/(2\kappa)\bigr)} W_{-1}\left(-\frac{b_\kappa/e}{1 - \log\bigl(1/(2\kappa)\bigr) + b_\kappa}\right) \ .$$

For $\kappa \to (1/(2e))^+$, we have $b_\kappa = 2\kappa \cdot \log\bigl(1/(2\kappa)\bigr) \to 1/e$ and hence this lower bound implies

$$\begin{aligned} F(\kappa) &\geq -\frac{1}{\log\bigl(1/(2\kappa)\bigr)} \cdot W_{-1}\left(-\frac{b_\kappa/e}{1 - \log\bigl(1/(2\kappa)\bigr) + b_\kappa}\right) \\ &\to -W_{-1}\left(-\frac{1/e^2}{1 - 1 + 1/e}\right) = 1 \ . \end{aligned}$$

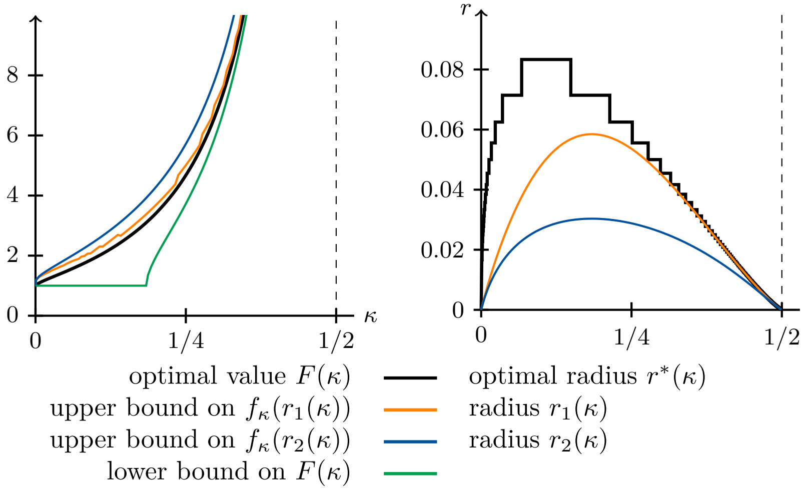


Figure 4.4: The left plot shows the (numerically calculated) optimal value $F(\kappa)$ together with the bounds from (4.10), Lemma 4.4.3, and Lemma 4.4.4. The right plot shows an (numerically calculated) optimal radius $r^*(\kappa)$ together with the radii defined in (4.9). Note that $\kappa \mapsto r^*(\kappa)$ is a step function because $1/(2r^*(\kappa)) \in \mathbb{N}$ according to Point (i) of Lemma 4.4.1.

Consequently, at $\kappa = 1/(2e)$ this lower bound coincides with the lower bound in (4.10). For a visualization of both bounds see Figure 4.4.

Proof. Analogously to the proof of the upper bounds in Lemma 4.4.3, we use the concavity of the logarithm to get a bound for the denominator of $f_\kappa(r)$ by its linearization, namely

$$\begin{aligned}\log\left(\frac{1}{2\kappa+2r}\right) &\leq \left(\frac{1}{2\kappa+2r} - \frac{1}{2\kappa}\right)\log'(1/(2\kappa)) + \log\left(\frac{1}{2\kappa}\right) \\ &= \log\left(\frac{1}{2\kappa}\right) - \frac{2r}{2\kappa+2r} \quad .\end{aligned}$$

Together with $\lceil 1/(2r) \rceil \geq 1/(2r)$, $2\kappa \leq 1$, and $b_\kappa = 2\kappa \log\bigl(1/(2\kappa)\bigr)$ we get

$$f_\kappa(r) \geq \frac{(2\kappa + 2r)\log\bigl(1 + \lceil 1/(2r) \rceil\bigr)}{(2\kappa + 2r)\log\bigl(1/(2\kappa)\bigr) - 2r} \geq 2\kappa \cdot \frac{(1 + 2r)\log\bigl(1 + 1/(2r)\bigr)}{b_\kappa - 2r\bigl(1 - \log(1/(2\kappa))\bigr)} \quad .$$

Using the function g_{1,c_κ} defined in Lemma 4.4.2 with

$$c_\kappa := \frac{b_\kappa}{1 - \log\bigl(1/(2\kappa)\bigr)} = \frac{2\kappa \log\bigl(1/(2\kappa)\bigr)}{1 - \log\bigl(1/(2\kappa)\bigr)}$$

this reads

$$f_\kappa(r) \geq \frac{2\kappa}{1 - \log\bigl(1/(2\kappa)\bigr)} \cdot g_{1,c_\kappa}\bigl(1/(2r)\bigr)$$

for $r < 1/2 - \kappa$. Since we want to apply Lemma 4.4.2, we need to assume $c_\kappa > 0$ which is equivalent to $\kappa > 1/(2e)$. Moreover, since $a = 1$ Point (iii) of Lemma 4.4.2 applies and plugging in the minimum provided by Lemma 4.4.2 proves the desired lower bound. □

The final lemma combines the lower bounds in Lemma 4.4.4 and (4.10) with the upper bounds in Lemma 4.4.3 to establish the optimality of our choices r_i in (4.9) for $\kappa \to 0^+$ and $\kappa \to (1/2)^-$.

4.4.5 Lemma (Asymptotic Behavior) *For F defined in* (4.8) *and r_i defined in* (4.9) *the following strong asymptotic equivalences are satisfied for $i = 1, 2$: $F(\kappa) \sim f_\kappa(r_i(\kappa)) \to 1$ for $\kappa \to 0^+$ and*

$$F(\kappa) \sim f_\kappa(r_i(\kappa)) \sim \frac{1}{1 - 2\kappa} \log\Bigl(\frac{1}{1 - 2\kappa}\Bigr) \to \infty$$

for $\kappa \to (1/2)^-$.

Proof. Let us recall some basic facts about the asymptotic behavior of Lambert's W-function and the logarithm, which we will use several times in this proof,

$$-W_{-1}(-t) \sim \log(1/t) \quad \text{for } t \to 0^+ \qquad \text{and} \qquad \log(1+t) \sim t \quad \text{for } t \to 0 \ ,$$

see e.g. Lemma 7.2.4 of Part II for details about Lambert's W-function. Moreover, note that the quantity $b_\kappa = 2\kappa \cdot \log\bigl(1/(2\kappa)\bigr) \to 0$ vanishes for $\kappa \to 0^+$ as well as for $\kappa \to (1/2)^-$.

First, we consider the radius r_1. Since $b_\kappa \to 0$ holds true, we find

$$r_1(\kappa) = -\frac{b_\kappa/2}{W_{-1}(-b_\kappa/e)} \sim \frac{b_\kappa/2}{\log(e/b_\kappa)} \to 0$$

for $\kappa \to 0^+$ as well as for $\kappa \to (1/2)^-$. As a result, the upper bound in Lemma 4.4.3 for r_1 gives

$$\begin{aligned} F(\kappa) \le f_\kappa(r_1(\kappa)) &\le \frac{-W_{-1}(-b_\kappa/e)}{\log\bigl(1/(2\kappa)\bigr)} \cdot \frac{\log\bigl(1 + \lceil 1/(2r_1(\kappa))\rceil\bigr)}{\log\bigl(1/(2r_1(\kappa))\bigr)} \\ &\sim \frac{\log(e/b_\kappa)}{\log\bigl(1/(2\kappa)\bigr)} \end{aligned} \tag{4.14}$$

for $\kappa \to 0^+$ as well as for $\kappa \to (1/2)^-$. Next, we investigate the right hand side of (4.14). To this end, we write

$$\frac{\log(e/b_\kappa)}{\log\bigl(1/(2\kappa)\bigr)} = \frac{1}{\log\bigl(1/(2\kappa)\bigr)} + 1 - \frac{\log\bigl(\log(1/(2\kappa))\bigr)}{\log\bigl(1/(2\kappa)\bigr)} \quad . \tag{4.15}$$

For $\kappa \to 0^+$, we have $\log\bigl(1/(2\kappa)\bigr) \to \infty$ and hence

$$\frac{\log(e/b_\kappa)}{\log\bigl(1/(2\kappa)\bigr)} \to 1 \quad . \tag{4.16}$$

For $\kappa \to (1/2)^-$, we have $\log\bigl(1/(2\kappa)\bigr) \sim 1/(2\kappa) - 1 \sim 1 - 2\kappa \to 0$ and hence the third term in (4.15) is dominating. This gives, for $\kappa \to (1/2)^-$,

$$\begin{aligned} \frac{\log(e/b_\kappa)}{\log\bigl(1/(2\kappa)\bigr)} &\sim \frac{1}{\log\bigl(1/(2\kappa)\bigr)} \log\left(\frac{1}{\log\bigl(1/(2\kappa)\bigr)}\right) \\ &\sim \frac{1}{1-2\kappa} \log\left(\frac{1}{1-2\kappa}\right) \quad . \end{aligned} \tag{4.17}$$

Second, we analyze the radius r_2 analogously. Since $b_\kappa \to 0$ holds true, we find

$$r_2(\kappa) = -\frac{b_\kappa/2}{(1+2b_\kappa)W_{-1}\big(-\frac{b_\kappa/e}{1+2b_\kappa}\big) + 2b_\kappa} \sim \frac{b_\kappa/2}{W_{-1}(-b_\kappa/e)} \to 0$$

for $\kappa \to 0^+$ as well as for $\kappa \to (1/2)^-$. As a result, the upper bound in Lemma 4.4.3 for r_2 gives

$$\begin{aligned} F(\kappa) \leq f_\kappa(r_2(\kappa)) &\leq -\frac{1}{\log\big(1/(2\kappa)\big)} W_{-1}\bigg(-\frac{b_\kappa/e}{1+2b_\kappa}\bigg) \\ &\sim \frac{\log(e/b_\kappa)}{\log\big(1/(2\kappa)\big)} + \frac{\log(1+2b_\kappa)}{\log\big(1/(2\kappa)\big)} \end{aligned}$$

for $\kappa \to 0^+$ as well as for $\kappa \to (1/2)^-$. The first term coincides with (4.14) and was already investigated. Since the second term behaves like

$$\frac{\log(1+2b_\kappa)}{\log\big(1/(2\kappa)\big)} \sim \frac{2b_\kappa}{\log\big(1/(2\kappa)\big)} = 4\kappa$$

in both cases, $\kappa \to 0^+$ and $\kappa \to (1/2)^-$, the second term does not influence the asymptotic behavior. As a result, the upper bound for $f_\kappa(r_2(\kappa))$ behaves asymptotically as the upper bound for $f_\kappa(r_1(\kappa))$ in both regimes, $\kappa \to 0^+$ and $\kappa \to (1/2)^-$.

Until now, we have only investigated the asymptotic behavior of the upper bounds on $f_\kappa(r_i(\kappa))$ with $i = 1, 2$. Consequently, it remains to establish the optimality by proving the same behavior for a lower bound. The lower bound in (4.10) gives $F(\kappa) \geq 1$ and hence, together with (4.16), the optimality of $F(\kappa) \sim f_\kappa(r_i(\kappa)) \to 1$ for $\kappa \to 0^+$ is proven.

For $\kappa \to (1/2)^-$ we use the lower bound from Lemma 4.4.4. In this case we have

$$\begin{aligned} F(\kappa) &\geq -\frac{1}{\log\big(1/(2\kappa)\big)} \cdot W_{-1}\bigg(-\frac{b_\kappa/e}{1-\log\big(1/(2\kappa)\big)+b_\kappa}\bigg) \\ &\sim \frac{\log(e/b_\kappa)}{\log\big(1/(2\kappa)\big)} + \frac{\log\big(1-\log\big(1/(2\kappa)\big)+b_\kappa\big)}{\log\big(1/(2\kappa)\big)} \end{aligned}$$

and again the behavior of the first term was already investigated. The second term behaves like

$$\frac{\log\bigl(1-\log\bigl(1/(2\kappa)\bigr)+b_\kappa\bigr)}{\log\bigl(1/(2\kappa)\bigr)} \sim \frac{b_\kappa-\log\bigl(1/(2\kappa)\bigr)}{\log\bigl(1/(2\kappa)\bigr)} = -(1-2\kappa) \to 0$$

for $\kappa \to (1/2)^-$. Consequently, the second term does not influence the asymptotic behavior and hence, together with (4.17), the optimality $F(\kappa) \sim f_\kappa(r_i(\kappa))$ and the claimed asymptotic behavior for $\kappa \to (1/2)^-$ are proven. □

Chapter 5

Generalizations

In this chapter we present various generalizations of the prototypical example that still allow to prove polynomial learning rates. To this end, we typically denote the prototypical example by $\tilde{P}$ and the classification problem of interest by P which will be in some sense a generalization of $\tilde{P}$. Although, in most cases all the upper rates of Section 4.3 for $\tilde{P}$ can be transferred to P, for convenience, we focus on the situation considered in Theorem 4.3.1, i.e. $\tilde{P}$ satisfies Assumption 4.1.1 for $\sigma_i = \sigma$, $\kappa_i \leq \kappa$ and $q_i \geq q$ for $i \geq 1$ with some $0 < \kappa < 1/2$ and $\sigma, q > 0$. Moreover, in many cases it is also possible to combine multiple generalizations but, again for convenience, we always take the assumptions of Theorem 4.3.1 as starting point for $\tilde{P}$.

5.1 Absolute Continuous Finite-Dimensional Distributions

In this section we consider two classification problems P and $\tilde{P}$ on $X \times \{\pm 1\}$, with a sequence space $X = \prod_{i \geq 1} X_i$ as defined in (3.1), which satisfy the following relation: There is a constant $C_0 \geq 1$ such that $\nu_{\pm,I} \ll \tilde{\nu}_{\pm,I}$ and

$$\left| \frac{\mathrm{d}\nu_{\pm,I}}{\mathrm{d}\tilde{\nu}_{\pm,I}} \right| \leq C_0^{|I|} \tag{5.1}$$

$\tilde{\nu}_{\pm,I}$-almost surely for all $I \in \mathcal{F}(\mathbb{N})$. Here and in the following we denote all quantities related to $\tilde{P}$ with a tilde and all quantities related to P without a tilde. For convenience, we assume $p_\pm = \tilde{p}_\pm$ throughout this section. The goal is to transfer learning rates for $\tilde{P}$ to P using (5.1). But let us start with an important class of distributions satisfying (5.1).

To this end, we assume that $X = \mathbb{R}^{\mathbb{N}}$, P satisfies Assumption 3.3.1, and $\nu_{\pm,i} \ll \lambda$ with

$$\frac{\mathrm{d}\nu_{\pm,i}}{\mathrm{d}\lambda}(t) \leq C_0 \cdot \frac{f_i(\kappa_i \pm t/\sigma_i)}{\sigma_i} \quad , \tag{5.2}$$

for λ-almost all $t \in \mathbb{R}$ and all $i \geq 1$, where κ_i, q_i, σ_i, and f_i are defined as in Assumption 4.1.1. Now, if $\tilde{P}$ satisfies Assumption 4.1.1 with the same parameters κ_i, q_i, and σ_i then (5.2) reads

$$\frac{\mathrm{d}\nu_{\pm,i}}{\mathrm{d}\lambda}(t) \leq C_0 \cdot \frac{\mathrm{d}\tilde{\nu}_{\pm,i}}{\mathrm{d}\lambda}(t)$$

for λ-almost all $t \in \mathbb{R}$. Consequently, (5.2) is equivalent to

$$\frac{\mathrm{d}\nu_{\pm,i}}{\mathrm{d}\tilde{\nu}_{\pm,i}} \leq C_0$$

$\tilde{\nu}_{\pm,i}$-almost surely for all $i \geq 1$. Since we assume that $\nu_\pm$ and $\tilde{\nu}_\pm$ are product measures, via Assumption 3.3.1, we find

$$\left|\frac{\mathrm{d}\nu_{\pm,I}}{\mathrm{d}\tilde{\nu}_{\pm,I}}\right| = \left|\frac{\mathrm{d}(\bigotimes_{i\in I} \nu_{\pm,i})}{\mathrm{d}(\bigotimes_{i\in I} \tilde{\nu}_{\pm,i})}\right| = \left|\bigotimes_{i\in I} \frac{\mathrm{d}\nu_{\pm,i}}{\mathrm{d}\tilde{\nu}_{\pm,i}}\right| \leq C_0^{|I|}$$

$\tilde{\nu}_{\pm,I}$-almost surely for all $I \in \mathcal{F}(\mathbb{N})$. This proves (5.1) and hence the following section especially generalizes the prototypical example from Assumption 4.1.1 to the class of distributions which satisfy only the inequality in (5.2) with $C_0 \geq 1$ instead of the equality with $C_0 = 1$ as required by Assumption 4.1.1.

The first lemma provides some basic implications of (5.1).

5.1.1 Lemma (Basic Properties) *Let X be a sequence space as defined in* (3.1) *and P, $\tilde{P}$ be probability distributions on $X \times \{\pm 1\}$ satisfying* (5.1). *Then the following statements are true, for $I \in \mathcal{F}(\mathbb{N})$:*

(i) $\nu_{\pm,I}(A) \leq C_0^{|I|} \tilde{\nu}_{\pm,I}(A)$ for all measurable $A \subseteq X_I$.

(ii) $\operatorname{supp} \nu_{\pm,I} \subseteq \operatorname{supp} \tilde{\nu}_{\pm,I}$.

Proof. This is a direct consequence of (5.1) and $\nu_{\pm,I} \ll \tilde{\nu}_{\pm,I}$. □

The next lemma relates the margin-noise function and the number or relevant cells of the two classification problems P and $\tilde{P}$.

5.1.2 Lemma (Upper Bounds) *Let $X = \prod_{i \geq 1} X_i$ be a sequence space as defined in* (3.1) *and P, $\tilde{P}$ be probability distributions on $X \times \{\pm 1\}$ satisfying* (5.1)*. Then the following statements are true, for $I \in \mathcal{F}(\mathbb{N})$:*

(i) If $\operatorname{supp} \nu_+ \cap \operatorname{supp} \nu_-$ is a ν_-- or ν_+-zero set and there is no noise in $\tilde{P}$ then there is a version of η such that the margin-noise functions satisfy, for every version of $\tilde{\eta}$ and $r \geq 0$,

$$MN_I(r) \leq C_0^{|I|} \tilde{MN}_I(r) \quad .$$

(ii) For a measurable partition $\mathcal{A}$ of $(X_i)_{i \geq 1}$ the relevant cells satisfy the inclusion

$$\mathcal{A}_{I,\nu} \subseteq \mathcal{A}_{I,\tilde{\nu}} \quad .$$

Proof. (i) Let $I \in \mathcal{F}(\mathbb{N})$ and $r \geq 0$ be fixed. Since $\operatorname{supp} \nu_+ \cap \operatorname{supp} \nu_-$ is a ν_+- or a ν_--zero set, Point (ii) of Lemma 2.3.4 is applicable. As a result, there is a version of η such that

$$\begin{aligned} MN_I(r) = {} & p_+ \nu_{+,I}\big(\operatorname{dist}(\,\cdot\,, \operatorname{supp} \nu_{-,I}) \leq 2r\big) \\ & + p_- \nu_{-,I}\big(\operatorname{dist}(\,\cdot\,, \operatorname{supp} \nu_{+,I}) \leq 2r\big) \quad . \end{aligned}$$

From Point (ii) of Lemma 5.1.1 we get

$$\big\{\operatorname{dist}(\,\cdot\,, \operatorname{supp} \nu_{-,I}) \leq 2r\big\} \subseteq \big\{\operatorname{dist}(\,\cdot\,, \operatorname{supp} \tilde{\nu}_{-,I}) \leq 2r\big\}$$

and together with Point (i) of Lemma 5.1.1 we find

$$\nu_{+,I}\big(\operatorname{dist}(\,\cdot\,, \operatorname{supp} \nu_{-,I}) \leq 2r\big) \leq C_0^{|I|} \tilde{\nu}_{+,I}\big(\operatorname{dist}(\,\cdot\,, \operatorname{supp} \tilde{\nu}_{-,I}) \leq 2r\big) \quad .$$

For symmetry reasons, this bound remains true if we interchange "+" and "−". Since there is no noise in $\tilde{P}$, Point (i) of Lemma 2.3.4 yields the assertion.

(ii) This statement is a direct consequence of $\nu_{\pm,I} \ll \tilde{\nu}_{I,\pm}$ and Point (iii) of Lemma 1.3.3. □

In the remaining part of this section we assume that $\tilde{P}$ satisfies Assumption 4.1.1 and transfer the bounds and learning rates presented in Section 4.1 and Section 4.3, respectively, from $\tilde{P}$ to P.

5.1.3 Lemma (No Noise) *Let $X = \mathbb{R}^{\mathbb{N}}$ be a sequence space as defined in* (3.1) *with $X_i = \mathbb{R}$ for $i \geq 1$ and P, $\tilde{P}$ be probability distributions on $X \times \{\pm 1\}$ satisfying* (5.1)*. Furthermore, let $\tilde{P}$ satisfy Assumption 4.1.1 with $\sigma_i = \sigma$, $q_i \geq q$, and $\kappa_i \leq \kappa$ for all $i \geq 1$ for some $\sigma, q > 0$, and $\kappa < C_0^{-1/q}/2$. Then the following identity is satisfied*

$$\nu(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = \tilde{\nu}(\operatorname{supp}\tilde{\nu}_+ \cap \operatorname{supp}\tilde{\nu}_-) = 0 \ .$$

Proof. Since $\tilde{P}$ satisfies Assumption 4.1.1, $\tilde{\nu}(\operatorname{supp}\tilde{\nu}_+ \cap \operatorname{supp}\tilde{\nu}_-) = 0$ is already stated in Point (ii) of Lemma 4.1.4. Now, let us fix some $I \in \mathcal{F}(\mathbb{N})$. For symmetry reasons it is enough to show $\nu_+(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = 0$. Using $A \subseteq \pi_I^{-1} \circ \pi_I(A)$ and $\pi_I(A \cap B) \subseteq \pi_I(A) \cap \pi_I(B)$, which hold for all $A, B \subseteq X$, we get

$$\begin{aligned}
\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_- &\subseteq \pi_I^{-1} \circ \pi_I(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) \\
&\subseteq \pi_I^{-1}\big(\pi_I(\operatorname{supp}\nu_+) \cap \pi_I(\operatorname{supp}\nu_-)\big) \\
&\subseteq \pi_I^{-1}(\operatorname{supp}\nu_{+,I} \cap \operatorname{supp}\nu_{-,I}) \ ,
\end{aligned}$$

where we used $\operatorname{supp}\nu_{\pm,I} = \overline{\pi_I(\operatorname{supp}\nu_\pm)}$ from Lemma B.2 in the last step. Since the set on the right hand side is measurable, we can plug this into ν_+ to find

$$\nu_+(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) \leq \nu_{+,I}(\operatorname{supp}\nu_{+,I} \cap \operatorname{supp}\nu_{-,I}) \ .$$

Together with Point (i) and (ii) of Lemma 5.1.1 we get

$$\nu_+(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) \le C_0^{|I|}\tilde{\nu}_{+,I}(\operatorname{supp}\tilde{\nu}_{+,I} \cap \operatorname{supp}\tilde{\nu}_{-,I}) \ .$$

Finally, using the explicit form of $\tilde{\nu}_+$ together with Point (i) of Lemma 4.1.2, for $r = 0$, gives us

$$\nu_+(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) \le \prod_{i\in I} C_0(2\kappa_i)^{q_i} \le \prod_{i\in I} C_0(2\kappa)^q \ .$$

Since this holds for all $I \in \mathcal{F}(\mathbb{N})$ and we assume $C_0(2\kappa)^q < 1$, taking the limit $|I| \to \infty$ yields the assertion. □

The final lemma transfers the learning rates of Theorem 4.3.1 from $\tilde{P}$ to P.

5.1.4 Lemma (Polynomial Learning Rates) *Let $X = \mathbb{R}^{\mathbb{N}}$ be a sequence space as defined in (3.1) with $X_i = \mathbb{R}$ for $i \ge 1$ and P, $\tilde{P}$ be probability distributions on $X \times \{\pm 1\}$ satisfying (5.1). Furthermore, let $\tilde{P}$ satisfy the assumptions of Theorem 4.3.1, where additionally $\kappa < C_0^{-1/q}/2$ and $0 < r < (C_0^{-1/q}/2 - \kappa)\sigma$ holds true. Moreover, define $\alpha := -\log(2\kappa + 2r/\sigma)$ and $\beta := \log\bigl(1 + \lceil \sigma/(2r)\rceil\bigr)$ as in Theorem 4.3.1. Then there are constants $x_0, C \ge 1$ with the following property:*
For $\tau \ge 1$, $n \ge x_0\tau$, and every feature set $I_n \in \mathcal{I}$ with

$$|I_n| = \left\lceil \frac{\log(n/\tau)}{q\alpha - \log(C_0) + \beta} \right\rceil - 1$$

the histogram using a cubic partition with radius r and the feature set I_n satisfies

$$\mathcal{R}_{\mathrm{Class},P}(h_{D,r,I}) - \mathcal{R}^*_{\mathrm{Class},P} < C \cdot \left(\frac{\tau}{n}\right)^{\frac{q\alpha - \log(C_0)}{q\alpha - \log(C_0) + \beta}}$$

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

Compared to Theorem 4.3.1, here the restriction on κ is stronger. Since we have $C_0 \ge 1$, the learning rate for P is worse than for $\tilde{P}$. Nevertheless, we get polynomial rates for both classification problems. If $\tilde{P}$ even satisfies the assumptions of Lemma 4.3.3 or Lemma 4.3.4, the polynomial learning rates of Lemma 4.3.3 and Lemma 4.3.4, respectively, can be transferred to P as well.

Proof. Let $I \in \mathcal{F}(\mathbb{N})$ be fixed. According to Lemma 4.1.4 and Lemma 5.1.3 we have

$$\nu(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = \tilde{\nu}(\operatorname{supp}\tilde{\nu}_+ \cap \operatorname{supp}\tilde{\nu}_-) = 0 \ \ .$$

As a result, Lemma 5.1.2 is applicable. Together with Corollary 4.1.3 for $\tilde{P}$ we receive

$$MN_I(r) \leq C_0^{|I|} \tilde{MN}_I(r) \leq \prod_{i\in I} C_0(2\kappa + 2r/\sigma)^q = \exp\bigl(-(q\alpha - \log(C_0))|I|\bigr)$$
$$|\mathcal{A}_{I,\nu}| \leq |\mathcal{A}_{I,\tilde{\nu}}| \leq 2 \cdot \prod_{i\in I}\bigl(1 + \lceil \sigma/(2r)\rceil\bigr) = 2e^{\beta|I|} \ \ .$$

Our restriction on r ensures $q\alpha - \log(C_0) > 0$. Consequently, proceeding analogously to the proof of Theorem 4.3.1 we get the assertion. □

5.2 Fourier Coefficients

In this section we consider the prototypical example of Assumption 4.1.1 for a decreasing sequence of scaling factors $\sigma_i \searrow 0$ for $i \to \infty$. But we start with an application that motivates the investigation of this scenario.

To this end, we consider a classification problem $\tilde{P}$ given by some probability measures $\tilde{\nu}_+$ and $\tilde{\nu}_-$ on an $\mathcal{L}_2$ space, i.e. we have two stochastic processes with square-integrable sample paths. However, we assume that we cannot observe point evaluations, but we have some information about the Fourier coefficients of the sample paths. This is a typical assumption in tractability studies, see e.g. [67] where even so-called linear information is available to the algorithm. To be more precise, we assume that $\tilde{\nu}_+$ and $\tilde{\nu}_-$ are given by the following construction:
Let $(T, \mathcal{T}, \mu)$ be some measure space and $(e_i)_{i\geq 1}$ be a sequence of functions $e_i : T \to \mathbb{R}$ such that their μ-equivalence classes $[e_i]_\mu$ form an orthonormal system (ONS) in $L_2(\mu)$. Moreover, we assume that the functions $(e_i)_{i\geq 1}$

are pointwise uniformly bounded, i.e. $\sup_{i\geq 1} |e_i(t)| < \infty$ for all $t \in T$. Note that these assumptions are especially satisfied for the following standard Fourier basis.

5.2.1 Example (Standard Fourier Basis) Let $T = [-\pi, \pi]$ be equipped with the Borel σ-algebra $\mathcal{T} = \mathcal{B}(T)$ and $\mu = 1/(2\pi) \cdot \lambda$ be the normalized Lebesgue measure. Then for the sequence $(e_i)_{i\geq 1}$ of functions given by

$$e_i(t) := \begin{cases} \sqrt{2} \cdot \sin\bigl(i/2 \cdot t\bigr), & i \in 2\mathbb{N} \\ \sqrt{2} \cdot \cos\bigl((i+1)/2 \cdot t\bigr), & i \in 2\mathbb{N} - 1, \end{cases}$$

for $t \in T$, the equivalence classes $([e_i]_\mu)_{i\geq 1}$ define an ONS in $L_2(\mu)$. Since the constant function $t \mapsto 1$ is omitted, this is not an orthonormal basis (ONB), see e.g. [72, p. 187] for details (note that in [72] the Lebesgue measure is not normalized).

Now, we define the measures $\tilde{\nu}_\pm$ as the push-forward measures

$$\tilde{\nu}_\pm := \nu_\pm \circ F^{-1}$$

of some measures ν_+ and ν_- on the Borel σ-algebra $\mathcal{B}(\ell_1)$ under the (linear) mapping $F \colon \ell_1(\mathbb{N}) \to \mathcal{L}_2(\mu)$ given by

$$F\bigl((x_i)_{i\geq 1}\bigr) := \sum_{i\geq 1} x_i e_i \quad . \tag{5.3}$$

In other words, the classification problem $\tilde{P}$ on $\mathcal{L}_2(\mu)$ is the transformed scenario of a classification problem P on ℓ_1, given by $\nu_\pm$, under the transformation F. Since we assume that $(e_i)_{i\geq 1}$ is pointwise uniformly bounded, the series in (5.3) converges pointwise. The pointwise convergence allows us to simulate some paths of $\tilde{\nu}_\pm$, see Figure 5.1. Moreover, using the ONS property of $([e_i]_\mu)_{i\geq 1}$ and $\|\cdot\|_{\ell_2} \leq \|\cdot\|_{\ell_1}$ we find $\|F(x)\|_{\mathcal{L}_2(\mu)} = \|x\|_{\ell_2} \leq \|x\|_{\ell_1}$ and hence F is a continuous mapping. As a result, F is measurable with respect to the Borel σ-algebras and the push-forward measures $\tilde{\nu}_\pm$ are well-defined. This motivates the investigation of classification problems on ℓ_1. To this end, the following lemma recalls some basic measurability properties of ℓ_1.

5.2.2 Lemma *The set $\ell_1 \subseteq \mathbb{R}^{\mathbb{N}}$ is measurable and the Borel σ-algebra on ℓ_1 satisfies*

$$\mathcal{B}(\ell_1) = \mathcal{B}\big(\mathbb{R}^{\mathbb{N}}\big)\big|_{\ell_1(\mathbb{N})} = \Bigg(\bigotimes_{i\geq 1} \mathcal{B}(\mathbb{R})\Bigg)\Bigg|_{\ell_1(\mathbb{N})} = \sigma\big(\pi_i \colon \ell_1 \to \mathbb{R} : \ i \geq 1\big) \ ,$$

where $\pi_i \colon \ell_1 \to \mathbb{R}$ are the projections onto the i-th coordinate.

Proof. First we show that every closed ℓ_1-ball $x + rB_{\ell_1} \subseteq \mathbb{R}^{\mathbb{N}}$, with center $x \in \ell_1$ and radius $r > 0$, is a measurable subset of $\mathbb{R}^{\mathbb{N}}$. Since the ball $\pi_{[d]}(x) + rB_{\ell_1^d}$ in $\mathbb{R}^d$ and the projection $\pi_{[d]} \colon \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^d$ are measurable, this is a direct consequence of the representation

$$x + rB_{\ell_1} = \bigcap_{d\geq 1} \pi_{[d]}^{-1}(\pi_{[d]}(x) + rB_{\ell_1^d}) \ .$$

Using the measurability of the closed ℓ_1-balls in $\mathbb{R}^{\mathbb{N}}$ together with $\ell_1 = \bigcup_{m\geq 1} mB_{\ell_1}$ we get the measurability of $\ell_1 \subseteq \mathbb{R}^{\mathbb{N}}$. Now, we consider the claimed identities of the σ-algebra.

The second equality is a consequence of $\mathcal{B}\big(\mathbb{R}^{\mathbb{N}}\big) = \bigotimes_{i\geq 1} \mathcal{B}(\mathbb{R})$, which holds even for general sequence spaces, see e.g. the discussion after (3.1).

For the third equality, note that the product σ-algebra is defined as the initial σ-algebra $\bigotimes_{i\geq 1} \mathcal{B}(\mathbb{R}) = \sigma(\pi_i \ : \ \ i \geq 1)$ with the projections $\pi_i \colon \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ and that the trace σ-algebra is the initial σ-algebra $\sigma(\iota)$ with the embedding $\iota \colon \ell_1 \to \mathbb{R}^{\mathbb{N}}$. Using the transitivity of building the initial σ-algebra, we get that $\big(\bigotimes_{i\geq 1} \mathcal{B}(\mathbb{R})\big)\big|_{\ell_1(\mathbb{N})}$ is the initial σ-algebra of the projections $\pi_i \circ \iota \colon \ell_1 \to \mathbb{R}$, see e.g. [31, Korollar III.5.4] for details. This is the third equality.

For the first equality, we use the already proven ones and show both inclusions separately.

"$\subseteq$" Since ℓ_1 is a separable metric space, every open set in ℓ_1 is a countable union of closed ℓ_1-balls, which are $\mathcal{B}\big(\mathbb{R}^{\mathbb{N}}\big)\big|_{\ell_1}$-measurable by our preliminary remark. This proves the inclusion "$\subseteq$".

"$\supseteq$" Since the projections $\pi\colon \ell_1 \to \mathbb{R}$, which are defined on ℓ_1, are continuous with respect to the ℓ_1-norm, we find $\sigma(\pi_i : \ i \geq 1) \subseteq \mathcal{B}(\ell_1)$ and hence this inclusion is proven. □

Next, we show that the Bayes risks of P on ℓ_1 and $\tilde{P}$ on $\mathcal{L}_2(\mu)$ coincide. Using the ONS property of $(e_i)_{i\geq 1}$, we see that F is injective with inverse $G : F(\ell_1) \to \ell_1$ on the image $F(\ell_1) \subseteq \mathcal{L}_2(\mu)$ given by

$$G(f) \coloneqq \big(\langle f, e_i\rangle_{\mathcal{L}_2(\mu)}\big)_{i\geq 1} \ .$$

Now, we show that G is measurable. According to Lemma 5.2.2 we have $\mathcal{B}(\ell_1) = \sigma(\pi_i : \ i \geq 1)$ and hence G is measurable if and only if $\pi_i \circ G : F(\ell_1) \to \mathbb{R}$ is measurable for all $i \geq 1$. Moreover, the σ-algebra on $F(\ell_1)$, which is the trace σ-algebra, equals the Borel σ-algebra of the pseudo-metric space $F(\ell_1) \subseteq \mathcal{L}_2(\mu)$, that is $\mathcal{B}(\mathcal{L}_2(\mu))|_{F(\ell_1)} = \mathcal{B}(F(\ell_1))$, see e.g. [31, Korollar I.4.6] for details. Since $\pi_i \circ G : F(\ell_1) \to \mathbb{R}$ given by $\pi_i \circ G(f) = \langle f, e_i\rangle_{\mathcal{L}_2(\mu)}$ is continuous, we get the measurability of $\pi_i \circ G$ for all $i \geq 1$. This proves the measurability of $G : F(\ell_1) \to \ell_1$. As a result, the transformation F satisfies Point (ii) and (iii) of Lemma 2.1.3 and hence we find $\sigma(F) = \mathcal{B}(\ell_1)$. Thus, Lemma 2.1.2 yields that the problems $\tilde{P}$ on $\mathcal{L}_2(\mu)$ and P on ℓ_1 are equal in terms of Bayes risks, i.e. $\mathcal{R}^*_{\text{Class},P} = \mathcal{R}^*_{\text{Class},\tilde{P}}$.

As another consequence of Lemma 5.2.2, every probability measure ν on ℓ_1 can be extended to $\mathbb{R}^{\mathbb{N}}$ with $\nu(\ell_1) = 1$ and vice verse. Consequently, we can consider the measures $\nu_\pm$ as measures on $\mathbb{R}^{\mathbb{N}}$ with $\nu_\pm(\ell_1) = 1$. Now, if we want to use Assumption 4.1.1 for $\nu_\pm$, the condition $\nu_\pm(\ell_1) = 1$ is satisfied if and only if the scale sequence $(\sigma_i)_{i\geq 1} \in \ell_1$ is summable. This motivates the investigation of Assumption 4.1.1 with $\sigma_i \searrow 0$ for $i \to \infty$.

Note that Assumption 4.1.1 implies that $\nu_\pm$ are product measures and this corresponds to the independence of the Fourier coefficients of the stochastic processes $\tilde{\nu}_\pm$, which is a reasonable assumption in this context.

Recall that the larger the scaling σ_i of a feature, the higher its influence on the learning algorithm. Since we want to use the most influential features and $(\sigma_i)_{i\geq 1}$ is non-increasing, we use the first features for learning, i.e. we consider feature sets of the form $I = [m] = \{1, \dots, m\}$ with some

$m \geq 1$. The following lemma presents polynomial learning rates under some regularity condition on the scale sequence $(\sigma_i)_{i\geq 1}$.

5.2.3 Lemma (Polynomial Learning Rates) *Let Assumption 4.1.1 be satisfied for $\kappa_i \leq \kappa$ and $q_i \geq q$ for all $i \geq 1$ with some $q > 0$ and $0 \leq \kappa < 1/2$. Furthermore, let $c \geq 1$ and $(\sigma_i)_{i\geq 1}$ satisfy, for all $i \geq 1$,*

$$\bigl(\sigma_1 \cdot \sigma_2 \cdot \ldots \cdot \sigma_i\bigr)^{1/i} \leq c \cdot \sigma_i \quad . \tag{5.4}$$

In addition, let $0 < s < 1/2 - \kappa$ be fixed and $\alpha := -\log(2\kappa + 2s)$, $\beta' := \log\bigl(2 + 1/(2s)\bigr)$. Then there are constants $x_0, C \geq 1$ with the following property:
For $\tau \geq 1$, $n \geq x_0\tau$, and every feature set $I_n = [m_n]$ with

$$m_n = \left\lceil \frac{\log(n/\tau)}{q\alpha + \log(c) + \beta'} \right\rceil - 1$$

the histogram using a cubic partition with radius $r_n = s \cdot \sigma_{m_n}$ and the feature set I_n satisfies

$$\mathcal{R}_{\mathrm{Class},P}(h_{D,r_n,I_n}) - \mathcal{R}^*_{\mathrm{Class},P} < C \cdot \left(\frac{\tau}{n}\right)^{\frac{q\alpha}{q\alpha+\log(c)+\beta'}}$$

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

The condition in (5.4) does not ensure that ν is concentrated on $\ell_1(\mathbb{N})$, e.g. (5.4) is satisfied for the constant sequence $\sigma_i = \sigma$. This means that the application of Lemma 5.2.3 is not limited to our introductory Fourier example. However, the condition in (5.4) remains valid for decreasing sequences if they do not decrease too fast. To be more precise, (5.4) is satisfied for an arbitrary $c \geq 1$ if and only if the *doubling condition* $\sigma_{2i} \asymp \sigma_i$ is satisfied, see Lemma 10.2.2 of Part III. Based on this characterization it is easy to see that polynomially decreasing sequences $\sigma_i \asymp i^{-a}$, with some $a > 0$, satisfy (5.4). In Example 5.2.4 below, we provide an explicit value of the constant c in (5.4) for such a polynomial decreasing sequence. Moreover, using Lemma 10.2.2 again, we find that the condition in (5.4) implies $\sigma_i \succcurlyeq i^{-a}$ for some $a > 0$.

Compared to Theorem 4.3.1, the definition of α coincide if we replace s by r/σ, hence we denote both quantities by α. However, β' is slightly larger than β, hence we added a prime. Moreover, note that the constant $c \geq 1$ of (5.4) appears in the learning rate. All these changes, lead to a slightly worse polynomial order, compared to the situation $\sigma_i = \sigma$ considered in Theorem 4.3.1, but we get polynomial rates in both cases. Finally, in contrast to Theorem 4.3.1, here we choose a varying radius $r_n = s \cdot \sigma_{m_n}$ depending on the scale sequence $(\sigma_i)_{i\geq 1}$.

Proof. The proof is an application of Corollary 3.2.7 and Lemma 4.2.1. To this end, let us fix some feature set $I = [m]$ with $m \geq 1$ and some cubic partition $\mathcal{A}$ with radius $r \coloneqq s\sigma_m$. According to Point (ii) of Lemma 4.1.4 the assumption of Corollary 4.1.3 is satisfied and hence we find

$$\begin{aligned} MN_I(r) &\leq \prod_{i\in I} (2\kappa + 2r/\sigma_i)^q \qquad \text{and} \\ |\mathcal{A}_{I,\nu}| &\leq 2 \cdot \prod_{i\in I} \bigl(1 + \lceil \sigma_i/(2r) \rceil\bigr) \ . \end{aligned}$$

Using the monotonicity of $(\sigma_i)_{i\geq 1}$ and $r = s\sigma_m$ we get

$$MN_I(r) \leq (2\kappa + 2r/\sigma_m)^{qm} \leq (2\kappa + 2s)^{qm} = e^{-q\alpha m} =: e^{q\alpha} \cdot a_m$$

with $a_m \coloneqq \exp\bigl(-q\alpha(m+1)\bigr)$. Now, we turn to the number of relevant cells. Using $\lceil x \rceil \leq x + 1$, $r = s\sigma_m$, and the monotonicity of $(\sigma_i)_{i\geq 1}$ we find

$$\bigl|\mathcal{A}_{I,\nu}\bigr| \leq 2 \cdot \prod_{i=1}^{m} \bigl(2 + \sigma_i/(2s\sigma_m)\bigr) \leq 2 \cdot \bigl(2 + 1/(2s)\bigr)^m \cdot \prod_{i=1}^{m} (\sigma_i/\sigma_m) \ .$$

The condition in (5.4) gives us $\prod_{i=1}^{m}(\sigma_i/\sigma_m) \leq c^m$ and hence we get

$$|\mathcal{A}_{I,\nu}| \leq 2 \cdot c^m \bigl(2 + 1/(2s)\bigr)^m = 2\exp\bigl((\log(c) + \beta')m\bigr) =: 2b_m \ .$$

Finally, proceeding analogously to the proof of Theorem 4.3.1 we get the assertion. □

In the case of a polynomial decreasing scale sequence, the following example presents an explicit constant $c \geq 1$ for the condition in (5.4).

5.2.4 Example (Polynomial Scaling) For $a > 0$, the sequence $(\sigma_i)_{i\geq 1}$ given by $\sigma_i := \lceil i/2 \rceil^{-a}$ satisfies (5.4) with $c = e^a$. In order to prove this we write

$$\left(\frac{(\sigma_1 \cdot \ldots \cdot \sigma_\ell)^{1/\ell}}{\sigma_\ell} \right)^{1/a} = \frac{\lceil \ell/2 \rceil}{\bigl(1 \cdot 1 \cdot 2 \cdot 2 \cdot \ldots \cdot \lceil \ell/2 \rceil\bigr)^{1/\ell}} \tag{5.5}$$

and consider the case $\ell = 2m+1$ with some $m \geq 0$ first. For $m = 0$, (5.5) equals 1 and hence it is bounded by e. For $m \geq 1$, (5.5) equals

$$\frac{m+1}{\bigl((m!)^2 \cdot (m+1)\bigr)^{1/\ell}} \quad .$$

Using Stirling's formula $m! \geq \sqrt{2\pi m} \cdot (m/e)^m$ and $2\pi m \geq e$ we find

$$\begin{aligned} \frac{m+1}{\bigl((m!)^2 \cdot (m+1)\bigr)^{1/\ell}} &\leq \left(\frac{(m+1)^{\ell-1}}{e \cdot (m/e)^{2m}} \right)^{1/\ell} \\ &= \Bigl((1+1/m)^{2m} \cdot e^{2m-1} \Bigr)^{1/\ell} \quad . \end{aligned}$$

Finally, since $1 + 1/m \leq \exp(1/m)$ we find

$$\frac{(\sigma_1 \cdot \ldots \cdot \sigma_\ell)^{1/\ell}}{\sigma_\ell} \leq e^a \quad .$$

For $\ell = 2m$ with $m \geq 1$, an analogous argument gives the same bound. Consequently, Lemma 5.2.3 is applicable for $\sigma_i := \lceil i/2 \rceil^{-a}$ with $a > 0$ and gives us a polynomial learning rate of order

$$\frac{q\alpha}{q\alpha + a + \beta'} \quad .$$

We used $\sigma_i := \lceil i/2 \rceil^{-a}$ in Example 5.2.4, instead of the simpler sequence $\sigma_i := i^{-a}$, to ensure that in combination with the standard Fourier basis of Example 5.2.1 the sine and cosine basis functions with the same frequency have the same scaling.

For learning problems on finite-dimensional spaces the scaling of the data typically only influences the constants but not the learning rate itself. In contrast Example 5.2.4 shows that there are infinite-dimensional learning problems where the scaling of the input values influences the learning rate. Moreover, the learning rate provided by Lemma 5.2.3 gets worse for increasing a. At first sight, this seems to be counter intuitive since in combination with the standard Fourier basis faster decreasing sequences $(\sigma_i)_{i\geq 1}$ correspond to smoother sample paths. But there is the following heuristic argument for this effect: The faster $(\sigma_i)_{i\geq 1}$ decreases, the more the learning algorithm is dominated by the first features. Consequently, the information provided by the last features is nearly invisible for the learning algorithm and hence with decreasing σ_i the algorithm gets less information.

For a visualization of some sample paths of $\tilde{\nu}_\pm = \nu_\pm \circ F^{-1}$, where F is given by (5.3) with the standard Fourier basis described in Example 5.2.1 and P satisfies Assumption 4.1.1 with $\sigma_i = \lceil i/2 \rceil^{-a}$ from Example 5.2.4, see Figure 5.1.

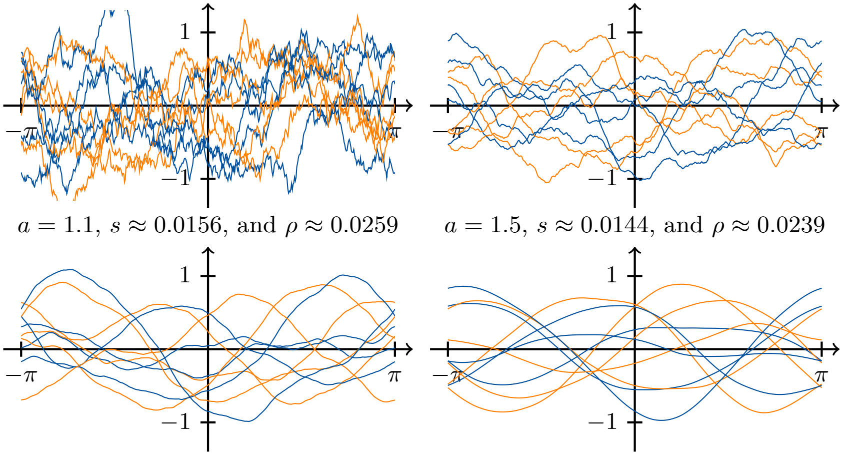
The final lemma of this section shows that the condition in (5.4) is—in some sense—almost sharp.

5.2.5 Lemma (No Polynomial Learning Rates) *Let Assumption 4.1.1 be satisfied for $0 < \kappa' \leq \kappa_i \leq \kappa$ and $q \leq q_i \leq q'$ for all $i \geq 1$ with some $0 < \kappa' \leq \kappa < 1/2$ and $0 < q' \leq q < \infty$. Furthermore, assume that*

$$\lim_{i\to\infty} \frac{(\sigma_1 \cdot \sigma_2 \cdot \ldots \cdot \sigma_i)^{1/i}}{\sigma_i} = \infty \tag{5.6}$$

is satisfied. Then Corollary 3.2.7 does not give polynomial learning rates for any choice of cubic partition sequence and feature set sequence $(I_n)_{i\geq 1}$ of the form $I_n = [m_n]$ with $m_n \geq 1$.

Since the condition in (5.4) for $c \geq 1$ is equivalent to $\sigma_i^{-1}(\sigma_1 \cdot \sigma_2 \cdot \ldots \cdot \sigma_i)^{1/i} \leq c$, there are no sequences $(\sigma_i)_{i\geq 1}$ satisfying both, (5.4) and (5.6). The condition (5.6) is typically satisfied for fast decreasing sequences, e.g. $\sigma_i \asymp \exp(-ai^\lambda)$ with $a, \lambda > 0$ satisfies (5.6).



$a = 1.1$, $s \approx 0.0156$, and $\rho \approx 0.0259$ $a = 1.5$, $s \approx 0.0144$, and $\rho \approx 0.0239$

$a = 2$, $s \approx 0.0132$, and $\rho \approx 0.0219$ $a = 3$, $s \approx 0.0113$, and $\rho \approx 0.0187$

Figure 5.1: Four Plots with 5 paths $\sim \tilde{\nu}_+$ in orange and 5 paths $\sim \tilde{\nu}_-$ in blue each. The distributions are $\tilde{\nu}_\pm = \nu_\pm \circ F^{-1}$ where F is given by (5.3) with the first $\ell = 1000$ elements of the standard Fourier basis described in Example 5.2.1 and $\nu_\pm$ satisfies Assumption 4.1.1 with $\kappa_i = 2/5$, $q_i = 2/3$, and $\sigma_i = \lceil i/2 \rceil^{-a}$ from Example 5.2.4 for all $i \geq 1$. Below each plot there is the used $a > 1$ and a (numerically calculated) optimal value for $s > 0$ maximizing the exponent in the learning rate $n^{-\rho}$ provided by Lemma 5.2.3.

Proof. The proof is an application of Lemma 4.2.2. Let $(r_n)_{n\geq 1}$ and $(I_n)_{n\geq 1}$ be sequences with $r_n > 0$ and $I_n = [m_n] \in \mathcal{F}(\mathbb{N})$ for all $n \geq 1$ and $MN_{I_n}(r_n) \to 0$ for $n \to \infty$. Moreover, let $\mathcal{A}$ be a cubic partition with radius r_n. According to Point (ii) of Lemma 4.1.4 the assumption of Corollary 4.1.3 is satisfied and hence we find

$$MN_{I_n}(r_n) \geq \prod_{i=1}^{m_n} \min\bigl\{2\kappa' + 2r_n/\sigma_i, 1\bigr\}^{q'} \quad \text{and} \quad |\mathcal{A}_{I_n,\nu}| \geq \prod_{i=1}^{m_n} \bigl\lceil \sigma_i/(2r_n) \bigr\rceil \ ,$$

where we additionally used $\kappa_i \geq \kappa'$ and $q_i \leq q'$. Let $1 \leq \ell_n \leq m_n$ be the maximal integer with $2\kappa' + 2r_n/\sigma_{\ell_n} < 1$. If there is no such integer, we set

$\ell_n := 0$. With this notation we find

$$MN_{I_n}(r_n) \geq \prod_{i=1}^{\ell_n} (2\kappa' + 2r_n/\sigma_i)^{q'} \qquad \text{and} \qquad |\mathcal{A}_{I_n,\nu}| \geq \prod_{i=1}^{\ell_n} \bigl(\sigma_i/(2r_n)\bigr) \ .$$

The lower bound on the margin-noise function implies

$$MN_{I_n}(r_n) \geq (2\kappa')^{q'\ell_n} =: a_n \ .$$

Since $MN_{I_n}(r_n) \to 0$ for $n \to \infty$, this implies $\ell_n \to \infty$ for $n \to \infty$. Consequently, we can assume $\ell_n \geq 1$ without loss of generality. As a result, we have $2\kappa' + 2r_n/\sigma_{\ell_n} \leq 1$ and hence $2r_n/\sigma_{\ell_n} \leq 1$ for all $n \geq 1$. This implies

$$|\mathcal{A}_{I_n,\nu}| \geq \sigma_{\ell_n}^{-\ell_n} \prod_{i=1}^{\ell_n} \sigma_i =: b_n \ .$$

All together we find

$$\frac{\log(b_n)}{-\log(a_n)} = \frac{1}{q' \log\bigl(1/(2\kappa')\bigr)} \cdot \log\left(\sigma_{\ell_n}^{-1} \left(\prod_{i=1}^{\ell_n} \sigma_i \right)^{1/\ell_n} \right) \to \infty$$

for $n \to \infty$ according to (5.6). Using Lemma 4.2.2 we get the assertion. □

Finally, note that the statement of Lemma 5.2.5 does not hold true in general if the sequences $(\kappa_i)_{i\geq 1}$ or $(q_i)_{i\geq 1}$ have a more favorable behavior, i.e. $\kappa_i \searrow 0$ or $q_i \nearrow \infty$ for $i \to \infty$.

5.3 Autoregressive Models

In this section we consider a specific type of stochastic processes, namely autoregressive models. Let us formulate the considered type of processes as an assumption.

5.3.1 Assumption (Autoregressive Model) Let $X = \prod_{i\geq 1} X_i$ be a sequence space as defined in (3.1) with $X_i = X_0 := \mathbb{R}^p$ for some $p \geq 1$ and all $i \geq 1$, where X_0 is equipped with the norm $\|\cdot\| := \|\cdot\|_{\ell_\infty^p}$. Furthermore, let

$m \geq 1$ and $f \colon X_0^m \to X_0$ be a Lipschitz continuous function with Lipschitz constant $L > 0$, where $X_0^m = X_0 \times \ldots \times X_0$ is equipped with the norm

$$\|(x_1, \ldots, x_m)\|_{\ell_\infty^m(X_0)} \coloneqq \max_{i=1,\ldots,m} \|x_i\| \enspace .$$

Then let $F \colon X \to X$ with $F\big((\varepsilon_i)_{i \geq 1}\big) \coloneqq (x_i)_{i \geq 1}$ be defined recursively by

$$x_i \coloneqq \begin{cases} \varepsilon_i, & i \leq m \\ \varepsilon_i + f(x_{i-1}, \ldots, x_{i-m}), & i > m. \end{cases}$$

Furthermore, for any probability distribution $\tilde{P}$ on $X \times \{\pm 1\}$ the distribution

$$P \coloneqq \tilde{P} \circ (F, \mathrm{id}_{\{\pm 1\}})^{-1}$$

is called the *autoregressive distribution* of $\tilde{P}$ and f.

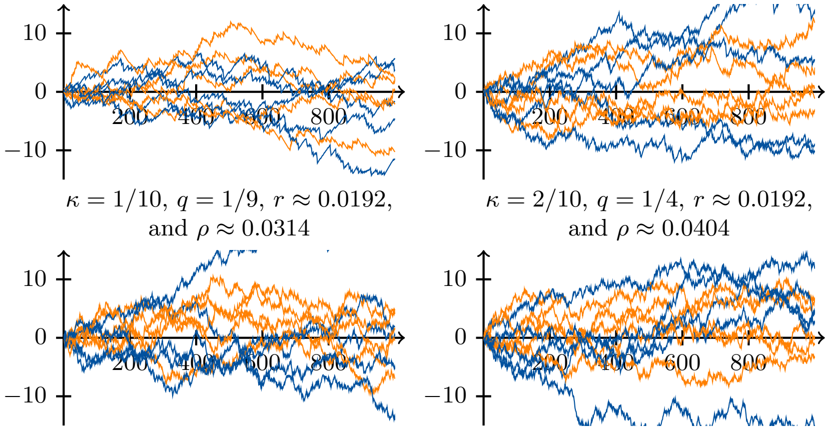
In the following we denote all quantities related to the classification problem $\tilde{P}$ with a tilde and all quantities related to the corresponding autoregressive classification problem P without a tilde. Note that P is the transformed learning scenario of $\tilde{P}$ under the function F. Moreover, to prove learning rates for P we have to assume that $\tilde{P}$ is the prototypical example from Assumption 4.1.1, but we can already relate important quantities of P and $\tilde{P}$ for general $\tilde{P}$. For this reason, Assumption 5.3.1 does not imply any restriction on $\tilde{P}$.

If the norm $\| \cdot \|_{\ell_\infty^p}$ on X_0 is replaced by another norm, the statements remain essentially the same, i.e. only the constants and the polynomial order possibly change.

Let us start with an example illustrating Assumption 5.3.1. To this end, let $m = 1$, $X_0 = \mathbb{R}$, and $f(x) \coloneqq x$. If $\varepsilon = (\varepsilon_i)_{i \geq 1}$ is a random variable with $\varepsilon \sim \tilde{\nu}_+$ then the random variable $x = (x_i)_{i \geq 1} \coloneqq F(\varepsilon)$ satisfies $x \sim \nu_+$. As a result, we find

$$x_i = \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_i$$

for all $i \geq 1$ and hence $x \sim \nu_+$ is a random walk on $\mathbb{R}$. If, in addition, $\tilde{\nu}_+$ is a product measure, the one-dimensional distributions $\nu_{+,i} = \nu_+ \circ \pi_i^{-1}$ of ν_+



$\kappa = 1/10$, $q = 1/9$, $r \approx 0.0192$, and $\rho \approx 0.0314$

$\kappa = 2/10$, $q = 1/4$, $r \approx 0.0192$, and $\rho \approx 0.0404$

$\kappa = 3/10$, $q = 3/7$, $r \approx 0.0143$, and $\rho \approx 0.0370$

$\kappa = 4/10$, $q = 2/3$, $r \approx 0.0071$, and $\rho \approx 0.0228$

Figure 5.2: Four Plots with 5 paths $\sim \nu_+$ in orange and 5 paths $\sim \nu_-$ in blue of length $\ell = 1000$ each. The distributions $\nu_\pm$ are given by Assumption 5.3.1 with $p = 1$, $m = 1$, $f(x) = x$ and $\tilde{P}$ satisfying Assumption 4.1.1, i.e. $\nu_\pm$ are random walks. The parameter sequences are $\kappa_i = \kappa$ and $q_i = q := \frac{\kappa}{1-\kappa}$, i.e $\mathbb{E}_{\varepsilon_i \sim \tilde{\nu}_{\pm,i}} \varepsilon_i = 0$, for all $i \geq 1$. Below each plot there is the used κ, q, and a (numerically calculated) optimal value for $r > 0$ maximizing the exponent in the learning rate $n^{-\rho}$ provided by Lemma 5.3.6.

are given by the convolution

$$\nu_{\pm,i} = \tilde{\nu}_{\pm,1} * \tilde{\nu}_{\pm,2} * \ldots * \tilde{\nu}_{\pm,i} \quad .$$

As a result, the supports and hence the scaling of the feature $x_i \sim \nu_i$ increases for $i \to \infty$. This is in contrast to the uniform scaling considered in Section 4.1. For a visualization of some sample paths of P, in the case that $\tilde{P}$ additionally satisfies Assumption 4.1.1, see Figure 5.2.

The first lemma provides basic properties of the transformation F.

5.3.2 Lemma (Basic Properties) *Let Assumption 5.3.1 be satisfied. Then the following statements are true:*

(i) The function F is continuous.

(ii) The function $G\colon X \to X$ with $G((x_i)_{i\geq 1}) := (\varepsilon_i)_{i\geq 1}$ defined by

$$\varepsilon_i := \begin{cases} x_i, & i \leq m \\ x_i - f(x_{i-1}, \ldots, x_{i-m}), & i > m \end{cases}$$

is the inverse of $F\colon X \to X$. Moreover, G is continuous.

(iii) For every $\ell \geq 1$ there is a finite-dimensional version $F_{[\ell]}\colon X_0^\ell \to X_0^\ell$ of F with

$$F_{[\ell]} \circ \pi_{[\ell]} = \pi_{[\ell]} \circ F \quad .$$

(iv) For every $\ell \geq 1$ there is a finite-dimensional version $G_{[\ell]}\colon X_0^\ell \to X_0^\ell$ of G with

$$G_{[\ell]} \circ \pi_{[\ell]} = \pi_{[\ell]} \circ G \quad .$$

Moreover, $G_{[\ell]}$ is Lipschitz continuous with constant $1 + L$ and $G_{[\ell]}$ is the inverse of $F_{[\ell]}$ for all $\ell \geq 1$.

Note that in the definition of F there are components of x and ε on the right hand side and hence F is defined recursively. But in the definition of G there are only components of x on the right hand side and hence G is not defined recursively.

Since the function F is continuous, it is measurable. Consequently, the autoregressive distribution $P = \tilde{P} \circ (F, \mathrm{id}_{\{\pm 1\}})^{-1}$ is well-defined. Moreover, since F is bijective with measurable (even continuous) inverse G, Point (ii) and (iii) of Lemma 2.1.3 are satisfied and hence we have $\sigma(F) = \mathcal{B}(X)$. Together with Lemma 2.1.2 we get $\mathcal{R}^*_{\mathrm{Class},P} = \mathcal{R}^*_{\mathrm{Class},\tilde{P}}$ and the learning problems $\tilde{P}$ and P are equal in terms of Bayes risks.

For the later use it is important that the Lipschitz constant of $G_{[\ell]}$ does not depend on ℓ.

Proof. (i) Since $X = X_0^{\mathbb{N}}$ is equipped with the product topology, the function $F\colon X \to X$ is continuous if and only if $\pi_i \circ F\colon X \to X_0$ is continuous for all $i \geq 1$. We prove the latter by induction. For $i = 1, \ldots, m$ we have

$\pi_i \circ F(\varepsilon) = \pi_i(\varepsilon)$ and hence $\pi_i \circ F$ is continuous. Now, assume that $\pi_i \circ F$ is continuous for all $i \leq k-1$ with some $k-1 \geq m$. Then we can write

$$\pi_k \circ F(\varepsilon) = \pi_k(\varepsilon) + f\bigl(\pi_{k-1} \circ F(\varepsilon), \dots, \pi_{k-m} \circ F(\varepsilon)\bigr) \ .$$

as a combination of continuous functions. Consequently, $\pi_k \circ F$ is continuous and the assertion is proven by induction.

(ii) To see the continuity of G we write

$$\pi_i \circ G = \begin{cases} \pi_i, & i \leq m \\ \pi_i - f(\pi_{i-1}, \dots, \pi_{i-m}), & i > m \end{cases}$$

as a combination of continuous functions. Next, we prove the identity $G \circ F = \mathrm{Id}_X$. To this end, we fix some $\varepsilon = (\varepsilon_i)_{i\geq 1}$ and set $x = (x_i)_{i\geq 1} \coloneqq F(\varepsilon)$ as well as $\varepsilon' = (\varepsilon'_i)_{i\geq 1} \coloneqq G(x) = G \circ F(\varepsilon)$. Then we have to show $\varepsilon_i = \varepsilon'_i$ for all $i \geq 1$. For $i = 1, \dots, m$ we have $\varepsilon'_i = x_i = \varepsilon_i$. For $i > m$ the definition of ε'_i and x_i gives us

$$\begin{aligned} \varepsilon'_i &= x_i - f(x_{i-1}, \dots, x_{i-m}) \\ &= \varepsilon_i + f(x_{i-1}, \dots, x_{i-m}) - f(x_{i-1}, \dots, x_{i-m}) = \varepsilon_i \ . \end{aligned}$$

(iii) This is a direct consequence of the fact that the k-th component x_k of $x = F(\varepsilon)$ depends only on the components ε_i for $i = 1, \dots, k$.

(iv) The existence of a finite-dimensional version $G_{[\ell]}$ follows by an analogous argument as in Point (iii). Next, we prove the Lipschitz continuity of $G_{[\ell]}$. To this end, let $x = (x_i)_{i=1}^{\ell} \in X_0^{\ell}$ and $x' = (x'_i)_{i=1}^{\ell} \in X_0^{\ell}$. Then we write

$$\bigl\|G_{[\ell]}(x) - G_{[\ell]}(x')\bigr\|_{\ell_\infty^\ell(X_0)} = \max_{i=1,\dots,\ell} \bigl\|\pi_i \circ G_{[\ell]}(x) - \pi_i \circ G_{[\ell]}(x')\bigr\| \ .$$

For $i = 1, \dots, m$ we have

$$\bigl\|\pi_i \circ G_{[\ell]}(x) - \pi_i \circ G_{[\ell]}(x')\bigr\| = \|x_i - x'_i\| \leq \|x - x'\|_{\ell_\infty^\ell(X_0)} \ .$$

For $i = m+1, \dots, \ell$, using the Lipschitz continuity of f we receive

$$\begin{aligned}
&\big\|\pi_i \circ G_{[\ell]}(x) - \pi_i \circ G_{[\ell]}(x')\big\| \\
&= \big\|\big(x_i - f((x_{i-k})_{k=1}^m)\big) - \big(x'_i - f((x'_{i-k})_{k=1}^m)\big)\big\| \\
&\leq \|x_i - x'_i\| + \big\|f\big((x_{i-k})_{k=1}^m\big) - f\big((x'_{i-k})_{k=1}^m\big)\big\| \\
&\leq \|x_i - x'_i\| + L \max_{k=1,\dots,m} \|x_{i-k} - x'_{i-k}\| \\
&\leq (1+L) \cdot \|x - x'\|_{\ell_\infty^\ell(X_0)} \quad .
\end{aligned}$$

All together this proves the Lipschitz continuity of $G_{[\ell]}$ with Lipschitz constant $1+L$.

Finally, we prove $G_{[\ell]} \circ F_{[\ell]} = \mathrm{Id}_{X_0^\ell}$. This can be done analogously to the proof of $G \circ F = \mathrm{Id}_X$ in Point (ii). Alternatively, since the projection $\pi_{[\ell]} \colon X \to X_0^\ell$ is surjective, it is enough to show $G_{[\ell]} \circ F_{[\ell]} \circ \pi_{[\ell]} = \pi_{[\ell]}$ which is a direct consequence of the previously shown properties of F, G, $F_{[\ell]}$, and $G_{[\ell]}$, namely

$$G_{[\ell]} \circ F_{[\ell]} \circ \pi_{[\ell]} = G_{[\ell]} \circ \pi_{[\ell]} \circ F = \pi_{[\ell]} \circ G \circ F = \pi_{[\ell]} \quad .$$

This concludes the proof. □

The next lemma relates the margin-noise functions of $\tilde{P}$ and the corresponding autoregressive distribution P.

5.3.3 Lemma (Margin-Noise Function) *Let Assumption 5.3.1 be satisfied. Then the following statements are true:*

(i) $\eta \circ F = \tilde{\eta}$ *$\tilde{\nu}$-almost surely.*

(ii) There is no noise in the autoregressive classification problem P if and only if there is no noise in the classification problem $\tilde{P}$.

If we define $\eta \coloneqq \tilde{\eta} \circ G$ for a fixed version of $\tilde{\eta}$ then the following statements are true:

(iii) $\tilde{X}_\pm = F^{-1}(X_\pm)$.

(iv) $MN_{[\ell]}(r) \leq \tilde{MN}_{[\ell]}\big(r(1+L)\big)$ *for all $\ell \geq 1$ and $r \geq 0$.*

(v) $M_{[\ell]}(r) \leq \tilde{M}_{[\ell]}\big(r(1+L)\big)$ *for all $\ell \geq 1$ and $r \geq 0$.*

Note that we use the same norm for $\tilde{\Delta}_{[\ell]}$ and $\Delta_{[\ell]}$.

Proof. (i) Since F is measurable and bijective with measurable inverse, we have $\sigma(F) = \mathcal{B}$ according to Lemma 2.1.3. Together with Lemma 2.1.6 we find $\eta \circ F = \mathbb{E}_{\tilde{\nu}}(\tilde{\eta}|\sigma(F)) = \tilde{\eta}$ $\tilde{\nu}$-almost surely.

(ii) As shown in Point (i) we have $\sigma(F) = \mathcal{B}$ and hence Lemma 2.1.2 gives $\mathcal{R}^*_{\text{Class},P} = \mathcal{R}^*_{\text{Class},\tilde{P}}$. Consequently, Lemma 1.2.2 gives the assertion.

(iii) This is a direct consequence of the definition $\eta := \tilde{\eta} \circ G$, namely

$$\tilde{X}_+ = \{\tilde{\eta} > 1/2\} = \{\eta \circ F > 1/2\} = F^{-1}\big(\{\eta > 1/2\}\big) = F^{-1}(X_+) \ .$$

The identity $\tilde{X}_- = F^{-1}(X_-)$ can be proven analogously.

(iv)+(v) First, we investigate $\Delta_{[\ell]} \circ F$. To this end, we fix some $\varepsilon \in X$ and assume $\varepsilon \in \tilde{X}_+$. In this case we have $F(\varepsilon) \in X_+$ according to Point (iii). Using the representation in (3.13), $X_- = F(\tilde{X}_-)$ from Point (iii), and $\pi_{[\ell]} \circ F = F_{[\ell]} \circ \pi_{[\ell]}$ from Lemma 5.3.2 we receive

$$\begin{aligned} \Delta_{[\ell]} \circ F(\varepsilon) &= \operatorname{dist}\big(\pi_{[\ell]} \circ F(\varepsilon), \pi_{[\ell]}(X_-)\big) \\ &= \operatorname{dist}\big(\pi_{[\ell]} \circ F(\varepsilon), \pi_{[\ell]} \circ F(\tilde{X}_-)\big) \\ &= \operatorname{dist}\big(F_{[\ell]} \circ \pi_{[\ell]}(\varepsilon), F_{[\ell]} \circ \pi_{[\ell]}(\tilde{X}_-)\big) \ . \end{aligned}$$

If we apply the Lipschitz continuity of $G_{[\ell]}$ and $G_{[\ell]} \circ F_{[\ell]} = \mathrm{Id}_{X_0^\ell}$ then we get

$$\begin{aligned} & \operatorname{dist}\big(F_{[\ell]} \circ \pi_{[\ell]}(\varepsilon), F_{[\ell]} \circ \pi_{[\ell]}(\tilde{X}_-)\big) \\ = \ & \inf_{\varepsilon' \in \tilde{X}_-} \big\|F_{[\ell]} \circ \pi_{[\ell]}(\varepsilon) - F_{[\ell]} \circ \pi_{[\ell]}(\varepsilon')\big\|_{\ell^\ell_\infty(X_0)} \\ \geq \ & \frac{1}{1+L} \inf_{\varepsilon' \in \tilde{X}_-} \big\|G_{[\ell]} \circ F_{[\ell]} \circ \pi_{[\ell]}(\varepsilon) - G_{[\ell]} \circ F_{[\ell]} \circ \pi_{[\ell]}(\varepsilon')\big\|_{\ell^\ell_\infty(X_0)} \\ = \ & \frac{1}{1+L} \inf_{\varepsilon' \in \tilde{X}_-} \big\|\pi_{[\ell]}(\varepsilon) - \pi_{[\ell]}(\varepsilon')\big\|_{\ell^\ell_\infty(X_0)} \\ = \ & \frac{1}{1+L} \operatorname{dist}\big(\pi_{[\ell]}(\varepsilon), \pi_{[\ell]}(\tilde{X}_-)\big) \ . \end{aligned}$$

Since we assume $\varepsilon \in \tilde{X}_+$, the right hand side equals $\tilde{\Delta}_{[\ell]}(\varepsilon)$ and all together we get

$$\Delta_{[\ell]} \circ F(\varepsilon) \geq \frac{1}{1+L} \cdot \tilde{\Delta}_{[\ell]}(\varepsilon) \tag{5.7}$$

for all $\varepsilon \in \tilde{X}_+$. This inequality can be shown analogously for $\varepsilon \in \tilde{X}_-$. For $\varepsilon \in (\tilde{X}_+ \cup \tilde{X}_-)^c$ we have $F(\varepsilon) \in (X_+ \cup X_-)^c$ and hence $\Delta_{[\ell]} \circ F(\varepsilon) = 0 = \frac{1}{1+L} \cdot \tilde{\Delta}_{[\ell]}(\varepsilon)$. As a result, (5.7) is satisfied for all $\varepsilon \in X$.

Now, we are ready to prove Point (iv) which is a direct consequence of $\nu = \tilde{\nu} \circ F^{-1}$, $\eta \circ F = \tilde{\eta}$, and (5.7), namely

$$\begin{aligned} MN_{[\ell]}(r) &= \int_{\{\Delta_{[\ell]} \leq 2r\}} |2\eta - 1| \,\mathrm{d}\nu \\ &= \int_{\{\Delta_{[\ell]} \circ F \leq 2r\}} |2\eta \circ F - 1| \,\mathrm{d}\tilde{\nu} \\ &\leq \int_{\{\tilde{\Delta}_{[\ell]} \leq 2r(1+L)\}} |2\tilde{\eta} - 1| \,\mathrm{d}\tilde{\nu} \\ &= M\tilde{N}_{[\ell]}(r(1+L)) \ . \end{aligned}$$

Finally, Point (v) can be proven analogously. □

The following lemma provides a bound on the number of relevant cells of a product partition with respect to $\nu_\pm$.

5.3.4 Lemma (Relevant Cells) *Let Assumption 5.3.1 be satisfied and $\mathcal{A}$ be a cubic partition of $(X_0)_{i\geq 1}$ with radius $r > 0$. Then, for $\ell \geq m$, the number of relevant cells is bounded by*

$$\left|\mathcal{A}_{[\ell],\nu_\pm}\right| \leq 3^{p\ell} \cdot \prod_{i=1}^{m} \mathcal{N}\big(\operatorname{supp} \tilde{\nu}_{\pm,i}, r\big) \prod_{i=m+1}^{\ell} \mathcal{N}\big(rLB_{X_0} + \operatorname{supp} \tilde{\nu}_{\pm,i}, r\big) \ .$$

Proof. For symmetry reasons it is enough to consider ν_+. Since the product partition $\mathcal{A}_{[\ell]}$ of $X_0^\ell \cong \mathbb{R}^{p\ell}$ is a cubic partition, the assumption in (1.21) is satisfied with $r_0 = r$ for the norm $\|\cdot\|_{\ell_\infty^\ell(X_0)}$. Consequently, Lemma 1.3.7

with $\varepsilon = r$ and $M = \operatorname{supp} \nu_{+,[\ell]}$ ensures

$$\left|\mathcal{A}_{[\ell],\nu_+}\right| \leq 3^{p\ell} \mathcal{N}(\operatorname{supp} \nu_{+,[\ell]}, r) \quad .$$

Since $F_{[\ell]}$ is continuous, Lemma B.2 gives us $\operatorname{supp} \nu_{+,[\ell]} = \overline{F_{[\ell]}(\operatorname{supp} \tilde{\nu}_{+,[\ell]})}$ and together with Lemma C.4 and Lemma B.3 we find

$$\begin{aligned} \mathcal{N}(\operatorname{supp} \nu_{+,[\ell]}, r) &= \mathcal{N}\big(\overline{F_{[\ell]}(\operatorname{supp} \tilde{\nu}_{+,[\ell]})}, r\big) \\ &= \mathcal{N}\big(F_{[\ell]}(\operatorname{supp} \tilde{\nu}_{+,[\ell]}), r\big) \\ &\leq \mathcal{N}\Big(F_{[\ell]}\big(\textstyle\prod_{i=1}^{\ell} \operatorname{supp} \tilde{\nu}_{+,i}\big), r\Big) \quad . \end{aligned}$$

Now, it remains to give an upper bound for the right hand side. To this end, we define $n_i := \mathcal{N}(\operatorname{supp} \tilde{\nu}_{+,i}, r)$ for $i = 1, \ldots, m$ and $n_i := \mathcal{N}(rLB_{X_0} + \operatorname{supp} \tilde{\nu}_{+,i}, r)$ for $i = m+1, \ldots, \ell$. Moreover, we choose corresponding r-nets $\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_i} \in X_0$, i.e. for $i = 1, \ldots, m$ we have

$$\operatorname{supp} \tilde{\nu}_{+,i} \subseteq \bigcup_{j=1}^{n_i} \varepsilon_{i,j} + rB_{X_0} \tag{5.8}$$

and for $i = m+1, \ldots, \ell$ we have

$$rLB_{X_0} + \operatorname{supp} \tilde{\nu}_{+,i} \subseteq \bigcup_{j=1}^{n_i} \varepsilon_{i,j} + rB_{X_0} \quad . \tag{5.9}$$

Then, for multi-indexes $j = (j_1, \ldots, j_\ell)$ with $1 \leq j_i \leq n_i$ for all $i = 1, \ldots, \ell$, we define $\varepsilon_j := (\varepsilon_{1,j_1}, \ldots, \varepsilon_{\ell,j_\ell})$ and $x_j := F_{[\ell]}(\varepsilon_j)$. This defines at most $n_1 \cdot \ldots \cdot n_\ell$ elements x_j and hence it remains to prove that they form an r-net of $F_{[\ell]}\big(\prod_{i=1}^{\ell} \operatorname{supp} \tilde{\nu}_{+,i}\big)$.

To this end, let $x = (x_i)_{i=1}^{\ell} \in F_{[\ell]}\big(\prod_{i=1}^{\ell} \operatorname{supp} \tilde{\nu}_{+,i}\big)$ and $\varepsilon = (\varepsilon_i)_{i=1}^{\ell} \in \prod_{i=1}^{\ell} \operatorname{supp} \tilde{\nu}_{+,i}$ with $F_{[\ell]}(\varepsilon) = x$. Then, we choose a multi-index $j = (j_1, \ldots, j_\ell)$ by selecting $1 \leq j_i \leq n_i$ with $\|x_i - x_{i,j_i}\| \leq r$ recursively for all $i = 1, \ldots, \ell$ as follows:

For $i = 1, \ldots, m$, according to (5.8) there is some $1 \leq j_i \leq n_i$ such that $\|\varepsilon_i - \varepsilon_{i,j_i}\| \leq r$ is satisfied. Since $x_i = \varepsilon_i$ and $x_{i,j_i} = \varepsilon_{i,j_i}$ are satisfied, we have

$$\|x_i - x_{i,j_i}\| \le r \quad .$$

Now, assume that $j_1, \dots, j_{k-1}$ for some $k-1 \ge m$ have already been defined such that $\|x_i - x_{i,j_i}\| \le r$ for $i = 1, \dots, k-1$. Then we have

$$\left\| f\big((x_{k-i})_{i=1}^m\big) - f\big((x_{k-i,j_{k-i}})_{i=1}^m\big) \right\| \le L \max_{i=1,\dots,m} \|x_{k-i} - x_{k-i,j_{k-i}}\| \le Lr$$

and the difference on the left hand side is contained in rLB_{X_0}. According to (5.9) and $\varepsilon_k \in \operatorname{supp} \tilde{\nu}_{+,k}$ there is some $1 \le j_k \le n_k$ with

$$\left\| f\big((x_{k-i})_{i=1}^m\big) - f\big((x_{k-i,j_{k-i}})_{i=1}^m\big) + \varepsilon_k - \varepsilon_{k,j_k} \right\| \le r \quad .$$

Since we have

$$\begin{aligned} x_k &= \varepsilon_k + f(x_{k-1}, \dots, x_{k-m}) \qquad \text{and} \\ x_{k,j_k} &= \varepsilon_{k,j_k} + f(x_{k-1,j_{k-1}}, \dots, x_{k-m,j_{k-m}}) \quad , \end{aligned}$$

this choice ensures $\|x_k - x_{k,j_k}\| \le r$. By induction, we get a multi-index $j = (j_1, \dots, j_\ell)$ with $1 \le j_i \le n_i$ for $i = 1, \dots, \ell$ and $\|x - x_j\| \le r$. This finishes the proof. □

The next lemma combines Lemma 5.3.3 and Lemma 5.3.4 with the assumption that $\tilde{P}$ satisfies Assumption 4.1.1.

5.3.5 Lemma (Upper Bounds) *Let Assumption 5.3.1 be satisfied for $p = 1$ and let $\tilde{P}$ satisfy Assumption 4.1.1. Then the following statements are true, for $\ell \ge 1$:*

(i) If $\operatorname{supp} \tilde{\nu}_+ \cap \operatorname{supp} \tilde{\nu}_-$ is a $\tilde{\nu}$-zero set then there is no noise in the autoregressive scenario P and there is a version of η with

$$MN_{[\ell]}(r) \le \prod_{i=1}^{\ell} \left(2\kappa_i + \frac{2(1+L)r}{\sigma_i} \right)$$

for all $r \ge 0$.

(ii) For $r > 0$ and a cubic partition $\mathcal{A}$ of $(X_i)_{i\geq 1}$ with radius r the number of relevant cells satisfies

$$|\mathcal{A}_{[\ell],\nu}| \leq 2 \cdot (3\lceil L \rceil)^{\ell} \cdot \prod_{i=1}^{\ell} \Bigl(1 + \Bigl\lceil \frac{\sigma_i}{2r} \Bigr\rceil\Bigr) \ .$$

Proof. (i) This is a direct consequence of Lemma 5.3.3 and Lemma 4.1.3.

(ii) First, we consider the covering numbers appearing in Lemma 5.3.4. Since $\operatorname{supp}\tilde{\nu}_{\pm,i} = \pm[-\kappa_i\sigma_i, (1-\kappa_i)\sigma_i]$ is an interval of length σ_i and rLB_{X_0} is an interval of length $2rL$, the sum $rLB_{X_0} + \operatorname{supp}\tilde{\nu}_{\pm,i}$ is an interval of length $\sigma_i + 2rL$. Consequently, the covering numbers are bounded by

$$\mathcal{N}\bigl(rLB_{X_0} + \operatorname{supp}\tilde{\nu}_{\pm,i}, r\bigr) = \Bigl\lceil L + \frac{\sigma_i}{2r} \Bigr\rceil \leq \lceil L \rceil + \Bigl\lceil \frac{\sigma_i}{2r} \Bigr\rceil \leq \lceil L \rceil \Bigl(1 + \Bigl\lceil \frac{\sigma_i}{2r} \Bigr\rceil\Bigr) \ ,$$

where we used $\lceil a + b \rceil \leq \lceil a \rceil + \lceil b \rceil$ which holds true for all $a, b > 0$. Analogously, we get

$$\mathcal{N}(\operatorname{supp}\tilde{\nu}_{\pm,i}, r) = \Bigl\lceil \frac{\sigma_i}{2r} \Bigr\rceil \leq \lceil L \rceil \Bigl(1 + \Bigl\lceil \frac{\sigma_i}{2r} \Bigr\rceil\Bigr) \ .$$

Together with $|\mathcal{A}_{[\ell],\nu}| \leq |\mathcal{A}_{[\ell],\nu_+}| + |\mathcal{A}_{[\ell],\nu_-}|$ from Lemma 1.3.3 and an application of Lemma 5.3.4 we get the assertion. □

The next lemma shows that if $\tilde{P}$ satisfies Assumption 4.1.1, we get polynomial learning rates for the corresponding autoregressive distribution P.

5.3.6 Lemma (Polynomial Learning Rates) *Let Assumption 5.3.1 be satisfied for $p = 1$ and let $\tilde{P}$ satisfy Assumption 4.1.1 with $\sigma_i = \sigma$, $q_i \geq q$, and $\kappa_i \leq \kappa$ for all $i \geq 1$ for some $\sigma, q > 0$, and $\kappa < 1/2$. Furthermore, let $0 < r < (1/2 - \kappa)\sigma/(1 + L)$ be fixed and $\alpha' := -\log\bigl(2\kappa + 2(1 + L)r/\sigma\bigr)$, $\beta := \log\bigl(1 + \lceil \sigma/(2r) \rceil\bigr)$. Then there are constants $x_0, C \geq 1$ with the following property:*
For $\tau \geq 1$, $n \geq x_0\tau$, and every feature set $I_n = [\ell_n]$ with

$$\ell_n = \Bigl\lceil \frac{\log(n/\tau)}{q\alpha' + \log(3\lceil L \rceil) + \beta} \Bigr\rceil - 1$$

the histogram using a cubic partition with radius r and the feature set I_n satisfies

$$\mathcal{R}_{\text{Class},P}(h_{D,r,I_n}) - \mathcal{R}^*_{\text{Class},P} < C \cdot \left(\frac{\tau}{n}\right)^{\frac{q\alpha'}{q\alpha' + \log(3\lceil L\rceil) + \beta}}$$

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

Compared to Theorem 4.3.1, the definitions of β coincide, but the definitions of α and α' coincide only if we replace r by $(1+L)r$, hence we added a prime. Moreover, note that the Lipschitz constant L of f appears in the learning rate. All these changes result in a slightly worse polynomial order for P, compared to the learning rate for $\tilde{P}$ provided by Theorem 4.3.1, but we get polynomial rates for both classification problems.

If $\tilde{P}$ even satisfies the assumptions of Lemma 4.3.3 or Lemma 4.3.4, the polynomial learning rates of these lemmas can be transferred to P as well. For a visualization of some paths of a random walk, i.e. $f(x) = x$ in Assumption 5.3.1, see Figure 5.2.

Proof. Let $\ell \geq 1$, $I = [\ell]$, and $r \geq 0$ be fixed. According to our assumptions on $\tilde{P}$ Lemma 4.1.4 gives us $\tilde{\nu}(\operatorname{supp}\tilde{\nu}_+ \cap \operatorname{supp}\tilde{\nu}_-) = 0$. As a result, Lemma 5.3.5 is applicable and gives us a version of η with

$$\begin{aligned} MN_I(r) &\leq \big(2\kappa + 2(1+L)r/\sigma\big)^{q\ell} = e^{-q\alpha'\ell} \\ |\mathcal{A}_{I,\nu}| &\leq 2 \cdot (3\lceil L\rceil)^{\ell} \cdot \prod_{i=1}^{\ell}\left(1 + \left\lceil \frac{\sigma}{2r} \right\rceil\right) = 2\exp\big((\log(3\lceil L\rceil) + \beta)\ell\big) \ . \end{aligned}$$

Finally, proceeding analogously to the proof of Theorem 4.3.1 we get the assertion. □

5.4 α-Mixing

In this section we use the concept of α-mixing conditions to relax the independence assumption used in Assumption 4.1.1.

Let us start with a brief introduction to the α-mixing coefficient. To this end, let $X = \prod_{i\geq 1} X_i$ be a sequence space as defined in (3.1) and P be a distribution on $X \times \{\pm 1\}$ with marginals $\nu_\pm$ of the positive and negative labeled data points, respectively. For two sub-σ-algebras $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{B}$ we define

$$\alpha_\pm(\mathcal{F}_1, \mathcal{F}_2) \coloneqq \sup_{\substack{A_1\in\mathcal{F}_1\\ A_2\in\mathcal{F}_2}} \bigl|\nu_\pm(A_1 \cap A_2) - \nu_\pm(A_1)\nu_\pm(A_2)\bigr| \ .$$

Note that the sub-σ-algebras $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{B}$ are independent with respect to $\nu_\pm$ if and only if the corresponding quantity $\alpha_\pm(\mathcal{F}_1, \mathcal{F}_2) = 0$ is zero. In this sense $\alpha_\pm(\mathcal{F}_1, \mathcal{F}_2)$ measures the dependence of $\mathcal{F}_1$ and $\mathcal{F}_2$ with respect to $\nu_\pm$. Moreover, for $1 \leq i \leq j \leq \infty$ we define the sub-σ-algebras

$$\mathcal{F}_i^j \coloneqq \sigma\bigl(\pi_k : \ i \leq k \leq j\bigr)$$

generated by the projections $\pi_k \colon X \to X_k$ with $i \leq k \leq j$. Then we define the *α-mixing coefficient* with respect to $\nu_\pm$ by

$$\alpha_\pm(k) \coloneqq \sup_{i\geq 1} \alpha_\pm\bigl(\mathcal{F}_1^i, \mathcal{F}_{i+k}^\infty\bigr) \ .$$

for $k \geq 1$. Note that we do not assume that $(\pi_k)_{k\geq 1}$ is a stationary sequence of random variables. Consequently, we need the supremum in the definition of $\alpha_\pm$. Moreover, we set $\alpha(k) \coloneqq \max\bigl\{\alpha_+(k), \alpha_-(k)\bigr\}$. A direct consequence of these definitions is that $(\alpha_\pm(k))_{k\geq 1}$ and $(\alpha(k))_{k\geq 1}$ are non-increasing sequences. If $\alpha_\pm(k) \searrow 0$ even vanishes for $k \to \infty$, we say that $\nu_\pm$ is *α-mixing.* For examples of α-mixing stochastic processes see e.g. [45, Section 3.1] and the references therein and for a general overview of mixing coefficients see e.g. the survey [14].

Furthermore, we denote the one-dimensional marginal distributions of $\nu_\pm$, for $i \geq 1$, by $\nu_{\pm,i} \coloneqq \nu_\pm \circ \pi_i^{-1}$ and define the corresponding product measures

$$\tilde{\nu}_\pm \coloneqq \bigotimes_{i\geq 1} \nu_{\pm,i}. \tag{5.10}$$

Consequently, $\tilde{\nu}_\pm$ defines a classification problem $\tilde{P}$ on $X \times \{\pm 1\}$ which we call the *independent classification problem* of P. Moreover, all quantities related to $\tilde{P}$ are denoted with a tilde and all quantities related to P are denoted without a tilde.

For $I \in \mathcal{F}(\mathbb{N})$, we denote the *inner distance* of I by $\delta(I) \coloneqq \inf_{i,j \in I: i \neq j} |i - j| \geq 1$. Using the convention $\inf \emptyset = \infty$ we have $\delta(I) = \infty$ for singletons $I = \{i\} \subseteq \mathbb{N}$.

The first lemma provides a basic property of the α-mixing coefficient.

5.4.1 Lemma (Basic Properties) *Let $X = \prod_{i \geq 1} X_i$ be a sequence space as defined in* (3.1)*, ν_+ be a probability measure on X, and $I = \{i_1, \ldots, i_m\} \subseteq \mathbb{N}$ be a feature set with $i_1 < i_2 < \ldots < i_m$, $m \geq 1$, and inner distance $\delta(I) \geq k \geq 1$. Then, for $\sigma(\pi_i)$-measurable sets $A_i \in \sigma(\pi_i)$, the following bound is satisfied*

$$\nu_+\Big(\bigcap_{i \in I} A_i\Big) \leq \alpha_+(k) \sum_{s=1}^{m} \prod_{j=1}^{s-1} \nu_+(A_{i_j}) + \prod_{i \in I} \nu_+(A_i) \quad .$$

Proof. The proof is based on the following telescope sum

$$\begin{aligned}
&\nu_+\Big(\bigcap_{i \in I} A_i\Big) - \prod_{j=1}^{m} \nu_+(A_{i_j}) \\
&= \sum_{s=1}^{m} \Big[\nu_+\Big(\bigcap_{j=s}^{m} A_{i_j}\Big) \cdot \prod_{j=1}^{s-1} \nu_+(A_{i_j}) - \nu_+\Big(\bigcap_{j=s+1}^{m} A_{i_j}\Big) \cdot \prod_{j=1}^{s} \nu_+(A_{i_j}) \Big] \quad .
\end{aligned}$$

For $1 \leq s \leq m$, the corresponding summand can be bounded by

$$\begin{aligned}
&\nu_+\Big(\bigcap_{j=s}^{m} A_{i_j}\Big) \cdot \prod_{j=1}^{s-1} \nu_+(A_{i_j}) - \nu_+\Big(\bigcap_{j=s+1}^{m} A_{i_j}\Big) \cdot \prod_{j=1}^{s} \nu_+(A_{i_j}) \\
&= \Big[\nu_+\Big(A_{i_s} \cap \bigcap_{j=s+1}^{m} A_{i_j}\Big) - \nu_+\Big(\bigcap_{j=s+1}^{m} A_{i_j}\Big) \cdot \nu_+(A_{i_s}) \Big] \cdot \prod_{j=1}^{s-1} \nu_+(A_{i_j}) \\
&\leq \alpha_+(i_{s+1} - i_s) \prod_{j=1}^{s-1} \nu_+(A_{i_j}) \quad ,
\end{aligned}$$

where we used $\bigcap_{j=s+1}^{m} A_{i_j} \in \mathcal{F}_{i_{s+1}}^{m}$ and $A_{i_s} \in \mathcal{F}_1^{i_s}$. Since we assume $\delta(I) \geq k$, we have $i_{s+1} - i_s \geq k$ and hence $\alpha_+(i_{s+1} - i_s) \leq \alpha_+(k)$ is satisfied. This gives the assertion. □

For the remaining part of this section we consider the case where the corresponding independent classification problem $\tilde{P}$ satisfies Assumption 4.1.1.

5.4.2 Lemma (No Noise) *Let $X = \prod_{i\geq 1} X_i$ be a sequence space as defined in* (3.1) *and P be a probability distribution on $X \times \{\pm 1\}$ with corresponding independent distribution $\tilde{P}$ given by* (5.10)*. Furthermore, let $\tilde{P}$ satisfy Assumption 4.1.1 with $\sigma_i = \sigma$, $q_i \geq q$, and $\kappa_i \leq \kappa$ for all $i \geq 1$ with some $\sigma, q > 0$, and $\kappa < 1/2$. If the measures $\nu_\pm$ are α-mixing then the following equality is satisfied*

$$\nu(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = \nu(\operatorname{supp}\tilde{\nu}_+ \cap \operatorname{supp}\tilde{\nu}_-) = 0 \ .$$

Proof. Since $\tilde{P}$ satisfies Assumption 4.1.1, $\tilde{\nu}(\operatorname{supp}\tilde{\nu}_+ \cap \operatorname{supp}\tilde{\nu}_-) = 0$ is already stated in Point (ii) of Lemma 4.1.4. Now, let $I := k \cdot [\ell] = \{k, 2k, 3k, \ldots, \ell k\}$ with some fixed $\ell, k \geq 1$ and

$$A_i := \big(\operatorname{supp}\nu_{+,i} \cap \operatorname{sup}\nu_{-,i}\big) \times \prod_{j\neq i} X_j \in \sigma(\pi_i)$$

for $i \geq 1$. For symmetry reasons it is enough to prove $\nu_+(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = 0$. According to Lemma B.3 we have $\operatorname{supp}\nu_\pm \subseteq \prod_{i\geq 1} \operatorname{supp}\nu_{\pm,i}$ and hence

$$\begin{aligned}\nu_+\big(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-\big) &\leq \nu_+\left(\prod_{i\geq 1} \operatorname{supp}\nu_{+,i} \cap \operatorname{supp}\nu_{-,i}\right) \\ &= \nu_+\left(\bigcap_{i\geq 1} A_i\right) \leq \nu_+\left(\bigcap_{i\in I} A_i\right) \ .\end{aligned}$$

Since $\delta(I) \geq k$ is satisfied, Lemma 5.4.1 yields

$$\nu_+(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) \leq \alpha_+(k) \sum_{s=1}^{\ell} \prod_{j=1}^{s-1} \nu_+(A_{kj}) + \prod_{j=1}^{\ell} \nu_+(A_{kj}) \ .$$

From Point (i) of Lemma 4.1.2 for $r = 0$ and our assumptions we get

$$\nu_+(A_i) \le (2\kappa_i)^{q_i} \le (2\kappa)^q =: a < 1 \ .$$

Together we find

$$\nu_+(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) \le \alpha_+(k)\sum_{s=1}^{\ell} a^{s-1} + a^\ell \le \frac{\alpha_+(k)}{1-a} + a^\ell \ .$$

Since this bound is satisfied for all $k, \ell \ge 1$ and ν_+ is α-mixing, we get the assertion. □

In order to establish bounds on the margin-noise function we need some restrictions on the (non-empty) feature set $I \in \mathcal{F}(\mathbb{N})$ used for learning. To this end, we introduce the following notation: For $k \ge 1$, we denote some arbitrary subset with maximal cardinality and inner distance of at least k by

$$I_k \in \operatorname{argmax}\bigl\{|I'| : \ \emptyset \ne I' \subseteq I \text{ with } \delta(I') \ge k\bigr\} \ .$$

Since $\delta(\{i\}) = \infty$ is infinite for singletons $\{i\} \subseteq \mathbb{N}$, there is always such a subset. Note that $(|I_k|)_{k\ge1}$ is a non-increasing sequence with $|I_k| \ge 1$ for all $k \ge 1$, $I_k = I$ for $k \le \delta(I)$, and $|I_k| = 1$ for $k > \max I - \min I$. Using this notation we define, for $0 < v \le 1$ and $k \ge 1$,

$$\mathcal{I}_{k,v} := \bigl\{I \in \mathcal{F}(\mathbb{N}) : \ |I_k| \ge v|I|\bigr\} \ .$$

Roughly speaking, $\mathcal{I}_{k,v}$ contains the feature sets I that are in some sense well separated. Larger values of v corresponds to a larger separation.

Let us briefly discuss the consequences for a *grid* $I = k \cdot [\ell]$ of distance $k \ge 1$ and length $\ell \ge 1$. Since $\delta(I) = k$ is satisfied, we have $I_{k'} = I$ for $k' \le k$. For $k' > k$, a subset $I' \subseteq I$ with $\delta(I') \ge k'$ and maximal cardinality can be chosen as follows: To this end, we denote the elements of I' by $I' = \{i_0, i_1, \ldots, i_{\ell'-1}\}$ with $i_0 < i_1 < \ldots i_{\ell'-1}$ for some $1 \le \ell' \le \ell$. Since $I' \subseteq I = k \cdot [\ell]$ is a subset, every i_j is of the form

$$i_j = k \cdot m_j$$

with some $1 \leq m_j \leq \ell$. Moreover, the distance of two consecutive elements $i_{j+1} - i_j = k(m_{j+1} - m_j) \geq k'$ must be at least k'. Dividing by k and using $m_{j+1} - m_j \in \mathbb{N}$ yields

$$m_{j+1} - m_j \geq \lceil k'/k \rceil \ .$$

Consequently, the subset $I' \subseteq I$ satisfies $\delta(I') \geq k'$ and is of maximal cardinality if we take the first element of I, i.e. $i_0 = k$ (with $m_0 := 1$), and choose $m_{j+1} := m_j + \lceil k'/k \rceil$ for $j \geq 0$ recursively. As a result, we get

$$m_j = 1 + j \cdot \lceil k'/k \rceil$$

for $j \geq 0$. This procedure can be continued as long as $i_j = k \cdot m_j \leq k\ell$ is satisfied. This condition is equivalent to

$$j \leq \left\lfloor \frac{\ell - 1}{\lceil k'/k \rceil} \right\rfloor =: \ell' - 1 \ .$$

Finally, we set $I_{k'} := I'$. To sum it up, for $\ell, k \geq 1$ and $k' > k$, we constructed a subset $I_{k'} \subseteq I = k \cdot [\ell]$ with $\delta(I_{k'}) \geq k'$ and

$$|I_{k'}| = \ell' = \left\lfloor \frac{\ell - 1}{\lceil k'/k \rceil} \right\rfloor + 1 \ . \tag{5.11}$$

Note that this identity remains valid for $1 \leq k' \leq k$. Finally, for $\ell \geq 2$ we can bound this by

$$|I_{k'}| \geq \frac{1}{2\lceil k'/k \rceil} |I|$$

and hence $I = k \cdot [\ell] \in \mathcal{I}_{k',v}$ with $v := 1/(2\lceil k'/k \rceil)$ for $k, k' \geq 1$ and $\ell \geq 2$.

Now, we are ready to prove the following lemma which establishes an oracle inequality for feature sets in $\mathcal{I}_{k,v}$ using the α-mixing coefficient.

5.4.3 Lemma (Oracle Inequality) *Let X be a sequence space as defined in* (3.1) *and P be a probability distribution on $X \times \{\pm 1\}$ with corresponding independent distribution $\tilde{P}$ given by* (5.10). *Furthermore, let $\tilde{P}$ satisfy Assumption 4.1.1 with $\sigma_i = \sigma$, $q_i \geq q$, and $\kappa_i \leq \kappa$ for all $i \geq 1$ with*

some $\sigma, q > 0$, $\kappa < 1/2$, and $\nu_\pm$ be α-mixing. Moreover, let $0 < v \leq 1$, $0 < r < (1/2-\kappa)\sigma$ be fixed and $\alpha := -\log(2\kappa+2r/\sigma)$, $\beta := \log\bigl(1+\lceil\sigma/(2r)\rceil\bigr)$. Then there are constants $x_0, C \geq 1$ with the following property:
For $\tau \geq 1$, $n \geq x_0\tau$, $k \geq 1$, and every feature set $I_n \in \mathcal{I}_{k,v}$ with

$$|I_n| = \left\lceil \frac{\log(n/\tau)}{vq\alpha+\beta} \right\rceil - 1$$

the histogram using a cubic partition with radius r and the feature set I_n satisfies

$$\mathcal{R}_{\mathrm{Class},P}(h_{D,r,I_n}) - \mathcal{R}^*_{\mathrm{Class},P} < \frac{\alpha(k)}{1-e^{-q\alpha}} + C \cdot \left(\frac{\tau}{n}\right)^{\frac{vq\alpha}{vq\alpha+\beta}}$$

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

If $\nu_\pm$ are α-mixing with $\alpha(k) > 0$ for all $k \geq 1$, we have to choose $k = k_n \to \infty$ for $n \to \infty$ to get consistency or learning rates from this bound.

Note that the definitions of α and β coincide with the corresponding definitions in Theorem 4.3.1.

Proof. Let $k, m \geq 1$ and $I \in \mathcal{I}_{k,v}$ with $|I| = m$ be fixed. First, we consider the margin-noise function for some fixed $r > 0$. According to Lemma 5.4.2 we have $\nu(\operatorname{supp}\nu_+ \cap \operatorname{supp}\nu_-) = 0$ and hence Lemma 2.3.4 is applicable. As a result, there is a version of η such that

$$\begin{aligned} MN_I(r) = {} & p_+\nu_{+,I}\bigl(\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{-,I}) \leq 2r\bigr) \\ & + p_-\nu_{-,I}\bigl(\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{+,I}) \leq 2r\bigr) \ . \end{aligned}$$

For symmetry reasons it is enough to investigate the first summand in the following. According to Lemma B.3 we have $\operatorname{supp}\nu_{-,I} \subseteq \prod_{i\in I}\operatorname{supp}\nu_{-,i}$ and hence Lemma 3.3.3 gives us

$$\begin{aligned} \nu_{+,I}\bigl(\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{-,I}) \leq 2r\bigr) & \leq \nu_{+,I}\Bigl(\operatorname{dist}\bigl(\,\cdot\,, \textstyle\prod_{i\in I}\operatorname{supp}\nu_{-,i}\bigr) \leq 2r\Bigr) \\ & = \nu_{+,I}\Bigl(\textstyle\prod_{i\in I}\bigl\{\operatorname{dist}(\,\cdot\,, \operatorname{supp}\nu_{-,i}) \leq 2r\bigr\}\Bigr) \ . \end{aligned}$$

Since we assume $I \in \mathcal{I}_{k,v}$, there is a subset $I_k \subseteq I$ with $|I_k| \geq vm$ and $\delta(I_k) \geq k$. Using the notation $A_i := \{\operatorname{dist}(\,\cdot\,, \operatorname{supp} \nu_{-,i}) \leq 2r\} \times \prod_{j \neq i} X_j \subseteq X$ for $i \in I$ we find

$$\nu_{+,I}\big(\operatorname{dist}(\,\cdot\,, \operatorname{supp} \nu_{-,I}) \leq 2r\big) \leq \nu_+\Bigg(\bigcap_{i \in I} A_i\Bigg) \leq \nu_+\Bigg(\bigcap_{i \in I_k} A_i\Bigg) \ .$$

Since $\tilde{P}$ satisfies Assumption 4.1.1 and $\nu_{+,i} = \tilde{\nu}_{+,i}$ holds true, Point (i) of Lemma 4.1.2 gives us

$$\nu_+(A_i) = \nu_{+,i}\big(\operatorname{dist}(\,\cdot\,, \operatorname{supp} \nu_{-,i}) \leq 2r\big) \leq (2\kappa + 2r/\sigma)^q = e^{-q\alpha} < 1 \ .$$

Combining both bounds with Lemma 5.4.1 yields

$$\begin{aligned}\nu_{+,I}\big(\operatorname{dist}(\,\cdot\,, \operatorname{supp} \nu_{-,I}) \leq 2r\big) &\leq \alpha_+(k) \sum_{s=1}^{|I_k|} e^{-(s-1)q\alpha} + e^{-q\alpha|I_k|} \\ &\leq \frac{\alpha_+(k)}{1 - e^{-q\alpha}} + e^{-vq\alpha m} \ .\end{aligned}$$

Altogether, we get for the margin-noise function

$$MN_I(r) \leq \frac{\alpha(k)}{1 - e^{-q\alpha}} + e^{-vq\alpha|I|} = \frac{\alpha(k)}{1 - e^{-q\alpha}} + a_m$$

with $a_m := e^{-vq\alpha m}$. Now, we consider the number of relevant cells. Since $\mathcal{A}_{I,\nu} = \mathcal{A}_{I,\nu_+} \cup \mathcal{A}_{I,\nu_-}$ holds true according to Lemma 1.3.3, it is enough to consider $|\mathcal{A}_{I,\nu_+}|$. Using Lemma 3.2.1 and Point (ii) of Lemma 4.1.2 we get

$$|\mathcal{A}_{I,\nu_+}| \leq \prod_{i \in I} \big|(\mathcal{A}_i)_{\nu_{+,i}}\big| \leq \big(1 + \lceil \sigma/(2r) \rceil\big)^{|I|} = e^{\beta m} =: b_m \ .$$

Consequently, we have $|\mathcal{A}_{I,\nu}| \leq 2b_m$. Finally, proceeding analogously to the proof of Theorem 4.3.1 we get the assertion. □

In order get learning rates from Lemma 5.4.3 we need explicit bounds on the α-mixing coefficient $\alpha(k)$ and the structure of the feature set I. In the following we consider the case where $I = k \cdot [\ell]$ is a grid and investigate under which conditions on $k = k_n$ Lemma 5.4.3 provides the same learning rate for P as Theorem 4.3.1 does for the independent learning Problem $\tilde{P}$, i.e. $n^{-\frac{q\alpha}{q\alpha+\beta}}$. Recall from (5.11), that for a grid $I_n = k_n \cdot [\ell_n]$ with $\ell_n = \lceil \log(n/\tau)/(q\alpha+\beta) \rceil - 1$ we have $I_n \in \mathcal{I}_{k_n,1}$ and hence Lemma 5.4.3 yields

$$\mathcal{R}_{\mathrm{Class},P}(h_{D,r,I_n}) - \mathcal{R}^*_{\mathrm{Class},P} < \frac{\alpha(k_n)}{1-e^{-q\alpha}} + C \cdot \left(\frac{\tau}{n}\right)^{\frac{q\alpha}{q\alpha+\beta}}$$

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

Now, let us assume that $\nu_\pm$ are *polynomially α-mixing*, i.e. there is a constant $c > 0$ and $\delta > 0$ with

$$\alpha_\pm(k) \le ck^{-\delta}$$

for all $k \ge 1$. As a result, we recover the learning rate $n^{-\frac{q\alpha}{q\alpha+\beta}}$ if there is a constant $c > 0$ with

$$k_n \ge c\Big(\frac{n}{\tau}\Big)^{\frac{q\alpha/\delta}{q\alpha+\beta}}$$

for all $n \ge 1$. Analogously, if we assume that $\nu_\pm$ are *geometrically α-mixing*, i.e. there is a constant $c > 0$ and $\delta, \lambda > 0$ with

$$\alpha_\pm(k) \le c\exp(-\delta k^\lambda)$$

for all $k \ge 1$ then we recover the learning rate $n^{-\frac{q\alpha}{q\alpha+\beta}}$ if there is a constant $c > 0$ with

$$k_n \ge \left(\frac{q\alpha/\delta}{q\alpha+\beta} \cdot \log(n/\tau)\right)^{1/\lambda}$$

for all $n \ge 1$. Note that geometrically α-mixing is stronger than polynomially α-mixing, but also the restriction on k_n is much weaker in the geometrical case than in the polynomial case. Roughly speaking, the faster the decay of the α-mixing coefficient the weaker the condition on k_n.

Finally, note that if the feature set $I = [\ell]$, for $\ell \ge 1$, is a block, i.e. a grid of distance 1, then (5.11) implies

$$\frac{1}{2k} \le \frac{|I_k|}{|I|} \le \frac{\frac{\ell-1}{k}+1}{\ell} \le \frac{1}{k} + \frac{1}{\ell}$$

for $k \geq 1$ and $\ell \geq 2$. Assuming that $\alpha(k_0) = 0$ for some $k_0 \geq 1$, we can apply Lemma 5.4.3 for $v = 1/(2k_0)$ and $k = k_0$ to get the polynomial learning rate

$$\left(\frac{\tau}{n}\right)^{\frac{q\alpha}{q\alpha+2k_0\beta}} .$$

This learning rate for P is worse than the rate for $\tilde{P}$ from Theorem 4.3.1. Nevertheless, we get polynomial rates for both classification problems.

However, if $\nu_\pm$ are α-mixing with $\alpha(k) > 0$ for all $k \geq 1$, we have to choose $k = k_n \to \infty$ for $n \to \infty$. Consequently, for every fixed $0 < v \leq 1$ there is some $n_0 \geq 1$ such that $I_n = [\ell_n] \notin \mathcal{I}_{k_n,v}$ for all $n \geq n_0$. As a result, Lemma 5.4.3 does not provide any consistency result or learning rate for the simple reason that it only applies for finitely many n. This is in contrast to the results in Section 5.2 and Section 5.3, where $I = [\ell]$ is a reasonable choice for a feature set.

Chapter 6

Discussion

In the first part of the thesis we provided a mathematical framework to rigorously treat high-dimensional learning scenarios. Moreover, we used this framework to investigate histograms on high-dimensional classification problems. However, these are only small steps towards understanding learning with high-dimensional data. Hence, we would like to draw attention to the following research questions.

To achieve the learning rates presented in Section 4.3, we chose the hyper parameters, i.e. the feature set I and radius r, based on parameters of the learning scenario at hand. However, these parameters are unknown in practical applications. So the question is, how to choose the hyper parameters adaptively, i.e. without knowing the parameters of the learning scenario. We conjecture that the classical cross-validation method is an appropriate choice for the radius. However, there is an own research field called *feature selection* for the feature set. It would be interesting to pair a concrete feature selection method with our framework and analyze its statistical properties.

The histogram method is probably the easiest learning algorithm to analyze theoretically, but in practice it is rarely used. This raises the question, how other learning methods, e.g. kernel methods, perform in the high-dimensional scenarios of Chapter 4 and Chapter 5. For methods based on Gaussian kernels, our investigations in Part II can possibly be useful.

Our developed framework in Chapter 2 and Chapter 3 indicates that other learning goals, i.e. different loss functions, can be treated in the same way. We already presented some results for LS regression in Lemma 2.1.4, Lemma 2.2.2, and Lemma 3.2.4. For regression problems in general an important question is, what are appropriate regularity assumptions for the Bayes function $f^*_{\mathrm{LS},P}$. Since $f^*_{\mathrm{LS},P}$ is a function on an infinite-dimensional space, there is no obvious choice. For example, there are various non-equivalent generalizations of Sobolev differentiable functions in the literature, see e.g. [12] and references therein. Another example is the classical Hölder continuity, but even for this type of regularity assumption different possibly non-equivalent metrics on X can be considered. Maybe a look into the tractability literature, which deals with high- or infinite-dimensional numerical problems, can give some inspiration for defining suitable regularity assumptions, see e.g. [67] for an introduction to this topic.

In the remaining part of this section we briefly discuss further approaches tackling high-dimensional learning problems:
One strand of research additionally assumes that the support of the data generating distribution is concentrated on a low-dimensional set, i.e. the learning problem is intrinsically low-dimensional. For various algorithms this approach is able to replace, in the learning rate, the dimension d of the ambient space $X = \mathbb{R}^d$ by the (fractal) dimension of the support $\operatorname{supp}\nu$. For kernel estimators in combination with small ball probabilities see e.g. [35, Section 13.3.1] and for Gaussian support vector machines (SVMs) in combination with the (upper) box-counting dimension see e.g. [44]. Moreover, the oracle inequalities in Lemma 1.3.1 and Lemma 1.3.2 in combination with Lemma 1.3.8 show that histograms using cubic partitions are capable to exploit a low intrinsic dimension as well.

Next, we compare our findings with those of the already mentioned book [35] and the article [36], which belong to the field of *functional data analysis (FDA)*. To this end, we start with a brief overview of these contributions. The authors of [35] and [36] consider pseudo-metric (there called semi-metric) spaces as input space X. In [35], for a fixed input value $x \in X$, pointwise convergence of various quantities of the conditional distribution is proven. These results are improved to uniform convergence in [36]. The following common problem occurs in both contributions for infinite-dimensional learning problems: As stated in [35, Section 13.3.2], most of the classical stochastic processes ν on $X = C_b([0,1])$, equipped with the uniform or the L_p-norm, are of *exponential-type*, see e.g. [35, Definition 13.4] for the definition, and hence the learning rates decrease only logarithmically, see e.g. [35, Proposition 13.5]. This fact is unsatisfactory from a statistical point of view. To overcome this shortcoming, [35] and [36] suggest the usage of a specially designed pseudo-metric on X of the form

$$d(x,x') := \left(\sum_{i=1}^{d} \langle x - x', e_i \rangle_X^2 \right)^{1/2} , \tag{6.1}$$

where X is assumed to be a Hilbert space and $(e_i)_{i\geq 1}$ an ONS in X. This pseudo-metric is called *projections type pseudo-metric* and depends only on finitely many Fourier coefficients, see [35, Lemma 13.6 i)]. Under some additional assumptions [35, Lemma 13.6 ii)], in combination with [35, Proposition 13.2], provides polynomial learning rates if X is equipped with the projections type pseudo-metric. However, in these additional assumptions the Hölder continuity of $f^*_{\mathrm{LS},P} \colon X \to \mathbb{R}$ is included and since $d(x, x + x') = 0$ for all $x \in \operatorname{span}\{e_1, \dots, e_d\}$ and $x' \perp \{e_1, \dots, e_d\}$ this implies

$$f^*_{\mathrm{LS},P}(x + x') = f^*_{\mathrm{LS},P}(x) \ . \tag{6.2}$$

As a result, the considered Bayes function depends only on finitely many Fourier coefficients with respect to the ONS $(e_i)_{i\geq 1}$. To summarize, the considerations of [35, 36] lead to polynomial learning rates if the Bayes function decision depends only on finitely many features or if the support $\operatorname{supp}\nu$ has a low-dimensional structure, as mentioned in the previous paragraph. Now, we compare these results with ours. The projections type pseudo-metric in (6.1) is closely related to our pull-back pseudo-metric defined in (2.10). To be more precise, it is the pull-back pseudo-metric with respect to the transformation $s \colon X \to \ell_2^d$ given by $s(x) := (\langle x, e_i \rangle_X)_{i=1}^d$. In contrast to [35, 36], we consider a whole family of transformations and hence a family of pull-back pseudo-metrics. More precisely, we treat the pseudo-metric as a hyper parameter of our learning method. Since we consider a different learning problem, namely classification instead of regression, and a different algorithm, it is an open question if our approach can weaken the assumption in (6.2), which is implicitly used in [35, 36] to get polynomial learning rates.

Another strand of research considers *functional linear regression*, i.e. the LS Bayes function is of the form

$$f^*_{\mathrm{LS},P}(x) = a + \int_0^T \beta(t)x(t)\ \mathrm{d}t = a + \langle \beta, x\rangle_{L_2(\lambda)}\ ,$$

for $x \in X \subseteq L_2(\lambda)$ with some function $\beta \in L_2(\lambda)$, see e.g. [16, 52] and references therein as well as [86] for a generalization. For example [16] proves, for a fixed input value $x \in X$, polynomial pointwise convergence of an algorithm, which has access to the infinite-dimensional data under regularity assumptions on β, x, and the covariance kernel $K(u, v) := \mathbb{E}_{x\sim\nu}x(u)x(v)$. Possibly, our framework can be helpful to overcome the unrealistic assumption of accessing the original infinite-dimensional data.

In [40], see also the related article [41], the authors consider a sequence of learning problems $(P_d)_{d\geq 1}$ on $\mathbb{R}^d \times Y$ and the sequence $(H_d)_{d\geq 1}$ of hypothesis classes given by

$$H_d := \left\{ (x_i)_{i=1}^d \mapsto \rho\biggl(\sum_{i=1}^d x_i\beta_i\biggr) :\ \beta_i \in \mathbb{R} \text{ and } \bigl|\{i :\ \beta_i \neq 0\}\bigr| = k_n \right\}$$

with a sequence $k_n = o\bigl(n/\log(n)\bigr)$ and a (measurable) function $\rho\colon \mathbb{R} \to \mathbb{R}$. Then for an ℓ_1-regularized empirical risk minimizer (ERM) over H_n and the sequence $d = d_n = n^\alpha$, with some $\alpha > 1$, convergences in probability of

$$\mathcal{R}_{L,P_d}(f_D) - \mathcal{R}^*_{L,P_d,H_d} \to 0$$

for $n \to \infty$ is proven under some additional technical assumptions, see [40, Theorem 2]. This type of convergence is called *persistence*, see [40, Definition 1] for a definition, and is a weakened type of consistency, since the Bayes risk $\mathcal{R}^*_{L,P}$ is replaced by the minimal risk $\mathcal{R}^*_{L,P_d,H_d}$ over H_d.

Next, let us have a look at *linear discriminant analysis*, see e.g. [15, Chapter 1] for an introduction to this topic. The most basic setting is a classification problem P_d on $\mathbb{R}^d \times \{\pm 1\}$, where P_d is given by $p_+ = p_- = 1/2$

and two normal distributions $\nu_{d,+} = \mathcal{N}(\mu_{d,+}, \Sigma_d)$ and $\nu_{d,-} = \mathcal{N}(\mu_{d,-}, \Sigma_d)$ with a common covariance matrix $\Sigma_d \in \mathbb{R}^{d\times d}$ and different mean vectors $\mu_{d,\pm} \in \mathbb{R}^d$. In most contributions, see e.g. [7, 33], the assumption that the distance between the mean vectors is bounded away from zero, i.e. $d(\mu_{d,+}, \mu_{d,-}) \geq c > 0$, plays an important role. Since the mean vectors coincide (and are identical zero) in our prototypical example from Chapter 4 for $q_i = \frac{\kappa_i}{1-\kappa_i}$, $i \geq 1$, this approach does not cover our results. Note that the *Mahalanobis distance*

$$d_{\Sigma_d}(\mu_{d,+}, \mu_{d,-}) := \left\langle \mu_{d,+} - \mu_{d,-}, \Sigma_d^{-1}(\mu_{d,+} - \mu_{d,-}) \right\rangle_{\ell_2^d}$$

and the Euclidean distance are popular distance measures for the discrepancy of the mean vectors. The starting point of most contributions is *Fisher's linear discriminant rule*, which is the learning method that assigns the label $+1$ to an unseen data point $x \in \mathbb{R}^d$ if

$$d_{\hat{\Sigma}_d}(x, \hat{\mu}_{d,+}) \leq d_{\hat{\Sigma}_d}(x, \hat{\mu}_{d,-}) \tag{6.3}$$

and otherwise the label -1. Here $\hat{\mu}_{d,\pm}$ and $\hat{\Sigma}_d$ denote the sample versions of the mean vectors and the covariance matrix. For fixed $d \geq 1$ it is well-known that Fisher's linear discriminant rule is consistent, see e.g. [15, Equation (3.4) in Chapter 1]. However, in [7, Theorem 1 (a)] it is shown that in the so-called *high-dimension low-sample-size (HDLSS)* setting, i.e. for $d = d_n$ with $d_n/n \to \infty$ as $n \to \infty$, Fisher's linear discriminant rule performs asymptotically not better than guessing. For this reason, there are various generalizations of Fisher's linear discriminant rule addressing this issue in the HDLSS setting. For example, in [7, Theorem 1 (b)] a modification called *independence rule* or *naive Bayes* is investigated. In this modification the off-diagonal entries of $\hat{\Sigma}_d$ are set to zero in (6.3). In this theorem it is proven that the independence rule performs better than guessing if, in addition, the eigenvalues of Σ_d are bounded away from zero and infinity and $\mu_{d,\pm}$ are contained in a compact subset of ℓ_2. However, this theorem provides consistency only if Σ_d is a multiple of the identity matrix. Moreover, in [33] the authors combine the independence rule with a feature selection method and investigate the statistical properties of the combined method. For further generalization see e.g. [2, 46, 47, 49] and the references therein. Unfortunately, in most contributions there are no learning rates provided.

Finally, we mention the so-called *distribution regression.* In this approach we assume that P is a distribution on $\mathcal{M}_1 \times Y$, i.e. the marginal ν is a distribution on the set of probability distributions $\mathcal{M}_1$ on a measurable space Z. However, the algorithm has only access to finitely many samples $z_1, \dots, z_N \sim x$ of each data point $x \in \mathcal{M}_1$, see e.g. [80, 81, 65] and references therein. This scenario can be applied to model high-dimensional learning problems as follows: Let $Z = [0,1] \times \mathbb{R}$ be the product of the unit interval and $\mathbb{R}$. Then every data point $x \in X$ itself defines a learning problem on Z, with input space $[0,1]$ and output space $\mathbb{R}$, and using the corresponding LS Bayes function $f^*_{\mathrm{LS},x} \colon [0,1] \to \mathbb{R}$ the samples are given by $z_i = (t_i, s_i)$ with $t_i \sim x \circ \pi^{-1}_{[0,1]}$ and

$$s_i = f^*_{\mathrm{LS},x}(t_i) + \varepsilon_i \quad ,$$

where ε_i is some error term for $i = 1, \dots, N$. In this regard, the samples can be interpreted as noisy evaluations of the (random) function $f^*_{\mathrm{LS},x}$ at random points. This approach is more general than our framework since the mapping from the infinite-dimensional object $f^*_{\mathrm{LS},x}$ to the finite-dimensional object $(z_1, \dots, z_N) \in ([0,1] \times \mathbb{R})^N$ includes another source of randomness. But in most contributions to this field the assumptions, which allow polynomial learning rates, are very abstract such that an interpretation or verification in a concrete example is a difficult task.

To sum it up, our prototypical example from Chapter 4 and its generalizations from Chapter 5 contribute—to the best of our knowledge—new polynomial learning rates to the field of high-dimensional classification.

Part II

Gaussian Kernels

In this part we consider the probably most popular kernel used for kernel-based learning methods, namely Gaussian kernels. For the statistical analysis of such learning methods the capacity of the corresponding RKHSs plays an important role. We provide some new bounds on the capacity of the Gaussian RKHS using so-called log-covering numbers of its ℓ_∞-embedding. In Chapter 7 we set up our notation and provide some preparatory material. Chapter 8 is based on the article [77], which establishes log-covering number bounds for Gaussian RKHS over finite-dimensional domains. These bounds provide the dependence on the kernel width and the dimension of the underlying space explicitly. In Chapter 9 we generalize these log-covering number bounds to some particular infinite-dimensional domains.

Chapter 7

Introduction and Preparation

In this chapter we introduce the notation and some preparatory material.

7.1 Definitions and Basic Properties

Since we consider Gaussian kernels on finite-dimensional as well as infinite-dimensional spaces, we use a unifying notation covering both cases. To this end, we consider sequences over an index set $I \subseteq \mathbb{N}$ and recall the definition of the sequence space $\ell_p(I) := \{x = (x_i)_{i \in I} \in \mathbb{R}^I : \ \|x\|_{\ell_2(I)} < \infty\}$ with the norm

$$\|x\|_{\ell_2(I)} := \left(\sum_{i \in I} |x_i|^2\right)^{1/2}$$

and the closed unit ball $B_{\ell_2(I)}$. Using this notation for $I = [d]$ and $I = \mathbb{N}$ we get the finite-dimensional case $\ell_2^d = \ell_2([d]) = \mathbb{R}^d$ and the infinite-dimensional case $\ell_2 = \ell_2(\mathbb{N})$, respectively.

Next, for bounded sequences $\boldsymbol{\sigma} = (\sigma_i)_{i \in I} \in \ell_\infty(I)$ with $\sigma_i > 0$ for all $i \in I$ we define the *diagonal operator* $D_{\boldsymbol{\sigma}} \colon \ell_2(I) \to \ell_2(I)$ by

$$D_{\boldsymbol{\sigma}}(x_i)_{i \in I} = (\sigma_i x_i)_{i \in I} \ .$$

Using this diagonal operator we define the *(anisotropic) Gaussian kernel* on $\ell_2(I)$ with width vector $\boldsymbol{\sigma}$ by

$$k_{\boldsymbol{\sigma}}(x, x') := \exp\bigl(-\|D_{\boldsymbol{\sigma}} x - D_{\boldsymbol{\sigma}} x'\|_{\ell_2(I)}^2\bigr) \ ,$$

for $x, x' \in \ell_2(I)$. For a subset $X \subseteq \ell_2(I)$, we denote the Gaussian kernel restricted to X again by $k_{\boldsymbol{\sigma}}$ and the corresponding Gaussian RKHS by $H_{\boldsymbol{\sigma}}(X)$. For a general introduction to reproducing kernels and RKHSs see e.g. [76, Chapter 4].

Since $k_{\boldsymbol{\sigma}}(x,x) = 1$ for all $x \in X$ the Gaussian kernel is bounded and hence the Gaussian RKHS $H_{\boldsymbol{\sigma}}(X)$ is a subset of the space $\ell_\infty(X)$ of bounded functions on X. To be more precise, according to [76, Lemma 4.23] the ℓ_∞-embedding

$$I_{\boldsymbol{\sigma}}[X] \colon H_{\boldsymbol{\sigma}}(X) \to \ell_\infty(X), \; f \mapsto f \tag{7.1}$$

is a bounded operator with operator norm $\|I_{\boldsymbol{\sigma}}[X]\| = 1$.

The goal of this part is to give bounds on the capacity of the Gaussian RKHS $H_{\boldsymbol{\sigma}}(X)$ in terms of log-covering number bounds for the ℓ_∞-embedding $I_{\boldsymbol{\sigma}}[X]$. To this end, recall that, for $\varepsilon > 0$, the *covering number* $\mathcal{N}(I_{\boldsymbol{\sigma}}[X], \varepsilon)$ is defined as the minimum number of closed $\ell_\infty(X)$-balls of radius ε needed to cover the closed unit ball $B_{H_{\boldsymbol{\sigma}}(X)} \subseteq \ell_\infty(X)$ of the Gaussian RKHS $H_{\boldsymbol{\sigma}}(X)$. Moreover, we denote log-covering numbers by

$$\mathcal{H}\big(I_{\boldsymbol{\sigma}}[X], \varepsilon\big) \coloneqq \log \mathcal{N}\big(I_{\boldsymbol{\sigma}}[X], \varepsilon\big) \enspace ,$$

see e.g. Appendix C for basic properties of (log-)covering numbers.

Finally, for a scalar $\sigma > 0$ we denote the *(isotropic) Gaussian kernel* on $\ell_2(I)$ of width σ by $k_\sigma \coloneqq k_{\boldsymbol{\sigma}}$ with $\boldsymbol{\sigma} = (\sigma, \sigma, \ldots)$. Analogously, we denote the corresponding RKHS by $H_\sigma(X)$ and the ℓ_∞-embedding of the isotropic Gaussian kernel on $X \subseteq \ell_2(I)$ by $I_\sigma[X]$.

7.2 Special Functions

In this section we introduce functions we use at several places. Especially, we give a brief introduction to a generalization of the binomial coefficient and Lambert's W-function. To this end, recall that for f, g defined in some neighborhood of $a \in \mathbb{R} \cup \{\pm\infty\}$ we write $f(t) \sim g(t)$ as $t \to a$ for the *strong asymptotic equivalence*, i.e. if $f(t)/g(t) \to 1$ for $t \to a$.

For $t > 0$ and $k \in \mathbb{N}$, we define the *generalized binomial coefficient* by

$$\binom{t}{k} := \frac{1}{k!} \prod_{i=1}^{k} (t - k + i) \ .$$

Note that this definition coincides with the classical definition for $t \in \mathbb{N}$. In the following generalized binomial coefficients mainly appear, for $d \in \mathbb{N}$ and $t > 0$, in the form

$$\binom{t+d}{d} = \frac{1}{d!} \prod_{i=1}^{d} (t + i) \ . \tag{7.2}$$

For the first lemma, which summarizes important monotonicity properties of (7.2), we use the *Gamma function* $\Gamma(t) := \int_0^\infty x^{t-1} e^{-x} \, \mathrm{d}x$ for $t > 0$.

7.2.1 Lemma (Binomial Coefficient) *For an integer $d \geq 1$ and a real number $t > 0$ the (generalized) binomial coefficient from* (7.2) *satisfies*

$$\binom{t+d}{d} = \frac{\Gamma(t+d+1)}{\Gamma(t+1)\Gamma(d+1)} \ .$$

Moreover, the following statements are true:

(i) For a fixed $d \geq 1$ the function $t \mapsto \binom{t+d}{d}$ is increasing and the function $t \mapsto \binom{t+d}{d} t^{-d}$ is decreasing with $\binom{t+d}{d} t^{-d} \to 1/d!$ for $t \to \infty$.

(ii) For a fixed $t > 0$ the sequence $d \mapsto \binom{t+d}{d}$ is increasing and the sequence $d \mapsto \binom{t+d}{d} d^{-t}$ is decreasing with $\binom{t+d}{d} d^{-t} \to 1/\Gamma(t+1)$ for $d \to \infty$.

Using the notion of strong asymptotic equivalence this lemma states

$$\begin{aligned} \binom{t+d}{d} &\sim \frac{t^d}{d!} && \text{for } t \to \infty && \text{and} \\ \binom{t+d}{d} &\sim \frac{d^t}{\Gamma(t+1)} && \text{for } d \to \infty \ . \end{aligned} \tag{7.3}$$

Proof. First note that $\Gamma(d+1) = d!$ and an d-fold application of $\Gamma(t+1) =$

$t \cdot \Gamma(t)$ gives us the first assertion, namely

$$\binom{t+d}{d} = \frac{1}{d!}\prod_{i=1}^{d}(t+i) = \frac{\Gamma(t+1)\prod_{i=1}^{d}(t+i)}{\Gamma(d+1)\Gamma(t+1)} = \frac{\Gamma(d+t+1)}{\Gamma(d+1)\Gamma(t+1)} \quad .$$

(i) From (7.2) we directly see that $t \mapsto \binom{t+d}{d}$ is increasing and that $t \mapsto \binom{t+d}{d}t^{-d}$ is decreasing with limit $1/d!$.

(ii) Since $\frac{t+(d+1)}{d+1} > 1$ and

$$\binom{t+(d+1)}{d+1} = \binom{t+d}{d} \cdot \frac{t+(d+1)}{d+1} \quad ,$$

the sequence $d \mapsto \binom{t+d}{d}$ is increasing. In the following we use the abbreviation $a_d := \binom{t+d}{d} \cdot d^{-t}$ for $d \geq 1$. A well-known property of the Gamma function is $a_d \to 1/\Gamma(t+1)$ for $d \to \infty$, see e.g. [72, Equation (95)]. Hence it remains to show the monotonicity of $(a_d)_{d\geq 1}$. Since

$$a_{d+1} = \frac{\Gamma(d+t+2)}{\Gamma(d+2)\Gamma(t+1)}(d+1)^{-t} = a_d \cdot \Big(\frac{d}{d+1}\Big)^t \cdot \frac{d+t+1}{d+1} \quad ,$$

$(a_d)_{d\geq 1}$ is decreasing if and only if $\frac{d+t+1}{d+1} < (\frac{d+1}{d})^t$ is satisfied for all $d \geq 1$ and $t > 0$. In order to prove this we fix some $d \geq 1$ and show that

$$f_d(t) := \Big(1+\frac{1}{d}\Big)^t - \Big(1+\frac{t}{d+1}\Big) > 0$$

is satisfied for all $t > 0$. To this end, we calculate the derivatives

$$f_d'(t) = \Big(1+\frac{1}{d}\Big)^t \log\Big(1+\frac{1}{d}\Big) - \frac{1}{d+1} \qquad \text{and}$$
$$f_d''(t) = \Big(1+\frac{1}{d}\Big)^t \log^2\Big(1+\frac{1}{d}\Big) \quad .$$

Using $\log(1+x) \geq \frac{x}{1+x}$, which holds for all $x > -1$, for $x = 1/d$ we get

$$f_d'(0) = \log\Big(1+\frac{1}{d}\Big) - \frac{1}{d+1} \geq \frac{1/d}{1+1/d} - \frac{1}{d+1} = 0 \quad .$$

Together with $f_d''(t) > 0$ we get $f_d'(t) > 0$ for all $t > 0$. Finally, $f_d(0) = 0$ and $f_d'(t) > 0$ gives $f_d(t) > 0$ for all $t > 0$ and hence the assertion is proven. □

The next lemma demonstrates our general approach to establish explicit bounds for the binomial coefficient.

7.2.2 Lemma (Bound for the Binomial Coefficient) *Let $f, h\colon I \to (0,\infty)$ be functions defined on some interval $I \subseteq \mathbb{R}$ with $f(\varepsilon) \geq t_0$ and $h(\varepsilon) \leq f(\varepsilon)$ for all $\varepsilon \in I$ then the binomial coefficient satisfies, for $\varepsilon \in I$,*

$$\binom{h(\varepsilon)+d}{d} \leq \binom{t_0+d}{d} \cdot \left(\frac{f(\varepsilon)}{t_0}\right)^d \ .$$

Proof. In order to prove this statement we use the auxiliary function $a_d(t)\colon (0,\infty) \to (0,\infty)$ defined by

$$a_d(t) \coloneqq \binom{t+d}{d} \cdot t^{-d} = \frac{1}{d!} \prod_{i=1}^{d} (1 + i/t) \ .$$

Since $t \mapsto \binom{t+d}{d}$ is increasing and $t \mapsto a_d(t)$ is decreasing we get

$$\binom{h(\varepsilon)+d}{d} \leq \binom{f(\varepsilon)+d}{d} = a_d(f(\varepsilon)) \cdot f^d(\varepsilon) \leq a_d(t_0) \cdot f^d(\varepsilon)$$

for all $\varepsilon \in I$, which gives the assertion. □

Now, we give a brief introduction to *Lambert's W-function*, which is defined as the inverse of the function $f(t) \coloneqq te^t$. To be more precise, the function f is not bijective on $\mathbb{R}$ but, it is bijective as function $f_0 \coloneqq f|_{[-1,\infty)}\colon [-1,\infty) \to [-1/e,\infty)$ as well as function $f_{-1} \coloneqq f|_{(-\infty,-1]}\colon (-\infty,-1] \to [-1/e,0)$. Consequently, Lambert's W-function consists of two branches, namely

$$\begin{aligned} W_0 &\coloneqq f_0^{-1}\colon [-1/e,\infty) \to [-1,\infty) \qquad \text{and} \\ W_{-1} &\coloneqq f_{-1}^{-1}\colon [-1/e,0) \to (-\infty,-1] \end{aligned}$$

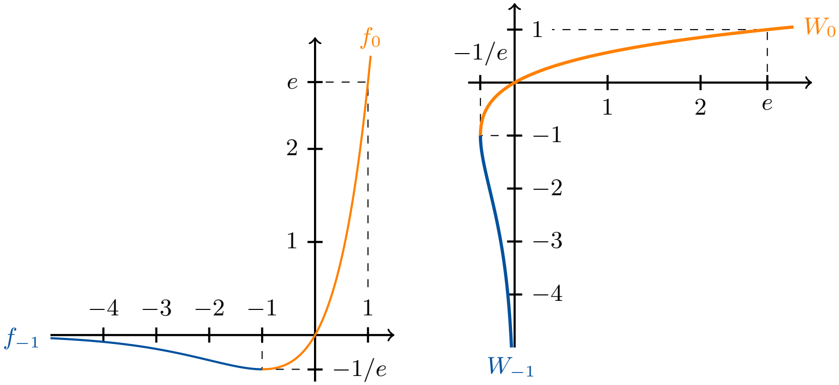


Figure 7.1: Plot of f_0 and f_{-1} on the left as well as plot of W_0 and W_{-1} on the right.

Note that for complex arguments there are even more branches, but we are only interested in real arguments. See Figure 7.1 for a plot of Lambert's W-function. The following lemma summarizes basic properties of W_0.

7.2.3 Lemma (Lambert's W_0-Function) *Lambert's W_0-function is differentiable on the interval $(-1/e, \infty)$ with derivative*

$$W_0'(x) = \frac{\exp\bigl(-W_0(x)\bigr)}{1 + W_0(x)} \overset{x \neq 0}{=} \frac{W_0(x)}{x\bigl(1 + W_0(x)\bigr)} \quad .$$

Moreover, the following statements are true:

(i) *W_0 is increasing.*

(ii) *$W_0(ye^y) = y$ for $y \geq -1$.*

(iii) *$W_0(x)e^{W_0(x)} = x$ for $x \geq -1/e$.*

(iv) *$W_0(0) = 0$ and $W(x) > 0$ for $x > 0$.*

(v) *$W_0(x) \sim \log(x)$ for $x \to \infty$.*

Proof. First note, that f_0 is differentiable on $[-1, \infty)$ with derivative $f_0'(y) = e^y(1 + y)$. Consequently, f_0 is increasing on $[-1, \infty)$ and hence the inverse

f_0^{-1} exists on $f_0([-1,\infty)) = [-1/e,\infty)$. In other words, W_0 is well-defined. Moreover, the monotonicity of f_0 already implies Point (i). Point (ii) and (iii) are explicit descriptions of the fact that W_0 is the inverse of f_0. Moreover, $f_0' > 0$ on $(-1,\infty)$ implies that W_0 is differentiable with derivative

$$W_0'(x) = \frac{1}{f_0'\big(W_0(x)\big)} = \frac{\exp\big(-W_0(x)\big)}{1+W_0(x)}$$

for all $x > -1/e$. For $x \neq 0$ an application of Point (iii) yields the second representation of the derivative. Point (iv) follows from $f_0(0) = 0$ and $f_0(y) > 0$ for $y > 0$. Finally, an application of L'Hôpital's rule

$$\lim_{x\to\infty} \frac{W_0(x)}{\log(x)} = \lim_{x\to\infty} \frac{W_0'(x)}{\log'(x)} = \lim_{x\to\infty} \frac{W_0(x)}{1+W_0(x)} = 1$$

gives Point (v). □

Analogously to the previous lemma, the next lemma summarizes basic properties of W_{-1}.

7.2.4 Lemma (Lambert's W_{-1}-Function) *Lambert's W_{-1}-function is differentiable on the interval $(-1/e, 0)$ with derivative*

$$W_{-1}'(x) = \frac{\exp\big(-W_{-1}(x)\big)}{1+W_{-1}(x)} = \frac{W_{-1}(x)}{x\big(1+W_{-1}(x)\big)} \quad .$$

Moreover, the following statements are true:

(i) W_{-1} *is decreasing.*

(ii) $W_{-1}(ye^y) = y$ *for* $y \leq -1$.

(iii) $W_{-1}(x)e^{W_{-1}(x)} = x$ *for* $-1/e \leq x < 0$.

(iv) $W_{-1}(-x) \sim \log(x)$ *for* $x \to 0^+$.

Proof. This can be proven analogously to Lemma 7.2.3. □

The next lemma presents an explicit bound comparing W_0 with a scaled version of the logarithm.

7.2.5 Lemma (Lambert's W_0-function vs. Logarithm) *For $\sigma > 0$ let $t^* := \sigma^{-2}\exp(\sigma^{-2})$ and $q_\sigma\colon (0,\infty) \to \mathbb{R}$ be given by*

$$q_\sigma(t) := \frac{\log(t \cdot e\sigma^2)}{W_0(t)} \quad .$$

Then q_σ is increasing on $(0, t^*]$ and decreasing on $[t^*, \infty)$. Moreover, q_σ has a unique global maximum at t^* with $q_\sigma(t^*) = 1 + \sigma^2$ and we have $\lim_{t\to\infty} q_\sigma(t) = 1$.

Proof. A simple but tedious calculation shows

$$q'_\sigma(t) = \frac{W_0(t) - \log(t\sigma^2)}{t \cdot W_0(t) \cdot \big(1 + W_0(t)\big)} \quad .$$

Since the denominator is positive for all $t > 0$ we can focus on the numerator in order to investigate the monotonicity properties of q_σ. Consequently, q_σ is decreasing, if and only if $W_0(t) < \log(t\sigma^2)$ and this is equivalent to

$$t = W_0(t)e^{W_0(t)} < \log(t\sigma^2) \cdot \exp \circ \log(t\sigma^2) = t\sigma^2 \log(t\sigma^2) \quad .$$

Rearranging this inequality for t shows that q_σ is decreasing on $[t^*, \infty)$. Analogously, we get that q_σ is increasing on $(0, t^*]$ and that q_σ has a unique global maximum at t^*. Since $W_0(t^*) = \sigma^{-2}$ and $\log(t^*) = -\log(\sigma^2) + \sigma^{-2}$ we find $q_\sigma(t^*) = 1 + \sigma^2$. Finally, $\lim_{t\to\infty} q_\sigma(t) = 1$ directly follows from $W_0(t) \sim \log(t)$ for $t \to \infty$, see Point (v) of Lemma 7.2.3. □

In the following lemma we use Lambert's W-function to introduce important functions that appear in our log-covering number bounds.

7.2.6 Lemma (Auxiliary Functions) *For $\sigma > 0$, $x > 0$, $y \geq -\sigma^2$,*

$$p_\sigma(x) := 2\bigg(\frac{2e\sigma^2}{x}\bigg)^{x/2} \quad , \qquad \text{and} \qquad h_\sigma(y) := 2e\sigma^2 \exp\bigg(W_0\Big(\frac{y}{e\sigma^2}\Big)\bigg)$$

the following statements are true:

(i) The function $p_\sigma \colon (0,\infty) \to (0,\infty)$ *is decreasing on* $(2\sigma^2,\infty)$ *and* $p_\sigma(x) \to 0$ *for* $x \to \infty$.

(ii) The function $h_\sigma \colon [-\sigma^2,\infty) \to [2\sigma^2,\infty)$ *is increasing and satisfies*

$$h_\sigma(y) = \frac{2y}{W_0\bigl(\frac{y}{e\sigma^2}\bigr)} \quad . \tag{7.4}$$

(iii) The function $p_\sigma \colon \bigl[2\sigma^2,\infty\bigr) \to \bigl(0, 2\exp(\sigma^2)\bigr]$ *is bijective with inverse* p_σ^{-1} *given by*

$$p_\sigma^{-1}(\varepsilon) = h_\sigma \circ \log(2/\varepsilon) \ .$$

Proof. (i) Some tedious calculations show that the derivative of p_σ, for $x > 0$, is given by

$$p'_\sigma(x) = \frac{p_\sigma(x)}{2} \log\biggl(\frac{2\sigma^2}{x}\biggr) \ .$$

From this identity for the derivative of p_σ the first assertion immediately follows. The second assertion is obvious.

(ii) The monotonicity of h_σ is a consequence of the monotonicity of W_0 and the definition of the function h_σ. Moreover, (7.4) follows from the identity $W_0(x)\exp(W_0(x)) = x$.

(iii) By Point (i) we already know that $p_\sigma \colon \bigl[2\sigma^2,\infty\bigr) \to \bigl(0, 2\exp(\sigma^2)\bigr]$ is bijective. To verify the formula for p_σ^{-1}, we fix some $0 < \varepsilon \leq 2\exp(\sigma^2)$ and write $y := \log(2/\varepsilon)$. This immediately gives $y \geq -\sigma^2$ and by the definition of h_σ we find

$$p_\sigma \circ h_\sigma(y) = 2\biggl(\frac{2e\sigma^2}{h_\sigma(y)}\biggr)^{h_\sigma(y)/2} = 2\exp\biggl(-W_0\Bigl(\frac{y}{e\sigma^2}\Bigr) \cdot \frac{h_\sigma(y)}{2}\biggr) = 2e^{-y} = \varepsilon \ ,$$

i.e. we have shown the assertion. □

Finally, the function considered in the next lemma appears at several places of Section 8.2.

7.2.7 Lemma *Let $t^* := e$ and the function $\beta\colon (1,\infty) \to (0,\infty)$ be defined by*

$$\beta(t) := \frac{t}{\log(t)} \quad .$$

Then β is decreasing on $(1, t^*]$ and increasing on $[t^*, \infty)$. Moreover, β has a unique global minimum at t^* with $\beta(t^*) = e$

Proof. Since the derivative of β equals

$$\beta'(t) = \frac{\log(t) - 1}{\log^2(t)} \quad ,$$

the assertion directly follows. □

Chapter 8

Gaussian Kernels on Finite-Dimensional Spaces

In this chapter we consider Gaussian RKHSs $H_\sigma(X)$ on bounded subsets $X \subseteq \mathbb{R}^d$ of the finite-dimensional space $\mathbb{R}^d$. To be more precise, we provide bounds on the log-covering number for the ℓ_∞-embedding $I_\sigma[X]$ of the Gaussian RKHS $H_\sigma(X)$ defined in (7.1). Unlike previous results in this direction we focus on small explicit constants as well as their dependence on crucial parameters such as the kernel width and the size and dimension of the underlying space. The content of this chapter is mainly taken from the article:

> I. Steinwart and S. Fischer. A closer look at covering number bounds for Gaussian kernels. *J. Complexity*, 62:101513, 2021.

8.1 A closer Look at Kühn's Proof

In this section we essentially repeat the key arguments of [59, Theorem 3] on the input space $X = B_{\ell_2^d}$ instead of $X = [0,1]^d$. Moreover, in contrast to [59, Theorem 3] we precisely keep track of the appearing constants.

Throughout this section the domain $X \coloneqq B_{\ell_2^d} \subseteq \mathbb{R}^d$ is fixed, and hence we simply write I_σ for the embedding $I_\sigma[B_{\ell_2^d}]\colon H_\sigma(B_{\ell_2^d}) \to \ell_\infty(B_{\ell_2^d})$. Our first result provides a general estimate for the log-covering numbers of I_σ.

8.1.1 Lemma (Kühn's Bound) *For $\sigma, \varepsilon > 0$ and integers $N \geq 1$ the following bound is satisfied*

$$\mathcal{H}\left(I_\sigma, \varepsilon + \sqrt{\frac{(2\sigma^2)^N}{N!}}\right) \leq \binom{N-1+d}{d} \cdot \log(1+2/\varepsilon) \ .$$

Proof. For fixed $\sigma > 0$, $\varepsilon > 0$, and $N \geq 1$ we define

$$\varepsilon_0 := \sqrt{\frac{(2\sigma^2)^N}{N!}} \ .$$

In order to repeat the argument of [59, Theorem 3], we begin by recalling some notation: For every multi-index $k = (k_1, \dots, k_d) \in \mathbb{N}_0^d$ we define the function $e_k : B_{\ell_2^d} \to \mathbb{R}$ by

$$e_k(x) := \sqrt{\frac{(2\sigma^2)^{|k|}}{k!}} x^k \exp\bigl(-\sigma^2 \|x\|_{\ell_2^d}^2\bigr)$$

where we use $|k| := k_1 + \ldots + k_d$, $k! := k_1! \cdot \ldots \cdot k_d!$, and $x^k := x_1^{k_1} \cdot \ldots \cdot x_d^{k_d}$ for $x = (x_1, \dots, x_d) \in B_{\ell_2^d}$. Since $B_{\ell_2^d}$ has a non-empty interior, the family of functions $(e_k)_{k \in \mathbb{N}_0^d}$ forms an ONB of $H_\sigma(B_{\ell_2^d})$ according to [76, Theorem 4.42]. Using this ONB we now consider, for $N \geq 1$, the orthogonal projections $P_N, Q_N \colon H_\sigma(B_{\ell_2^d}) \to H_\sigma(B_{\ell_2^d})$ onto $\overline{\operatorname{span}}\bigl\{e_k : \ |k| < N\bigr\}$ and $\overline{\operatorname{span}}\bigl\{e_k : \ |k| \geq N\bigr\}$, respectively. From the first equation on page 494 of [59] we know

$$\|I_\sigma \circ Q_N\| \leq \sup_{x \in B_{\ell_2^d}} \sqrt{\frac{\bigl(2\sigma^2 \|x\|_{\ell_2^d}^2\bigr)^N}{N!}} = \sqrt{\frac{(2\sigma^2)^N}{N!}} = \varepsilon_0 \ . \tag{8.1}$$

Using (C.3), Point (vi) of Lemma C.9, and $\|I_\sigma \circ P_N\| = 1$ we find

$$\begin{aligned} \mathcal{H}\bigl(I_\sigma, \varepsilon + \varepsilon_0\bigr) &= \mathcal{H}\Bigl(I_\sigma \circ P_N + I_\sigma \circ Q_N, \varepsilon + \varepsilon_0\Bigr) \\ &\leq \mathcal{H}\bigl(I_\sigma \circ P_N, \varepsilon\bigr) \\ &\leq \operatorname{rank}(P_N) \log(1 + 2/\varepsilon) \ . \end{aligned}$$

Together with the formula

$$\operatorname{rank}(P_N) = \binom{N-1+d}{d} \ ,$$

which was derived in [59, Remark 4], we thus obtain the assertion. □

Our next goal is to find suitable values of $N \geq 1$ for the bound established in Lemma 8.1.1. To this end, we use the functions p_σ and h_σ introduced in Lemma 7.2.6.

8.1.2 Lemma (Log-Covering Number Bound) *For $\sigma > 0$ and $0 < \varepsilon \leq 1$, the following bound is satisfied*

$$\mathcal{H}(I_\sigma, \varepsilon) \leq \binom{(h_\sigma \circ \log)(4/\varepsilon) + d}{d} \cdot \log(4/\varepsilon) \ .$$

This lemma is the basis for all log-covering number bounds presented in Chapter 8 as well as Chapter 9 below.

Proof. For a fixed $0 < \varepsilon \leq 1$ we write $y := \log(4/\varepsilon)$ and $x := h_\sigma(y)$. Since $y > 1$ holds true, we have $x > 2\sigma^2 > 0$, and hence there is a unique integer $N \geq 1$ with $N - 1 < x \leq N$. Using Lemma 8.1.1 with $2\varepsilon/3$ instead of ε, the monotonicity of $t \mapsto \binom{t+d}{d}$, and $1 \leq 1/\varepsilon$ we find

$$\begin{aligned} \mathcal{H}\biggl(I_\sigma, \frac{2\varepsilon}{3} + \sqrt{\frac{(2\sigma^2)^N}{N!}}\biggr) &\leq \binom{N-1+d}{d} \cdot \log(1+3/\varepsilon) \\ &\leq \binom{x+d}{d} \cdot \log(4/\varepsilon) \ . \end{aligned}$$

Consequently, it remains to show that $\sqrt{(2\sigma^2)^N/N!} \leq \varepsilon/3$ holds true. To this end, we use Stirling's formula $N! \geq \sqrt{2\pi N}\,(N/e)^N$ to get

$$\sqrt{\frac{(2\sigma^2)^N}{N!}} \leq \frac{1}{(2\pi N)^{1/4}} \cdot \left(\frac{2e\sigma^2}{N}\right)^{N/2} \leq \frac{p_\sigma(N)}{2(2\pi)^{1/4}} \ .$$

Moreover, the already observed $x > 2\sigma^2$ together with Point (i) and (iii) of

Lemma 7.2.6 yields

$$p_\sigma(N) \le p_\sigma(x) = p_\sigma\big(h_\sigma(y)\big) = p_\sigma\big(h_\sigma \circ \log(4/\varepsilon)\big) = \varepsilon/2 \ .$$

Combining both estimates and $(2\pi)^{-1/4} \le 4/3$ we get the assertion. □

Note that by an easy adaption of the above proof we can replace the 4 in $y = \log(4/\varepsilon)$ by $\gamma = 7/2$ if we choose $4\varepsilon/5$ instead of $2\varepsilon/3$ and use the bound $(2\pi)^{-1/4} \approx 0.6316 \le 7/10$. Moreover, some tedious calculations show that the argument still works for

$$\gamma := \frac{3(2\pi)^{1/4} + 1 + \sqrt{9(2\pi)^{1/2} + 2(2\pi)^{1/4} + 1}}{2(2\pi)^{1/4}} \approx 3.4485 \ .$$

Since these improvements have little impact we stick to $\gamma = 4$ for convenience.

8.2 Isotropic Gaussian Kernels on $B_{\ell_2^d}$

In this section we exploit the log-covering number bound from Lemma 8.1.2 to derive further bounds which are (probably) easier to interpret. Recall that Lemma 8.1.2 is valid for the domain $X = B_{\ell_2^d}$ and isotropic Gaussian kernels with width $\sigma > 0$. Moreover, we use again the abbreviation $I_\sigma := I_\sigma[B_{\ell_2^d}]$ in the proofs.

8.2.1 Theorem *For $d \ge 1$, $\sigma > 0$, $0 < \varepsilon \le 1$, and*

$$K_{d,\sigma} := \binom{2e(1+\sigma^2) + d}{d} \cdot e^{-d}$$

the following log-covering number bound is satisfied

$$\mathcal{H}\big(I_\sigma[B_{\ell_2^d}], \varepsilon\big) \le K_{d,\sigma} \cdot \frac{\log^{d+1}(4/\varepsilon)}{\log\log^d(4/\varepsilon)} \ .$$

Note that this result recovers the asymptotic behavior of $\varepsilon \mapsto \mathcal{H}\big(I_\sigma[B_{\ell_2^d}], \varepsilon\big)$ found by Kühn in [59], which in turn improves the upper bound of [87, Proposition 1]. Moreover, by presenting a corresponding lower bound on the log-covering numbers, [59] further shows that this behavior in ε is optimal and hence improves the lower bound of [88, Proposition 1]. Unlike the upper bound in [59], however, Theorem 8.2.1 provides an upper bound for the behavior in σ and d that is expressed by the constant $K_{d,\sigma}$.

Proof. Let us define $\varepsilon_0 := 1$ and $y_0 := \log(4/\varepsilon_0)$. For $0 < \varepsilon \leq \varepsilon_0$ we further write $y := \log(4/\varepsilon) \geq y_0 > 1$. Moreover, recall the representation of $h_\sigma(y) = 2y/W_0\big(\frac{y}{e\sigma^2}\big)$ in Lemma 7.2.6. An application of Lemma 7.2.5 then yields

$$\begin{aligned}(h_\sigma \circ \log)(4/\varepsilon) &= \frac{2y}{W_0\big(\frac{y}{e\sigma^2}\big)} \\ &= \frac{2y}{\log(y)} \cdot \frac{\log\big(\frac{y}{e\sigma^2} \cdot e\sigma^2\big)}{W_0\big(\frac{y}{e\sigma^2}\big)} \\ &\leq 2\,(1+\sigma^2) \cdot \frac{y}{\log(y)} =: f(\varepsilon)\end{aligned}$$

for all $0 < \varepsilon \leq \varepsilon_0$. According to Lemma 7.2.7 we have $y/\log(y) \geq e$ and hence we get $f(\varepsilon) \geq 2e(1+\sigma^2) =: t_0$. Finally, combining Lemma 8.1.2 and Lemma 7.2.2 for $h = (h_\sigma \circ \log)(4/\,\cdot\,)$ gives the assertion. □

To better understand the behavior of the constant $K_{d,\sigma}$ defined in Theorem 8.2.1 we fix some $\sigma > 0$ and consider its asymptotic behavior for $d \to \infty$. From (7.3) we get

$$K_{d,\sigma} = \binom{2e(1+\sigma^2)+d}{d} \cdot e^{-d} \sim \frac{d^{2e(1+\sigma^2)}}{\Gamma(2e(1+\sigma^2)+1)} \cdot e^{-d}$$

for $d \to \infty$. Moreover, the monotonicity properties of the binomial coefficient presented in Lemma 7.2.1 gives the upper bound

$$\begin{aligned}K_{d,\sigma} &\leq C_{d_0,\sigma} \cdot d^{2e(1+\sigma^2)} e^{-d} \qquad \text{with} \\ C_{d_0,\sigma} &:= \binom{2e(1+\sigma^2)+d_0}{d_0} \cdot d_0^{-2e(1+\sigma^2)}\end{aligned} \tag{8.2}$$

for all $d \geq d_0 \geq 1$. Note that the smaller the range of d, i.e. the larger d_0, the smaller the constant $C_{d_0,\sigma}$. In the case $\sigma = 1$, which is of special interest in Section 8.4 below, we have $C_{1,1} = 4e+1 \approx 11.8731$ and $C_{2,1} \approx 0.0407 \leq 0.05$, i.e. for $d \geq 2$ we have

$$K_{d,1} \leq C_{2,1} \cdot d^{4e} e^{-d} \leq 0.05 \cdot d^{4e} e^{-d} \quad .$$

For a plot of $d \mapsto K_{d,1}$ see Figure 8.1.

Now, let us consider the behavior of $K_{d,\sigma}$ in σ for some fixed $d \geq 1$. To this end, we first observe that (7.3) gives

$$K_{d,\sigma} = \binom{2e(1+\sigma^2)+d}{d} \cdot e^{-d} \sim \frac{\big(2e(1+\sigma^2)\big)^d}{d!} \cdot e^{-d} \sim \frac{2^d}{d!} \cdot \sigma^{2d}$$

for $\sigma \to \infty$. Consequently the constant $K_{d,\sigma}$ grows like σ^{2d} for $\sigma \to \infty$. This improves the bound in [87, Proposition 1], which provides an σ^{2d+2}-behavior. However, in Section 8.4 below we will see that we can find another constant for the estimate of Theorem 8.2.1 that only grows like σ^d for $\sigma \to \infty$. Moreover, $K_{d,\sigma}$ is increasing in σ and the representation of the binomial coefficient in (7.2) directly gives, for all $\sigma > 0$ satisfying $2e(1+\sigma^2) \geq d$, the bounds

$$\frac{2^d}{d!}\big(1+\sigma^2\big)^d \leq K_{d,\sigma} \leq \frac{4^d}{d!}\big(1+\sigma^2\big)^d \quad . \tag{8.3}$$

Our next goal is to show that the constant in Theorem 8.2.1 is significantly influenced by the considered range of ε. More precisely, Theorem 8.2.1 considers the maximal range $0 < \varepsilon \leq 1$, since we have $\|I_\sigma[B_{\ell_2^d}]\| = 1$, and thus Point (ii) of Lemma C.9 gives

$$\mathcal{H}\big(I_\sigma[B_{\ell_2^d}], \varepsilon\big) = 0$$

for all $\varepsilon \geq 1$. Our next theorem shows that by considering a smaller range for ε, we can substantially decrease the constant appearing in the estimate of Theorem 8.2.1.

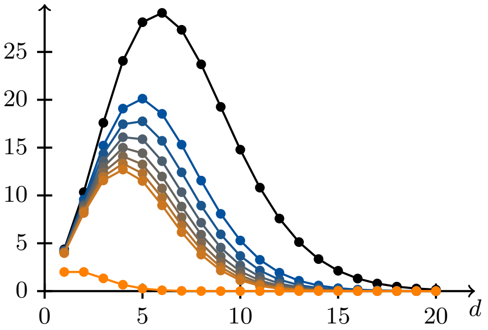


Figure 8.1: Plot of $K_{d,1}$ from Theorem 8.2.1 and $K_{d,1,\varepsilon_0}$ from Theorem 8.2.2 for different values of ε_0 as well as in the limit case $\varepsilon_0 \to 0^+$ from (8.4).

8.2.2 Theorem (Small ε-Range) *For $d \geq 1$, $\sigma > 0$, and $0 < \varepsilon_0 \leq 4 \cdot \exp\bigl(-e^{1+\sigma^{-2}}\bigr)$ consider $y_0 := \log(4/\varepsilon_0)$, $x_0 := 2y_0/W_0\bigl(\frac{y_0}{e\sigma^2}\bigr)$, and*

$$K_{d,\sigma,\varepsilon_0} := \binom{x_0+d}{d} \cdot \left(\frac{\log(y_0)}{y_0}\right)^d \quad .$$

Then the following log-covering number bound is satisfied, for $0 < \varepsilon \leq \varepsilon_0$,

$$\mathcal{H}\bigl(I_\sigma[B_{\ell_2^d}], \varepsilon\bigr) \leq K_{d,\sigma,\varepsilon_0} \cdot \frac{\log^{d+1}(4/\varepsilon)}{\log\log^d(4/\varepsilon)} \quad .$$

Proof. For $0 < \varepsilon \leq \varepsilon_0$ we write $y := \log(4/\varepsilon) \geq y_0$. Note that the restriction on ε_0 ensures $y_0 \geq \exp\bigl(1+\sigma^{-2}\bigr)$ and hence $\frac{y_0}{e\sigma^2} \geq \sigma^{-2}\exp\bigl(\sigma^{-2}\bigr)$. As a consequence, the function $y \mapsto \log(y)/W_0\bigl(\frac{y}{e\sigma^2}\bigr)$ is decreasing on $[y_0, \infty)$ according to Lemma 7.2.5 and we get

$$(h_\sigma \circ \log)(4/\varepsilon) = \frac{2\log(y)}{W_0\bigl(\frac{y}{e\sigma^2}\bigr)} \cdot \frac{y}{\log(y)} \leq \frac{2\log(y_0)}{W_0\bigl(\frac{y_0}{e\sigma^2}\bigr)} \cdot \frac{y}{\log(y)} =: f(\varepsilon) \quad .$$

Now, from Lemma 7.2.7 we know that the function $\beta(t) = t/\log(t)$ is increasing on $[e, \infty)$ and the already observed $y_0 \geq \exp\bigl(1+\sigma^{-2}\bigr) > e$ gives us

$$f(\varepsilon) \geq \frac{2\log(y_0)}{W_0\bigl(\frac{y_0}{e\sigma^2}\bigr)} \cdot \frac{y_0}{\log(y_0)} = \frac{2y_0}{W_0\bigl(\frac{y_0}{e\sigma^2}\bigr)} = x_0$$

for all $0 < \varepsilon \leq \varepsilon_0$. Finally, combining Lemma 8.1.2 and Lemma 7.2.2 for $h = (h_\sigma \circ \log)(4/\,\cdot\,)$ and $t_0 = x_0$ gives the assertion. □

To appreciate Theorem 8.2.2 we note that for $\varepsilon_0 \to 0^+$ we have $y_0 \to \infty$ and $x_0 \to \infty$. Since $\binom{t+d}{d} \sim t^d/d!$ and $W_0(t) \sim \log(t)$ for $t \to \infty$, see Lemma 7.2.1 and Lemma 7.2.3, respectively, we find

$$\lim_{\varepsilon_0 \to 0^+} K_{d,\sigma,\varepsilon_0} = \lim_{\varepsilon_0 \to 0^+} \binom{x_0+d}{d} \cdot x_0^{-d} \cdot \left(\frac{2\log(y_0)}{W_0\bigl(\frac{y_0}{e\sigma^2}\bigr)}\right)^d = \frac{2^d}{d!} \quad . \tag{8.4}$$

This sharpens the result of [59, Remark 4] by a factor of approximately $\sqrt{2\pi d}$. For fixed $\sigma > 0$ and $0 < \varepsilon_0 \leq 4\exp\bigl(-e^{1+\sigma^{-2}}\bigr)$ the quantity $K_{d,\sigma,\varepsilon_0}$ behaves like

$$K_{d,\sigma,\varepsilon_0} \sim \frac{d^{x_0}}{\Gamma(x_0+1)} \cdot \left(\frac{\log(y_0)}{y_0}\right)^d$$

for $d \to \infty$. Since $\log(y_0)/y_0 < e^{-1}$ is satisfied according to Lemma 7.2.7, the quantity $K_{d,\sigma,\varepsilon_0}$ vanishes faster than the quantity $K_{d,\sigma}$ of Theorem 8.2.1 for $d \to \infty$. Finally, note that, for $\sigma = 1$ and $\varepsilon_0 := 4\exp(-e^2) \approx 0.0025$, we have $y_0 = e^2$ and $x_0 = 2e^2$. Using the monotonicity of $d \mapsto \binom{x_0+d}{d} d^{-x_0}$ we find, for $d \geq 1$,

$$K_{d,1,\varepsilon_0} \leq (1+2e^2) \cdot d^{2e^2} \cdot (2/e^2)^d \leq 16 \cdot d^{2e^2} \cdot (2/e^2)^d \quad .$$

For a plot of $d \mapsto K_{d,1,\varepsilon_0}$ for different values of ε_0 see Figure 8.1.

For some applications, see e.g. [84, 34], it is sufficient and more convenient to work with a weaker bound in ε, namely a polynomial bound in ε. For this reason, the following theorem establishes a polynomial upper bound with an explicit constant.

8.2.3 Theorem (Polynomial Bound) *For $d \geq 1$, $\sigma > 0$, and $p > 0$ consider*

$$K_{d,\sigma,p} := \binom{t_0+d}{d} \cdot \frac{d+1}{ep} \cdot 4^{\frac{p}{d+1}} \qquad \textit{with}$$

$$t_0 := \frac{2(d+1)\cdot 4^{\frac{p}{d+1}}}{ep \cdot W_0\big(\frac{d+1}{p\sigma^2}\big)} \exp\Bigg(\frac{1}{W_0\big(\frac{d+1}{p\sigma^2}\big)}\Bigg) \ .$$

Then the following log-covering number bound is satisfied, for $0 < \varepsilon \leq 1$,

$$\mathcal{H}\big(I_\sigma[B_{\ell_2^d}], \varepsilon\big) \leq K_{d,\sigma,p} \cdot \varepsilon^{-p} \ .$$

Proof. For $0 < \varepsilon \leq 1$ we again write $y := \log(4/\varepsilon) \geq \log(4)$. In order to give a polynomial upper bound for $\mathcal{H}(I_\sigma, \varepsilon)$ we use Lemma 8.1.2 and estimate the two factors, $\binom{h_\sigma(y)+d}{d}$ and $y = \log(4/\varepsilon)$, appearing in Lemma 8.1.2, separately by a polynomial bound. To bound the first factor we fix $q_1 > 0$ and define the function

$$g_1(t) := 2\frac{te^{-q_1 t}}{W_0\big(\frac{t}{e\sigma^2}\big)}$$

for $t > 0$. Using $e^{-q_1 y} = (4/\varepsilon)^{-q_1}$ we then get

$$(h_\sigma \circ \log)(4/\varepsilon) = \frac{2y}{W_0\big(\frac{y}{e\sigma^2}\big)} \Big(\frac{4}{\varepsilon}\Big)^{-q_1} \cdot \Big(\frac{4}{\varepsilon}\Big)^{q_1} \leq \Big(\frac{4}{\varepsilon}\Big)^{q_1} \sup_{t>0} g_1(t) =: f(\varepsilon)$$

and $f(\varepsilon) \geq 4^{q_1} \cdot \sup_{t>0} g_1(t) =: t_1$. A simple but tedious calculation shows

$$g_1'(t) = \frac{g_1(t)}{1+W_0\big(\frac{t}{e\sigma^2}\big)} \Bigg(\sigma^{-2} \exp\Bigg(-\Big(1+W_0\Big(\frac{t}{e\sigma^2}\Big)\Big)\Bigg) - q_1\Big(1+W_0\Big(\frac{t}{e\sigma^2}\Big)\Big)\Bigg) \ .$$

Another tedious calculation shows that g_1 has a unique global maximum at

$$t^* := \frac{1}{q_1}\Bigg(1 - \frac{1}{W_0\big(\frac{1}{q_1\sigma^2}\big)}\Bigg) \ .$$

Using $t/W_0(t) = \exp\bigl(W_0(t)\bigr)$ for $t = \frac{1}{q_1\sigma^2}$ we find

$$\begin{aligned} t^* &= \left(W_0\Bigl(\frac{1}{q_1\sigma^2}\Bigr) - 1\right) \cdot \frac{1/q_1}{W_0\bigl(\frac{1}{q_1\sigma^2}\bigr)} \\ &= \sigma^2 \cdot \left(W_0\Bigl(\frac{1}{q_1\sigma^2}\Bigr) - 1\right) \cdot \exp \circ W_0\Bigl(\frac{1}{q_1\sigma^2}\Bigr) \ . \end{aligned}$$

This representation together with $W_0(xe^x) = x$ for $x = W_0\bigl(\frac{1}{q_1\sigma^2}\bigr) - 1$ implies

$$\frac{t^*}{W_0\bigl(\frac{t^*}{e\sigma^2}\bigr)} = \frac{t^*}{W_0\bigl(\frac{1}{q_1\sigma^2}\bigr) - 1} = \frac{1}{q_1 \cdot W_0\bigl(\frac{1}{q_1\sigma^2}\bigr)} \ .$$

Using this identity we directly get

$$\begin{aligned} t_1 = 4^{q_1} g_1(t^*) &= 2 \cdot 4^{q_1} \cdot \frac{e^{-q_1 t^*}}{q_1 \cdot W_0\bigl(\frac{1}{q_1\sigma^2}\bigr)} \\ &= \frac{2 \cdot 4^{q_1}}{e q_1 \cdot W_0\bigl(\frac{1}{q_1\sigma^2}\bigr)} \exp\left(1/W_0\Bigl(\frac{1}{q_1\sigma^2}\Bigr)\right) \end{aligned}$$

and Lemma 7.2.2 for $h = (h_\sigma \circ \log)(4/\,\cdot\,)$ and $t_0 = t_1$ gives us

$$\begin{aligned} \binom{(h_\sigma \circ \log)(4/\varepsilon) + d}{d} &\leq \binom{t_1 + d}{d} \cdot \left(\frac{f(\varepsilon)}{t_1}\right)^d \\ &= \binom{t_1 + d}{d} \cdot 4^{-q_1 d} \cdot (4/\varepsilon)^{q_1 d} \ . \end{aligned} \tag{8.5}$$

Now, we estimate the second factor $y = \log(4/\varepsilon)$ by a polynomial bound of order $q_2 > 0$. To this end, we define the function $g_2(t) \coloneqq te^{-q_2 t}$, for $t > 0$, and estimate

$$y = (4/\varepsilon)^{q_2} \cdot y \cdot (4/\varepsilon)^{-q_2} \leq (4/\varepsilon)^{q_2} \cdot \sup_{t>0} g_2(t) \ .$$

An easy calculation shows that the derivative of g_2 is given by $g_2'(t) = g_2(t) \cdot (1/t - q_2)$ and consequently g_2 has a global maximum at $t^* \coloneqq 1/q_2$

with $g_2(t^*) = \frac{1}{eq_2}$. Therefore, we get

$$y \le \frac{(4/\varepsilon)^{q_2}}{eq_2} \quad . \tag{8.6}$$

Finally, combining Lemma 8.1.2 with (8.5) and (8.6) yields

$$\mathcal{H}(I_\sigma, \varepsilon) \le \binom{t_1 + d}{d} \cdot \frac{1}{eq_2 \cdot 4^{q_1 d}} \cdot (4/\varepsilon)^{q_1 d + q_2} \quad ,$$

and for $q_1 = q_2 = \frac{p}{d+1}$ we find $t_1 = t_0$ as well as the claimed bound. □

Since we do not directly see the behavior of $K_{d,\sigma,p}$ in d, the next lemma shows that $K_{d,\sigma,p}$ grows more slowly than any exponential function in d, i.e. for all $a > 0$ we have $K_{d,\sigma,p} \cdot e^{-ad} \to 0$ for $d \to \infty$.

8.2.4 Lemma *For $\sigma, p > 0$ there are constants $c_{\sigma,p}, C_{\sigma,p} > 0$ such that $K_{d,\sigma,p}$ defined in Theorem 8.2.3 satisfies, for $d \ge 1$,*

$$K_{d,\sigma,p} \le C_{\sigma,p} \cdot \sqrt{d \log(d)} \cdot \exp\!\left(c_{\sigma,p} \cdot d \cdot \frac{\log\log(d)}{\log(d)} \right) \quad .$$

Proof. Using $W_0(t) \sim \log(t)$ from Lemma 7.2.3 we get, for $d \to \infty$,

$$t_0 = \frac{2(d+1) \cdot 4^{\frac{p}{d+1}}}{ep \cdot W_0\big(\frac{d+1}{p\sigma^2}\big)} \exp\!\left(\frac{1}{W_0\big(\frac{d+1}{p\sigma^2}\big)} \right) \asymp \frac{d}{W_0\big(\frac{d+1}{p\sigma^2}\big)} \asymp \frac{d}{\log(\frac{d+1}{p\sigma^2})} \asymp \frac{d}{\log(d)} \quad .$$

Since $t_0 \to \infty$ for $d \to \infty$, Lemma 7.2.1 together with Stirling's formula $\Gamma(t+1) \sim \sqrt{2\pi t}\,(t/e)^t$ yields

$$\binom{t_0 + d}{d} = \frac{\Gamma(t_0 + d + 1)}{\Gamma(t_0 + 1)\Gamma(d+1)} \asymp \Big(\frac{1}{t_0} + \frac{1}{d} \Big)^{1/2} \Big(1 + \frac{t_0}{d} \Big)^{d} \Big(1 + \frac{d}{t_0} \Big)^{t_0}$$

for $d \to \infty$. Using the inequality $1 + t \le e^t$, which holds for all $t \in \mathbb{R}$, for $t = t_0/d$ we get

$$\binom{t_0 + d}{d} \preccurlyeq \sqrt{\frac{\log(d)}{d}} \cdot e^{t_0} \cdot (1 + d/t_0)^{t_0}$$

for $d \to \infty$. Consequently, we find a constant $C_{\sigma,p} > 0$ with

$$K_{d,\sigma,p} = \binom{t_0+d}{d} \cdot \frac{d+1}{ep} \cdot 4^{\frac{p}{d+1}} \leq C_{\sigma,p} \cdot \sqrt{d \log(d)} \cdot \exp\biggl(t_0 \cdot \log\Bigl(e + e\frac{d}{t_0}\Bigr)\biggr) \ .$$

Since the exponent behaves like

$$t_0 \cdot \log\Bigl(e + e\frac{d}{t_0}\Bigr) \asymp \frac{d}{\log(d)} \cdot \log\Bigl(e + e\log(d)\Bigr) \asymp d \cdot \frac{\log\log(d)}{\log(d)}$$

for $d \to \infty$, there is a constant $c_{\sigma,p} > 0$ independent of d with the desired property. □

The next lemma establishes a bound on $K_{d,\sigma,p}$ in d and p for fixed $\sigma = 1$.

8.2.5 Lemma *For $0 < p_0 \leq 1/e$ and $\sigma = 1$ the quantity $K_{d,\sigma,p}$ defined in Theorem 8.2.3 satisfies, for $0 < p \leq p_0$ and $d \geq 1$,*

$$K_{d,1,p} \leq 1/2 \cdot C_{p_0}^d \cdot \sqrt{d} \cdot \frac{(1/p)^{d+1}}{\log^d(1/p)} \ ,$$

where C_{p_0} is given by

$$C_{p_0} := ep_0 \cdot \log(1/p_0) + (2+1/e)2^{1+p_0} \exp\biggl(\frac{1}{W_0(2/p_0)}\biggr) \ .$$

For fixed p, this bound is weaker in d than the bound in Lemma 8.2.4. However, this bound provides the dependence on p explicitly. In particular, C_{p_0} only depends on p_0, and for $p_0 := 1/e$ we find $C_{p_0} \approx 13.6481$. In addition, C_{p_0} converges to $4 + 2/e \approx 4.7358$ for $p_0 \to 0^+$.

Moreover, this lemma shows that Theorem 8.2.3 with $p = 1/\log(4/\varepsilon)$ recovers the optimal behavior of $\varepsilon \mapsto \mathcal{H}(I_1[B_{\ell_2^d}], \varepsilon)$ for $\varepsilon \to 0^+$ from Theorem 8.2.1 and Theorem 8.2.2 in the case $\sigma = 1$.

Proof. As a first step we bound t_0 defined in Theorem 8.2.3. To this end, we write

$$g(p_0) := 2^{1+p_0} \cdot \exp\biggl(\frac{1}{W_0(2/p_0)}\biggr) \cdot (2 + 1/e) \geq 4$$

and bound W_0 by log with the help of Lemma 7.2.5, that is

$$\frac{\log(1/p)}{W_0\big(\frac{d+1}{p}\big)} = \frac{\log\Big(\frac{d+1}{p} \cdot e^{\frac{1}{(d+1)e}}\Big)}{W_0\big(\frac{d+1}{p}\big)} \leq 1 + \frac{1}{(d+1)e} \leq d \cdot \frac{2+1/e}{d+1} \quad ,$$

where we used $d+1 \leq 2d$ in the last step. Since $p \leq p_0$ and $d \geq 1$ hold true, we have $4^{\frac{p}{d+1}} \leq 2^{p_0}$ and together we find

$$t_0 = \frac{2(d+1) \cdot 4^{\frac{p}{d+1}}}{ep \cdot W_0\big(\frac{d+1}{p}\big)} \exp\left(\frac{1}{W_0\big(\frac{d+1}{p}\big)}\right) \leq \frac{g(p_0)}{e} \cdot d \cdot \frac{1/p}{\log(1/p)} =: f(p) \quad .$$

Since $\beta(t) = t/\log(t)$ is increasing on $[e, \infty)$ according to Lemma 7.2.7 and $1/p \geq 1/p_0 \geq e$ holds true, we get $f(p) \geq f(p_0) = g(p_0)/e \cdot d \cdot z_0$, where

$$z_0 := \frac{1/p_0}{\log(1/p_0)} \quad .$$

Then Lemma 7.2.2 for $h = t_0$ and $t_0 = f(p_0) = g(p_0)/e \cdot d \cdot z_0$ together with the already observed bounds, $d+1 \leq 2d$ and $4^{\frac{p}{d+1}} \leq 2^{p_0}$, yields

$$\begin{aligned} K_{d,1,p} &= \binom{t_0 + d}{d} \cdot \frac{d+1}{ep} \cdot 4^{\frac{p}{d+1}} \\ &\leq \binom{f(p_0) + d}{d} \cdot \left(\frac{f(p)}{f(p_0)}\right)^d \cdot d \cdot \frac{2^{1+p_0}}{e} \cdot 1/p \\ &= \binom{f(p_0) + d}{d} \cdot z_0^{-d} \cdot d \cdot \frac{2^{1+p_0}}{e} \cdot \frac{(1/p)^{d+1}}{\log^d(1/p)} \quad . \end{aligned}$$

Consequently, we have extracted the dependence on p from the binomial coefficient and it remains to consider the dependence on d. To this end, we use the representation from Lemma 7.2.1 for the binomial coefficient together with Stirling's formula, for $x > 0$,

$$\sqrt{2\pi x} \cdot \left(\frac{x}{e}\right)^x \leq \Gamma(x+1) \leq \exp\left(\frac{1}{12x}\right) \cdot \sqrt{2\pi x} \cdot \left(\frac{x}{e}\right)^x \quad ,$$

$f(p_0) \geq 4$, $d \geq 1$, and $1+t \leq e^t$ for $t = d/f(p_0)$

$$\begin{aligned}
&\binom{f(p_0)+d}{d} \cdot z_0^{-d} \cdot d \\
&\leq \frac{e^{1/60}}{\sqrt{2\pi}} \cdot \left(\frac{1}{d} + \frac{1}{f(p_0)}\right)^{1/2} \left(1 + \frac{f(p_0)}{d}\right)^{d} \left(1 + \frac{d}{f(p_0)}\right)^{f(p_0)} \cdot z_0^{-d} \cdot d \\
&\leq \frac{e^{1/60}}{\sqrt{2\pi}} \cdot \left(1 + \frac{e}{g(p_0) \cdot z_0}\right)^{1/2} \cdot \left(\frac{1 + g(p_0)/e \cdot z_0}{z_0}\right)^{d} \cdot e^d \cdot d^{1/2} \quad .
\end{aligned}$$

Since $C_{p_0} = e/z_0 + g(p_0)$ holds true and the arising quantities that are independent of p and d satisfy

$$\begin{aligned}
&\frac{e^{1/60}}{\sqrt{2\pi}} \cdot \left(1 + \frac{e}{g(p_0) \cdot z_0}\right)^{1/2} \cdot \frac{2^{1+p_0}}{e} \\
\leq\ & \frac{e^{1/60}}{\sqrt{2\pi}} \cdot \left(1 + \frac{1}{2(2+1/e)}\right)^{1/2} \cdot \frac{2^{1+1/e}}{e} \\
\approx\ & 0.4239 \leq 1/2 \quad ,
\end{aligned}$$

the assertion is proven. □

Finally, we note that the constant appearing in Theorem 8.2.3 can again be substantially improved if we restrict our consideration to a smaller range $0 < \varepsilon \leq \varepsilon_0$.

8.3 Translation Invariant Kernels

In this section we exploit the translation invariance of the Gaussian kernel to reduce the problem of bounding the log-covering numbers $\mathcal{H}\big(I_\sigma[X], \varepsilon\big)$ of the ℓ_∞-embedding $I_\sigma[X]$ with a bounded subset $X \subseteq \mathbb{R}^d$ to the estimation of $\mathcal{H}\big(I_\sigma[B_{\ell_2^d}], \varepsilon\big)$. In other words, this section enables us to generalize the log-covering number bounds presented in Section 8.2 to anisotropic Gaussian kernels on general bounded subsets X. Since this reduction is possible for general bounded and translation invariant kernels on a Banach space, we formulate this section in its natural generality.

Let us start with some notation. For a fixed bounded kernel k defined on a set X we often consider its restriction to different subsets $Y \subseteq X$. Consequently, we highlight the considered domain by writing $H(Y)$ for the corresponding RKHS and

$$I[Y]\colon H(Y) \to \ell_\infty(Y),\ f \mapsto f \tag{8.7}$$

for the corresponding ℓ_∞-embedding. Recall that $I[Y]$ is well-defined according to [76, Lemma 4.23]. The first lemma considers the behavior of the covering numbers of $I[X]$ under transformations of the kernel.

8.3.1 Lemma (Transformed Kernels) *Let $T\colon Y \to X$ be a mapping between two non-empty sets and k be a bounded kernel on X with RKHS $H(X)$. Then*

$$k_T(y, y') := k\bigl(T(y), T(y')\bigr) \tag{8.8}$$

for $y, y' \in Y$ defines a bounded kernel on Y with RKHS $H_T(Y) = \bigl\{f \circ T : f \in H(X)\bigr\}$ and the corresponding RKHS-norm satisfies

$$\|f \circ T\|_{H_T(Y)} \leq \|f\|_{H(X)}$$

for $f \in H(X)$. Moreover, the covering numbers satisfy, for $\varepsilon > 0$,

$$\mathcal{N}\bigl(\mathrm{Id}\colon H_T(Y) \to \ell_\infty(Y), \varepsilon\bigr) \leq \mathcal{N}\bigl(\mathrm{Id}\colon H(X) \to \ell_\infty(X), \varepsilon\bigr)\ . \tag{8.9}$$

If, in addition, T is bijective then equality holds in (8.9).

An easy, but important, application of this lemma is the case $Y \subseteq X$ with the embedding $T = \mathrm{id}\colon Y \to X$. In this case we have $k_T = k|_{Y\times Y}$, $H_T(Y) = H(Y)$, and, for $\varepsilon > 0$,

$$\mathcal{N}\bigl(I[Y], \varepsilon\bigr) \leq \mathcal{N}\bigl(I[X], \varepsilon\bigr)\ . \tag{8.10}$$

Proof. Let $\Phi\colon X \to H(X)$ be the canonical feature map of k, that is $\Phi(x) := k(x, \cdot)$ for $x \in X$. Then it is easy to see that $\Phi_T := \Phi \circ T$ is a feature map for k_T. Consequently, k_T is a kernel on Y, and according to [76, Theorem 4.21] the RKHS of k_T has the claimed form, the claimed norm inequality is satisfied, and $S_H\colon H(X)\to H_T(Y)$ defined by $f\mapsto f\circ T$ is a metric surjection, i.e. $S_H\mathring{B}_{H(X)}=\mathring{B}_{H_T(Y)}$. Consequently, it remains to prove the covering number bound. To this end, we define the mapping $S_\infty\colon \ell_\infty(X)\to\ell_\infty(Y)$ by $f\mapsto f\circ T$ and recall that $I[X]$ and $I_T[Y]$ denote the ℓ_∞-embeddings of $H(X)$ and $H_T(Y)$, respectively. These mappings satisfy the commutative diagram

$$\begin{tikzcd}
H(X) \arrow[r, "S_H"] \arrow[d, "{I[X]}"'] & H_T(Y) \arrow[d, "{I_T[Y]}"] \\
\ell_\infty(X) \arrow[r, "S_\infty"'] & \ell_\infty(Y)\ ,
\end{tikzcd}$$

i.e. $I_T[Y]\circ S_H=S_\infty\circ I[X]$. Together with Point (i) of Lemma C.9 and the metric surjectivity of S_H we get, for $\varepsilon>0$,

$$\mathcal{N}\big(I_T[Y],\varepsilon\big)=\mathcal{N}\big(I_T[Y]\circ S_H,\varepsilon\big)=\mathcal{N}\big(S_\infty\circ I[X],\varepsilon\big)\ .$$

Since $\|S_\infty f\|_{\ell_\infty(Y)}=\sup_{y\in Y}|f(T(y))|\le\|f\|_{\ell_\infty(X)}$ is satisfied for all $f\in\ell_\infty(X)$, we have $\|S_\infty\|\le 1$ and together with (C.3) this yields the assertion. If T is bijective, we can exchange the role of X and Y and hence we get the claimed equality. □

The next lemma investigates the behavior of the covering numbers under a partition of the domain.

8.3.2 Lemma (Partition of the Domain) *Let $X=X_1\uplus X_2$ be the disjoint union of non-empty sets X_1,X_2 and k be a bounded kernel on X with RKHS $H(X)$ and ℓ_∞-embedding $I[X]$. Then the covering numbers satisfy, for $\varepsilon>0$,*

$$\mathcal{N}\big(I[X_1\uplus X_2],\varepsilon\big)\le\mathcal{N}\big(I[X_1],\varepsilon\big)\cdot\mathcal{N}\big(I[X_2],\varepsilon\big)\ .$$

Proof. Let $m:=\mathcal{N}\big(I[X_1],\varepsilon\big)$ and $n:=\mathcal{N}\big(I[X_2],\varepsilon\big)$. Moreover, choose corresponding ε-nets $f_1,\dots,f_m\in\ell_\infty(X_1)$ and $g_1,\dots,g_n\in\ell_\infty(X_2)$. Then for each $i \in \{1, \ldots, m\}$ and each $j \in \{1, \ldots, n\}$ we define

$$h_{i,j}(x) := \begin{cases} f_i(x), & x \in X_1 \\ g_j(x), & x \in X_2 \end{cases} ,$$

for $x \in X$. This defines at most $m \cdot n$ different elements of $\ell_\infty(X)$ and it remains to show that $h_{i,j}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$ defines an ε-net of $B_{H(X)}$.

For $h \in H(X)$ with $\|h\|_{H(X)} \leq 1$ Lemma 8.3.1 gives us $h|_{X_\ell} \in H(X_\ell)$ with $\|h|_{X_\ell}\|_{H(X_\ell)} \leq 1$ for $\ell = 1, 2$. Consequently, there is an $i \in \{1, \ldots, m\}$ and a $j \in \{1, \ldots, n\}$ with $\|h|_{X_1} - f_i\|_{\ell_\infty(X_1)} \leq \varepsilon$ and $\|h|_{X_2} - g_j\|_{\ell_\infty(X_2)} \leq \varepsilon$, respectively. For this choice of i and j we have

$$\|h - h_{i,j}\|_{\ell_\infty(X)} = \max\bigl\{\|h|_{X_1} - f_i\|_{\ell_\infty(X_1)}, \|h|_{X_2} - g_j\|_{\ell_\infty(X_2)}\bigr\} \leq \varepsilon$$

and hence the assertion is proven. □

So far, we considered bounded kernels on general sets. In the following we investigate bounded kernels $k : V \times V \to \mathbb{R}$ on a vector space V. The kernel k is called *translation invariant* along a vector $a \in V$ if

$$k(v + a, v' + a) = k(v, v')$$

is satisfied for all $v, v' \in V$. In this case the transformation $T(x) := x + a$ does not change the kernel. With the notation from (8.8) this means $k_T = k$. Since T is bijective as a mapping $X \to a + X$, Lemma 8.3.1 gives the following identity for the covering numbers of the ℓ_∞-embeddings

$$\mathcal{N}\bigl(I[X], \varepsilon\bigr) = \mathcal{N}\bigl(I[X + a], \varepsilon\bigr) \tag{8.11}$$

for all $\varepsilon > 0$. If k is translation invariant along all $a \in U \subseteq V$ for some subspace $U \subseteq V$ then we call k translation invariant along U. The following lemma bounds the covering number of the ℓ_∞-embedding of a translation invariant kernel and is the main result of this section.

8.3.3 Lemma (Translation Invariant Kernels) *Let $(V, \|\cdot\|)$ be a Banach space with complemented subspaces $V_1, V_2 \subseteq V$, i.e. $V = V_1 + V_2$ and $V_1 \cap V_2 = \{0\}$. Moreover, let $X_i \subseteq V_i$ be non-empty subsets, for $i = 1, 2$, and k be a bounded kernel on V with ℓ_∞-embeddings $I[\,\cdot\,]$ of its restrictions defined in (8.7). If k is translation invariant along V_1 and X_1 is precompact then the log-covering numbers satisfy, for $\delta > 0$ and $\varepsilon > 0$,*

$$\mathcal{H}\big(I[X_1 + X_2], \varepsilon\big) \le \mathcal{N}(X_1, \delta) \cdot \mathcal{H}\big(I[\delta B_{V_1} + X_2], \varepsilon\big) \ .$$

Proof. Let us fix some $\varepsilon, \delta > 0$ and set $n := \mathcal{N}(X_1, \delta)$. For a minimal δ-net of X_1 denoted by $x_{1,1}, \ldots, x_{1,n} \in V_1$ we choose a partition $X_{1,1}, \ldots, X_{1,n}$ of X_1 with $X_{1,i} \subseteq x_{1,i} + \delta B_{V_1}$ for all $i = 1, \ldots, n$. Since we chose a minimal δ-net, $X_{1,i} \neq \emptyset$ is non-empty for all $i = 1, \ldots, n$. Because $X_i \subseteq V_i$, for $i = 1, 2$, and V_1, V_2 are complemented subspaces the sets $X_{1,i} + X_2$, for $i = 1, \ldots, n$, form a partition of $X_1 + X_2$ with $X_{1,i} + X_2 \subseteq x_{1,i} + \delta B_{V_1} + X_2$. A repeated application of Lemma 8.3.2 and (8.10) yield

$$\mathcal{H}\big(I[X], \varepsilon\big) \le \sum_{i=1}^{n} \mathcal{H}\Big(I\big[X_{1,i} + X_2\big], \varepsilon\Big) \le \sum_{i=1}^{n} \mathcal{H}\Big(I\big[x_{1,i} + (\delta B_{V_1}) + X_2\big], \varepsilon\Big) \ .$$

Since k is translation invariant along V_1, (8.11) gives the assertion. □

8.4 Anisotropic Gaussian Kernels on General Domains

The goal of this section is to analyze how the constants in the log-covering number bounds depend on the kernel width σ and the size of the input space X. To this end, we use Lemma 8.3.3 to reduces the problem of bounding the log-covering numbers of an *anisotropic* Gaussian RKHSs to the estimation of the log-covering numbers of the *isotropic* Gaussian RKHS on the closed unit ball $B_{\ell_2^d}$ with width $\sigma = 1$.

However, let us start with an easy observation: With the notation introduced in (8.8) the anisotropic Gaussian kernel reads $k_{\boldsymbol{\sigma}} = k_{D_{\boldsymbol{\sigma}}}$ if k and

$D_{\boldsymbol{\sigma}}$ denote the isotropic Gaussian kernel of width $\sigma = 1$ and the diagonal operator, respectively. Since $D_{\boldsymbol{\sigma}}\colon X \to D_{\boldsymbol{\sigma}}X$ is bijective, Lemma 8.3.1 yields, for $\varepsilon > 0$,

$$\mathcal{H}\big(I_{\boldsymbol{\sigma}}[X], \varepsilon\big) = \mathcal{H}\big(I_1[D_{\boldsymbol{\sigma}}X], \varepsilon\big) \ . \tag{8.12}$$

Note that we did not use that X is a subset of a finite-dimensional space and hence (8.12) holds true for Gaussian kernels on infinite-dimensional spaces.

The following theorem is a direct consequence of (8.12) and Lemma 8.3.3.

8.4.1 Theorem *For a bounded subset $X \subseteq \mathbb{R}^d$, $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in (0,\infty)^d$, and $0 < \varepsilon \leq 1$ the following log-covering number bound is satisfied*

$$\mathcal{H}\big(I_{\boldsymbol{\sigma}}[X], \varepsilon\big) \leq \mathcal{N}(D_{\boldsymbol{\sigma}}X, 1) \cdot \mathcal{H}\big(I_1[B_{\ell_2^d}], \varepsilon\big) \ ,$$

where the covering numbers $\mathcal{N}(D_{\boldsymbol{\sigma}}X, 1)$ of $D_{\boldsymbol{\sigma}}X \subseteq \mathbb{R}^d$ are with respect to the Euclidean norm.

Proof. Using (8.12) and Lemma 8.3.3 for $\delta = 1$, $V_1 = \mathbb{R}^d$ (equipped with the Euclidean norm), $V_2 = \{0\}$, and $X_1 = D_{\boldsymbol{\sigma}}X$, $X_2 = \{0\}$ we get the assertion. □

The proof suggests that this bound can be improved if we optimize over δ instead of using a fixed value for δ. But this is not an easy task and the possible suboptimal choice $\delta = 1$ has the advantage that the dependence of the bound on $\boldsymbol{\sigma}$ and ε is in some sense factorized.

Now, to illustrate the impact of Theorem 8.4.1 we note that X is assumed to be bounded, and hence there is an $a \in \mathbb{R}^d$ and an $R > 0$ with $X \subseteq a + RB_{\ell_2^d}$. In the case $\min_i \sigma_i \geq 1/R$ Point (iii) of Lemma C.7 gives us

$$\begin{aligned}\mathcal{N}\big(D_{\boldsymbol{\sigma}}X, 1\big) &\leq \mathcal{N}\big(D_{\boldsymbol{\sigma}}B_{\ell_2^d}, 1/R\big)\\ &\leq \frac{\lambda^d\big(D_{\boldsymbol{\sigma}}B_{\ell_2^d}\big)}{\lambda^d\big(B_{\ell_2^d}\big)} \cdot \left(\frac{1}{\min_i \sigma_i} + 2R\right)^d\\ &\leq \sigma_1 \cdot \ldots \cdot \sigma_d \cdot (3R)^d \ .\end{aligned} \tag{8.13}$$

For the sake of completeness, we further mention that in the case $\max_i \sigma_i \leq 1/R$ we have $\mathcal{N}(D_{\boldsymbol{\sigma}} X, 1) = 1$. Now, we can combine Theorem 8.4.1 and (8.13) with one of the theorems presented in Section 8.2.

For example, together with Theorem 8.2.1 we obtain

$$\begin{aligned}\mathcal{H}\big(I_{\boldsymbol{\sigma}}[X], \varepsilon\big) &\leq \sigma_1 \cdot \ldots \cdot \sigma_d \cdot (3R)^d \cdot \mathcal{H}\big(I_1[B_{\ell_2^d}], \varepsilon\big) \\ &\leq \tilde{K}_{d,\boldsymbol{\sigma},R} \cdot \frac{\log^{d+1}(4/\varepsilon)}{\log\log^d(4/\varepsilon)}\end{aligned} \tag{8.14}$$

for $0 < \varepsilon \leq 1$, $R > 0$, and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d) \in \big[1/R, \infty\big)^d$ with

$$\begin{aligned}\tilde{K}_{d,\boldsymbol{\sigma},R} &\coloneqq K_{d,1} \cdot \sigma_1 \cdot \ldots \cdot \sigma_d \cdot (3R)^d \\ &= \binom{4e+d}{d} \cdot (3R/e)^d \cdot \sigma_1 \cdot \ldots \cdot \sigma_d \\ &\leq C_{1,1} \cdot d^{4e} \cdot (3R/e)^d \cdot \sigma_1 \cdot \ldots \cdot \sigma_d \ .\end{aligned}$$

Recall that $C_{1,1}$ is defined in (8.2) with $C_{1,1} = 4e + 1 \approx 11.8731$. We mention that in the case $R \geq 1$ and an isotropic Gaussian kernel with width $\sigma \geq 1$ the constant $\tilde{K}_{d,\boldsymbol{\sigma},R}$ grows like σ^d for $\sigma \to \infty$. In contrast, recall from (8.3) that the constant $K_{d,\sigma}$ obtained in Theorem 8.2.1 grows like σ^{2d}. Consequently, (8.14) improves Theorem 8.2.1 in the dependence on σ by a factor of 2 in the exponent. In this respect, note that [84, Lemma 4.5] obtained the same behavior in σ but for a bound that does *not* include the double logarithmic factor $\log\log^d(4/\varepsilon)$ of (8.14). Moreover, [76, Theorem 6.27] achieves the same behavior in σ for a polynomial bound. Of course, the latter two results can be recovered from (8.14), and in addition, the results in [76, 84] do not take care of the explicit form of the constants and their dependence on d.

Next, a combination of Theorem 8.4.1 with Theorem 8.2.3 yields

$$\mathcal{H}\big(I_{\boldsymbol{\sigma}}[X], \varepsilon\big) \leq K_{d,1,p} \cdot \mathcal{N}(D_{\boldsymbol{\sigma}} X, 1) \cdot \varepsilon^{-p} \tag{8.15}$$

for $d \geq 1$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d) \in (0, \infty)^d$, $p > 0$, and $0 < \varepsilon \leq 1$ with the constant $K_{d,1,p}$ defined in Theorem 8.2.3.

Finally, we compare our result with [87, Proposition 1] since this is one of the few results that provides explicit constants for the log-covering number bound. As preparation, note that [87] considers the Gaussian kernel on the set $X = [0,1]^d$. To apply our result, we use $[0,1]^d \subseteq 1/2 + \sqrt{d}/2 \cdot B_{\ell_2^d}$. Then (8.14) gives us

$$\mathcal{H}\big(I_{\boldsymbol{\sigma}}\big[[0,1]^d\big], \varepsilon\big) \le C_{1,1} \cdot d^{4e} \cdot (2e/3)^{-d} \cdot d^{d/2} \cdot \sigma^d \cdot \frac{\log^{d+1}(4/\varepsilon)}{\log\log^d(4/\varepsilon)}$$

for $0 < \varepsilon \le 1$, $\sigma > 0$, and $d \ge 1$ with $\sigma \ge 2/\sqrt{d}$. As a result, we approximately get an $d^{d/2}$-behavior of the constant for $d \to \infty$. Since the bound in [87] behaves like d^{d+1} for $d \to \infty$, our result improves this by roughly a factor of 2 in the exponent.

The result in [87] establishes an improved bound for smaller ranges $0 < \varepsilon \le \varepsilon_0$ as well. To be more precise, they proved, for $\sigma > 0$ and $d \ge 1$,

$$\mathcal{H}\big(I_{\boldsymbol{\sigma}}\big[[0,1]^d\big], \varepsilon\big) \le 4^d(6d+2) \cdot \log^{d+1}(1/\varepsilon)$$

for all $0 < \varepsilon \le \exp(-90d^2\sigma^2 - 11d - 3)$. In order to compare our results with [87] we need the following lemma.

8.4.2 Lemma *For $C > 0$, $d \ge 2Ce^2$, and $\varepsilon_0 := 4\exp\big(-\frac{d}{2C}\log\big(\frac{d}{2eC}\big)\big)$ the condition $\varepsilon_0 \le 4\exp\big(-e^{1+\sigma^{-2}}\big)$ in Theorem 8.2.2 is satisfied for $\sigma = 1$ and the quantity $K_{d,1,\varepsilon_0}$ defined in Theorem 8.2.2 satisfies*

$$K_{d,1,\varepsilon_0} \le (2\pi)^{-1/2} \cdot (4e)^d (1+C)^d \cdot d^{-(d+1/2)} \quad .$$

Moreover, for $\frac{1}{2e^2} \ge C \ge \frac{1}{\sqrt{360e}}$ we have

$$\exp\big(-90d^2 - 11d - 3\big) \le \varepsilon_0 \quad .$$

Combining Theorem 8.4.1 with Theorem 8.2.2 and this lemma for $C = 1/\sqrt{360e} \approx 0.0320$ we obtain, for $d \ge 1$ and $\sigma \ge 2/\sqrt{d}$,

$$\mathcal{H}\big(I_{\boldsymbol{\sigma}}\big[[0,1]^d\big], \varepsilon\big) \le (2\pi)^{-1/2} \cdot (6e(1+C))^d \cdot d^{-(d+1)/2} \cdot \sigma^d \cdot \frac{\log^{d+1}(4/\varepsilon)}{\log\log^d(4/\varepsilon)}$$

for all $0 < \varepsilon \leq 4\exp\bigl(-\frac{d}{2C}\log\bigl(\frac{d}{2eC}\bigr)\bigr)$ where $6e(1+C) \approx 16.8311$. This improves the result from [87] in both, the ε range and the behavior for $d \to \infty$

Proof. Let us recall the definitions of $y_0 := \log(4/\varepsilon_0)$ and $x_0 := \frac{2y_0}{W_0(y_0/e)} = h_1(y_0)$ from Theorem 8.2.2 as well as the definitions of h_1 and p_1 from Lemma 7.2.6. Using the function p_1 we can write

$$\varepsilon_0 = 4\Bigl(\frac{2eC}{d}\Bigr)^{\frac{d}{2C}} = 2 \cdot p_1(d/C) \ .$$

Since $d/C \geq 2e^2 \geq 2$ holds true, Point (iii) of Lemma 7.2.6, which states $p_1^{-1} = h_1 \circ \log(2/\cdot)$, is applicable and hence

$$x_0 = h_1 \circ \log(4/\varepsilon_0) = h_1 \circ \log\biggl(\frac{2}{p_1(d/C)}\biggr) = d/C$$

is satisfied. Next, we prove the inequality $\varepsilon_0 \leq u := 4\exp(-e^2)$. To this end, note that we have $h_1 \circ \log(4/u) = h_1(e^2) = 2e\exp\bigl(W_0(e)\bigr) = 2e^2$ since $W_0(e) = 1$. Our assumption $d \geq 2Ce^2$ implies

$$h_1 \circ \log(4/\varepsilon_0) = d/C \geq 2e^2 = h_1 \circ \log(4/u)$$

and since h_1 is increasing according to Point (ii) of Lemma 7.2.6 we get $\varepsilon_0 \leq u = 4\exp(-e^2)$. Now, we prove the bound on $K_{d,1,\varepsilon_0}$. To this end, we rewrite $K_{d,1,\varepsilon_0}$ using the representation of the binomial coefficient from (7.2)

$$\begin{aligned} K_{d,1,\varepsilon_0} &= \binom{x_0+d}{d} x_0^{-d} \cdot \biggl(\frac{x_0\log(y_0)}{y_0}\biggr)^d \\ &= \frac{1}{d!}\prod_{i=1}^{d}(1+i/x_0) \cdot \biggl(\frac{2\log(y_0)}{W_0(y_0/e)}\biggr)^d \ . \end{aligned}$$

If we bound the first factor by using $i/x_0 \leq d/x_0 = C$ and if we bound the second factor by using $\log(y_0) \leq 2W_0(y_0/e)$ from Lemma 7.2.5 then we get

$$K_{d,1,\varepsilon_0} \leq \frac{4^d(1+C)^d}{d!} \quad .$$

Together with Stirling's formula $d! \geq \sqrt{2\pi d}\,(d/e)^d$ this gives the desired bound. Finally, note that $\log(t) \leq t$ and $C \geq 1/\sqrt{360e}$ yields

$$\varepsilon_0 \geq \exp\Bigl(-\frac{d^2}{4eC^2}\Bigr) \geq \exp(-90d^2) \quad ,$$

which proves the lower bound on ε_0. □

8.5 Conversion to Dyadic Entropy Number Bounds

Since in some situations it is more convenient to work with dyadic entropy number bounds instead of log-covering number bounds we show in this section how to convert some bounds of the previous sections into dyadic entropy number bounds. Recall, that the *dyadic entropy number* $e_n(I_{\boldsymbol{\sigma}}[X])$ is defined as the infimum over all $r > 0$ that allows to cover the closed unit ball $B_{H_\sigma(X)}$ with 2^{n-1} translates of $rB_{\ell_\infty(X)}$.

This conversion is easiest for polynomial bounds. To this end, we start with the polynomial bound in (8.15).

8.5.1 Theorem (Polynomial Entropy Number Bound) *For a bounded subset $X \subseteq \mathbb{R}^d$, $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in (0,\infty)^d$, and $p > 0$ the following dyadic entropy number bound is satisfied, for $n \geq 1$,*

$$e_n\bigl(I_{\boldsymbol{\sigma}}[X]\bigr) \leq \bigl(3 \cdot K_{d,1,p} \cdot \mathcal{N}(D_{\boldsymbol{\sigma}}X, 1)\bigr)^{1/p} \cdot n^{-1/p} \quad ,$$

where $K_{d,1,p}$ denotes the constant from Theorem 8.2.3 and the covering numbers $\mathcal{N}(D_{\boldsymbol{\sigma}}X, 1)$ are with respect to the Euclidean norm.

Proof. According to (8.15) we have $\mathcal{H}\bigl(I_{\boldsymbol{\sigma}}[X], \varepsilon\bigr) \leq C \cdot \varepsilon^{-p} =: F^{-1}(\varepsilon)$ for all $0 < \varepsilon \leq 1$, where we used the abbreviation $C := K_{d,1,p} \cdot \mathcal{N}(D_{\boldsymbol{\sigma}}X, 1)$ and defined the function $F^{-1}\colon (0,1] \to [C,\infty)$. Since F^{-1} is decreasing and bijective with $F(t) = (C/t)^{1/p}$, Point (ii) of Lemma C.8 gives us

$$e_n\big(I_{\boldsymbol{\sigma}}[X]\big) \leq F\big(\log(2)\cdot(n-1)\big) = \big(C/\log(2)\big)^{1/p}\cdot(n-1)^{-1/p}$$

for all $n \geq C/\log(2)+1$. This bound remains true for $2 \leq n < C/\log(2)+1$ since in this case the right hand side is larger than 1 and we have the trivial bound $e_n(I_{\boldsymbol{\sigma}}[X]) \leq e_1(I_{\boldsymbol{\sigma}}[X]) = \|I_{\boldsymbol{\sigma}}[X]\| = 1$. Using $(n-1)^{-1/p} \leq 2^{1/p}\cdot n^{-1/p}$ for $n \geq 2$ and $2/\log(2) \approx 2.8854 \leq 3$ we get the assertion for $n \geq 2$.

To treat the case $n = 1$ we use the following observation. According to Point (iii) of Lemma C.9 we have $\log(2) \leq \mathcal{H}(I_{\boldsymbol{\sigma}}[X],\varepsilon) \leq C\cdot\varepsilon^{-p}$ for $\varepsilon < \|I_{\boldsymbol{\sigma}}[X]\| = 1$. Letting $\varepsilon \nearrow 1$ we find $C \geq \log(2)$ and hence $e_1(I_{\boldsymbol{\sigma}}[X]) = 1 \leq (C/\log(2))^{1/p}$ proves the desired bound for $n = 1$. □

In order to provide a dyadic entropy number bound decreasing faster than any polynomial we need the following auxiliary lemma.

8.5.2 Lemma *For $d \geq 1$ the function $f\colon [0,\infty) \to [0,\infty)$ given by*

$$f(t) \coloneqq t\cdot\exp\big(d\cdot W_0(t)\big)$$

is increasing, bijective, satisfies

$$f(t) = \frac{t^{d+1}}{W_0^d(t)}$$

for $t > 0$, and its inverse is given by

$$\begin{aligned} f^{-1}(t) &= \frac{W_0\big((d+1)t\big)}{d+1}\cdot\exp\left(\frac{W_0\big((d+1)t\big)}{d+1}\right)\\ &= t^{\frac{1}{d+1}}\cdot\left(\frac{W_0\big((d+1)t\big)}{d+1}\right)^{\frac{d}{d+1}} \quad . \end{aligned}$$

Proof. Since Lambert's W-function W_0 is increasing and bijective from $[0,\infty)$ to $[0,\infty)$, f is increasing and bijective, too. The second representations of f and f^{-1} are direct consequences of the identity $\exp(W_0(x)) =$

$x/W_0(x)$ for $x > 0$. Consequently, it remains to check that the given f^{-1} is actually the inverse of f. To this end, we calculate $f \circ f^{-1}(t)$ for $t > 0$. Using the second representation for f and plugging the first representation of f^{-1} into the denominator and the second one into the enumerator we directly get $f \circ f^{-1}(t) = t$, i.e. the assertion is proven for $t > 0$. For $t = 0$ the identity $f \circ f^{-1}(t) = t$ is obvious. □

The following theorem provides a dyadic entropy number bound which decreases faster than every polynomial.

8.5.3 Theorem (Entropy Number Bound on $B_{\ell_2^d}$) *For $d \geq 1$, $\sigma > 0$, and $0 < \varepsilon_0 \leq 1$ consider $y_0 := \log(4/\varepsilon_0)$, $x_0 := 2y_0/W_0\big(\frac{y_0}{e\sigma^2}\big)$,*

$$C_{d,\sigma,\varepsilon_0} := \binom{x_0 + d}{d} \cdot (2/x_0)^d \cdot \frac{1}{\log(2)} \ , \qquad \textit{and}$$

$$n_0 := \binom{x_0 + d}{d} \cdot y_0/\log(2) + 1 \ .$$

Then the following dyadic entropy number bound is satisfied, for $n \geq n_0$,

$$e_n\big(I_\sigma[B_{\ell_2^d}]\big) \leq 4\exp\left(-\left(\frac{n-1}{C_{d,\sigma,\varepsilon_0}(d+1)^d}\right)^{\frac{1}{d+1}} W_0^{\frac{d}{d+1}}\left(\frac{(n-1)(d+1)}{C_{d,\sigma,\varepsilon_0}(e\sigma^2)^{d+1}}\right)\right) \ .$$

For $\varepsilon_0 \searrow 0$ we have $y_0 \nearrow \infty$ and $x_0 \nearrow \infty$. Together with Lemma 7.2.1 we have $C_{d,\sigma,\varepsilon_0} \searrow 2^d/d! \cdot 1/\log(2)$ for $\varepsilon_0 \searrow 0$. This means that the constant $C_{d,\sigma,\varepsilon_0}$ and hence the bound improves for decreasing ε_0 but also the range in which this bound applies decreases, i.e. $n_0 \nearrow \infty$.

For convenience, the consideration of Theorem 8.5.3 is restricted to the isotropic Gaussian RKHS on $B_{\ell_2^d}$. However, this bound can be easily generalized to the anisotropic Gaussian RKHS on general bounded sets $X \subseteq \mathbb{R}^d$.

Moreover, for $\varepsilon_0 = 1$ this bound remains valid for the whole range $n \geq 1$. This can be proven by proceeding analogously to the proof of Theorem 8.5.1. To be more precise, for $n \leq n_0 = F^{-1}(\varepsilon_0)/\log(2) + 1$, the right hand side is larger than $\varepsilon_0 = 1$ and hence the trivial bound $e_n(I_{\boldsymbol{\sigma}}[X]) \leq e_1(I_{\boldsymbol{\sigma}}[X]) = 1$ gives the desired estimate for $n \geq n_0$.

Proof. The proof is an application of Lemma C.8 and hence we need a bound on the log-covering numbers which is decreasing and bijective as a function on ε. Unfortunately, the function $\varepsilon \mapsto \frac{\log^{d+1}(4/\varepsilon)}{\log\log^{d}(4/\varepsilon)}$ is neither bijective nor decreasing on $(0,1]$. To this end, we first slightly modify the log-covering number bound from Theorem 8.2.1 using Lemma 8.1.2 such that the resulting bound satisfies our needs.

For $0 < \varepsilon \le \varepsilon_0$ we write $y := \log(4/\varepsilon) \ge y_0 > 1$ and recall the definition $h_\sigma(y) = 2y/W_0\big(\frac{y}{e\sigma^2}\big)$ from Lemma 7.2.6. Since h_σ is increasing, we have $h_\sigma(y) \ge h_\sigma(y_0) = x_0$. Combining Lemma 8.1.2 and Lemma 7.2.2 for $h = f = (h_\sigma \circ \log)(4/\,\cdot\,)$ and $t_0 = x_0$ gives us

$$\mathcal{H}\big(I_\sigma[B_{\ell_2^d}],\varepsilon\big) \le \binom{h_\sigma(y)+d}{d} \cdot y \le \binom{x_0+d}{d} \cdot (2/x_0)^d \cdot \frac{y^{d+1}}{W_0^d\big(\frac{y}{e\sigma^2}\big)} \ .$$

Using the auxiliary function $f\colon [0,\infty) \to [0,\infty)$ from Lemma 8.5.2 this bound reads

$$\mathcal{H}\big(I_\sigma[B_{\ell_2^d}],\varepsilon\big) \le \log(2) \cdot C_{d,\sigma,\varepsilon_0} \cdot (e\sigma^2)^{d+1} \cdot f\Big(\frac{y}{e\sigma^2}\Big) =: F^{-1}(\varepsilon)$$

for all $0 < \varepsilon \le \varepsilon_0$. Since the function $F^{-1}\colon (0,\varepsilon_0] \to [F^{-1}(\varepsilon_0),\infty)$ satisfies

$$F(t) = 4 \cdot \exp\bigg(-e\sigma^2 \cdot f^{-1}\Big(\frac{t}{\log(2)\cdot C_{d,\sigma,\varepsilon_0} \cdot (e\sigma^2)^{d+1}}\Big)\bigg) \ ,$$

a combination of Lemma 8.5.2 and Point (ii) of Lemma C.8 gives the desired dyadic entropy number bound for $n \ge F^{-1}(\varepsilon_0)/\log(2) + 1 = n_0$. □

Chapter 9

Gaussian Kernels on Infinite-Dimensional Spaces

In this chapter we transfer some of the log-covering number bounds for Gaussian kernels on finite-dimensional domains presented in Chapter 8 to Gaussian kernels on infinite-dimensional domains $X \subseteq \ell_2$. To be more precise, (8.15) states that on a bounded (precompact) set $X \subseteq \mathbb{R}^d$ the log-covering numbers $\mathcal{H}(I_\sigma[X], \varepsilon)$ increase slower than any polynomial for $\varepsilon \to 0^+$. The goal of this chapter is the generalization of this statement to some specific precompact but infinite-dimensional sets $X \subseteq \ell_2$.

9.1 Product Kernels

In this section we investigate product kernels. For such product kernels we partly reduce the problem of bounding the log-covering numbers of the ℓ_∞-embedding of their RKHS to the estimation of the same quantity for one factor only. In Section 9.2 below we use these results for Gaussian kernels.

Let us start with some notation: X_1 and X_2 are non-empty sets and $k_1\colon X_1 \times X_2 \to \mathbb{R}$ and $k_2\colon X_2 \times X_2 \to \mathbb{R}$ are kernels on X_1 and X_2 with RKHSs H_1 and H_2, respectively. According to [76, Lemma 4.6] the product kernel $k\colon X \times X \to \mathbb{R}$ on $X \coloneqq X_1 \times X_2$ given by

$$k\big((x_1, x_2), (x_1', x_2')\big) \coloneqq k_1(x_1, x_1') \cdot k_2(x_2, x_2') \tag{9.1}$$

for $x_1, x_1' \in X_1$ and $x_2, x_2' \in X_2$ is actually a kernel whose RKHS we denote by H. Then we define the *tensor product* of the RKHSs H_1 and H_2 by

$$H_1 \otimes H_2 := H \quad . \tag{9.2}$$

Moreover, for bounded kernels k_1 and k_2 we denote the corresponding ℓ_∞-embeddings by $I[X_1]\colon H_1 \to \ell_\infty(X_1)$ and $I[X_2]\colon H_2 \to \ell_\infty(X_2)$, respectively. In this section we investigate the log-covering numbers of the ℓ_∞-embedding $I[X_1 \times X_2]\colon H_1 \otimes H_2 \to \ell_\infty(X_1 \times X_2)$ of the product kernel k. But, we start with some preparatory results. The first lemma provides an ONB for the RKHS $H_1 \otimes H_2$.

9.1.1 Lemma (ONB of $H_1 \otimes H_2$) *Let X_1, X_2 be non-empty sets and k_1, k_2 be kernels on X_1 and X_2 with RKHSs H_1 and H_2, respectively. Then the following statements are true:*

(i) For ONBs $(e_i)_{i\in I}$ and $(f_j)_{j\in J}$ of H_1 and H_2, respectively, the functions $(e_i \otimes f_j)_{i\in I, j\in J}$ defined by

$$(e_i \otimes f_j)(x_1, x_2) := e_i(x_1) f_j(x_2)$$

for $x_1 \in X_1$, $x_2 \in X_2$ form an ONB of $H_1 \otimes H_2$.

(ii) $\|h_1 \otimes h_2\|_{H_1\otimes H_2} = \|h_1\|_{H_1}\|h_2\|_{H_2}$ for all $h_1 \in H_1$ and $h_2 \in H_2$.

For closed subspaces $U_1 \subseteq H_1$ and $U_2 \subseteq H_2$, which are again RKHSs, Lemma 9.1.1 directly gives $U_1 \otimes U_2 \subseteq H_1 \otimes H_2$.

Proof. (i) Let $\Phi_1\colon X_1 \to H_1$ and $\Phi_2\colon X_2 \to H_2$ be the canonical feature maps of k_1 and k_2, respectively, i.e. $\Phi_i(x_i) := k_i(x_i, \cdot\,)$ for $x_i \in X_i$ and $i = 1, 2$. Recall that $H_1 \hat{\otimes}_{\mathrm{hs}} H_2$ denotes the Hilbert space of Hilbert-Schmidt operators from H_1 to H_2. For a definition and basic properties of Hilbert-Schmidt operators see e.g. [63, Definition on p. 152]. From the proof of [76, Lemma 4.6] we know, that $\Phi\colon X_1 \times X_2 \to H_1 \hat{\otimes}_{\mathrm{hs}} H_2$ defined by

$$\Phi(x_1, x_2) := \Phi_1(x_1) \,\hat{\otimes}_{\mathrm{hs}}\, \Phi_2(x_2) := \big\langle \Phi_1(x_1), \cdot\, \big\rangle_{H_1} \Phi_2(x_2) \quad ,$$

for $(x_1, x_2) \in X_1 \times X_2$, is a feature map of the product kernel k. According to [76, Theorem 4.21] the mapping $V \colon H_1 \hat{\otimes}_{\text{hs}} H_2 \to H$ defined by

$$(Vw)(x_1, x_2) = \left\langle \Phi_1(x_1) \hat{\otimes}_{\text{hs}} \Phi_2(x_2), w \right\rangle_{H_1 \hat{\otimes}_{\text{hs}} H_2}$$

is a metric surjection. We show that V is injective and hence an isometric isomorphism. To this end, let $w \in H_1 \hat{\otimes}_{\text{hs}} H_2$ be fixed with $Vw = 0$. It is well-known that $(e_i \hat{\otimes}_{\text{hs}} f_j)_{i\in I, j\in J}$ forms an ONB of $H_1 \hat{\otimes}_{\text{hs}} H_2$ and consequently there is an (unique) $a = (a_{i,j})_{i\in I, j\in J} \in \ell_2(I \times J)$ with $w = \sum_{i\in I, j\in J} a_{i,j}(e_i \hat{\otimes}_{\text{hs}} f_j)$. Plugging this representation of w into the definition of V yields

$$\begin{aligned} 0 = (Vw)(x_1, x_2) &= \sum_{i\in I, j\in J} a_{i,j} \left\langle \Phi_1(x_1) \hat{\otimes}_{\text{hs}} \Phi_2(x_2), e_i \hat{\otimes}_{\text{hs}} f_j \right\rangle_{H_1 \hat{\otimes}_{\text{hs}} H_2} \\ &= \sum_{i\in I, j\in J} a_{i,j} \left\langle \Phi_1(x_1), e_i \right\rangle_{H_1} \left\langle \Phi_2(x_2), f_j \right\rangle_{H_2} \qquad (9.3) \\ &= \sum_{i\in I} e_i(x_1) \sum_{j\in J} a_{i,j} f_j(x_2) \end{aligned}$$

for all $x_1 \in X_1$,$x_2 \in X_2$. Since $(a_{i,j})_{j\in J} \in \ell_2(J)$ holds true for every fixed $i \in I$, we can define the sequence $b(x_2) = (b_i(x_2))_{i\in I}$ by

$$b_i(x_2) \coloneqq \sum_{j\in J} a_{i,j} f_j(x_2)$$

for every fixed $x_2 \in X_2$. Using Minkowski's inequality (for integrals), Hölder's inequality, and the representation from [76, Theorem 4.20] we find

$$\begin{aligned} \left(\sum_{i\in I} b_i^2(x_2) \right)^{1/2} &\leq \sum_{j\in J} |f_j(x_2)| \left(\sum_{i\in I} a_{i,j}^2 \right)^{1/2} \\ &\leq \left(\sum_{j\in J} f_j^2(x_2) \right)^{1/2} \|a\|_{\ell_2(I\times J)} \qquad (9.4) \\ &= k_2^{1/2}(x_2, x_2) \cdot \|a\|_{\ell_2(I\times J)} < \infty \end{aligned}$$

and hence $b(x_2) \in \ell_2(I)$ for all $x_2 \in X_2$. Consequently, for all $x_2 \in X_2$ the series $\sum_{i\in I} b_i(x_2)e_i$ defines an element in H_1, which is identical zero according to (9.3). Since $(e_i)_{i\in I}$ is an ONB of H_1, we get $b_i(x_2) = 0$ for all $x_2 \in X_2$ and all $i \in I$. Thus, for a fixed $i \in I$, $b_i = \sum_{j\in J} a_{i,j} f_j$ defines an element of H_2 which is zero. Using the ONB property of $(f_j)_{j\in J}$ in H_2 we get $a_{i,j} = 0$ for all $j \in J$. Because $i \in I$ was arbitrary we get $w = 0$ and hence V is an isometric isomorphism.

Consequently, the image of the ONB $(e_i \hat{\otimes}_{\mathrm{hs}} f_j)_{i\in I, j\in J}$ of $H_1 \hat{\otimes}_{\mathrm{hs}} H_2$ under V is an ONB of H. Since $V(e_i \hat{\otimes}_{\mathrm{hs}} f_j) = e_i \otimes f_j$ for all $i \in I$ and $j \in J$, this gives the first assertion.

(ii) The second assertion is a consequence of an application of the isometric isomorphism V to the well-known equality $\|h_1 \hat{\otimes}_{\mathrm{hs}} h_2\|_{H_1 \hat{\otimes}_{\mathrm{hs}} H_2} = \|h_1\|_{H_1}\|h_2\|_{H_2}$. □

Note that with an analogous calculation as in (9.4) we can show that for $h \in H_1 \otimes H_2$ and $x_2 \in X_2$ the section $h_{x_2}(x_1) := h(x_1, x_2)$ defines an element of H_1 with $\|h_{x_2}\|_{H_2} \le \|h\|_H \cdot k_2^{1/2}(x_2, x_2)$.

The next lemma provides an identity for the operator norm of the ℓ_∞-embedding restricted to subspaces, cf. [76, Theorem 4.20 and Lemma 4.23].

9.1.2 Lemma (Operator Norm) *Let $k\colon X \times X \to \mathbb{R}$ be a bounded kernel on a non-empty set X with RKHS H, $(e_k)_{k\in K}$ be an ONS in H, and $I\colon H \to \ell_\infty(X)$ be the ℓ_∞-embedding of H. Then, for the orthogonal projection $P\colon H \to H$ onto $\overline{\operatorname{span}}\{e_k : k \in K\}$, the following identity holds true*

$$\|I \circ P\|^2 = \sup_{x\in X} \sum_{k\in K} e_k^2(x) \quad .$$

Proof. Obviously, $U := \overline{\operatorname{span}}\{e_k : k \in K\}$ is a RKHS and the corresponding kernel $k_U\colon X \times X \to \mathbb{R}$ is given by

$$k_U(x, x') = \sum_{k\in K} e_i(x)e_i(x') \quad ,$$

see [76, Theorem 4.20]. Since $PB_H = B_U$ holds true, we have $\|I \circ P\| = \|\operatorname{Id} : U \to \ell_\infty(X)\|$. Moreover, $\|\operatorname{Id} : U \to \ell_\infty(X)\|$ is given by [76, Lemma 4.23] and hence

$$\|I \circ P\|^2 = \sup_{x \in X} k_U(x,x) = \sup_{x \in X} \sum_{k \in K} e_k^2(x)$$

proves the assertion. □

The following lemma combines Lemma 9.1.2 with Lemma 9.1.1 to get an identity for the operator norm of a product kernel's ℓ_∞-embedding.

9.1.3 Lemma (Operator Norm for Product Kernels) *Let X_1, X_2 be non-empty sets and k_1, k_2 be bounded kernels on X_1 and X_2 with RKHSs H_1 and H_2 as well as ℓ_∞-embeddings $I[X_1]$ and $I[X_2]$, respectively. Furthermore, let k be the product kernel of k_1 and k_2 on $X_1 \times X_2$ defined in* (9.1) *with ℓ_∞-embedding $I[X_1 \times X_2]$. If $U_1 \subseteq H_1$, $U_2 \subseteq H_2$ are closed subspaces, $P(U_1), P(U_2)$ the orthogonal projections onto U_1 and U_2, respectively, and $P(U_1 \otimes U_2)$ the orthogonal projection onto $U_1 \otimes U_2 \subseteq H_1 \otimes H_2$ then the following equation is satisfied*

$$\big\|I[X_1 \times X_2] \circ P(U_1 \otimes U_2)\big\| = \big\|I[X_1] \circ P(U_1)\big\| \cdot \big\|I[X_2] \circ P(U_2)\big\| \ .$$

Note that for $U_1 = H_1$ and $U_2 = H_2$ Lemma 9.1.3 gives $\|I[X_1 \times X_2]\| = \|I[X_1]\| \cdot \|I[X_2]\|$.

Proof. If we choose ONBs $(e_i)_{i \in I}$ and $(f_j)_{j \in J}$ of U_1 and U_2, respectively, then $(e_i \otimes f_j)_{i \in I, j \in J}$ forms an ONB of $U_1 \otimes U_2$ according to Lemma 9.1.1. Using Lemma 9.1.2 we find

$$\begin{aligned}\big\|I[X_1 \times X_2] \circ P(U_1 \otimes U_2)\big\| &= \sup_{(x_1,x_2) \in X_1 \times X_2} \sum_{i \in I, j \in J} (e_i \otimes f_j)^2(x_1,x_2) \\ &= \sup_{x_1 \in X_1} \sum_{i \in I} e_i^2(x_1) \sup_{x_2 \in X_2} \sum_{j \in J} f_j^2(x_2) \ .\end{aligned}$$

A two-fold application of Lemma 9.1.2 yields the assertion. □

The final lemma provides the already mentioned covering number bound for product kernels and is the main result of this section.

9.1.4 Lemma (Log-Covering Number Bound) *Let X_1, X_2 be non-empty sets and k_1, k_2 be bounded kernels on X_1 and X_2 with RKHSs H_1 and H_2 as well as ℓ_∞-embeddings $I[X_1]$ and $I[X_2]$, respectively. Furthermore, let k be the product kernel of k_1 and k_2 on $X_1 \times X_2$ defined in* (9.1) *with ℓ_∞-embedding $I[X_1 \times X_2]$. If $U_1 = \operatorname{span}\{u_1\} \subseteq H_1$ with $\|u_1\|_{H_1} = 1$ is an one-dimensional subspace and $P(U_1 \otimes H_2)$ the orthogonal projection onto $U_1 \otimes H_2 \subseteq H_1 \otimes H_2$ then the following log-covering number bounds are satisfied, for $\varepsilon > 0$,*

$$\mathcal{H}\big(I[X_1 \times X_2] \circ P(U_1 \otimes H_2), \varepsilon\big) \geq \mathcal{H}\big(I[X_2], 2\varepsilon / \|u_1\|_{\ell_\infty(X_1)}\big)$$
$$\mathcal{H}\big(I[X_1 \times X_2] \circ P(U_1 \otimes H_2), \varepsilon\big) \leq \mathcal{H}\big(I[X_2], \varepsilon / \|u_1\|_{\ell_\infty(X_1)}\big) \ .$$

Note that the upper bound can be generalized to general finite-dimensional subspaces U_1 using bounds for the covering numbers of so-called vector-valued diagonal operators. Since this generalization does not improve our application in Section 9.2 below, we stick to this version for convenience.

Proof. For the proof we use the commutative diagram

$$\begin{tikzcd}
U_1 \otimes H_2 \arrow[r, "{I[X_1 \times X_2]}"] \arrow[d, "Q"'] & \ell_\infty(X_1 \times X_2) \\
H_2 \arrow[r, "{I[X_2]}"'] & \ell_\infty(X_2) \ . \arrow[u, "M"']
\end{tikzcd}$$

To this end, we first introduce and investigate the operators Q and M.

Since u_1 is an ONB of U_1, we know from Lemma 9.1.1 that for every $h \in U_1 \otimes H_2$ there is a unique $h_2 \in H_2$ with $h = u_1 \otimes h_2$. Using this representation we define the operator $Q \colon U_1 \otimes H_2 \to H_2$ by $Q(u_1 \otimes h_2) := h_2$. Conversely, for every $h_2 \in H_2$ we have $u_1 \otimes h_2 \in U_1 \otimes H_2$ and $Q(u_1 \otimes h_2) = h_2$. Together we find that Q is bijective. Moreover, $\|u_1\|_{H_1} = 1$ and Lemma 9.1.1 give us

$$\|Q(u_1 \otimes h_2)\|_{H_2} = \|h_2\|_{H_2} = \|u_1\|_{H_1} \cdot \|h_2\|_{H_2} = \|u_1 \otimes h_2\|_{H_1 \otimes H_2}$$

and hence Q is an isometric isomorphism.

The operator $M\colon \ell_\infty(X_2) \to \ell_\infty(X_1 \times X_2)$ is defined by $Mh = u_1 \otimes h$ for $h \in \ell_\infty(X_2)$. Consequently, we have

$$\|Mh\|_{\ell_\infty(X_1\times X_2)} = \|u_1\|_{\ell_\infty(X_1)}\|h\|_{\ell_\infty(X_2)}$$

and M is the multiple of an isometric mapping with $\|M\| = \|u_1\|_{\ell_\infty(X_1)}$.

It is easy to check that the above diagram commutes. Since Q is an isometric isomorphism, we get from Lemma C.9 the claimed upper bound, namely

$$\begin{aligned}\mathcal{H}\big(I[X_1 \times X_2] \circ P(U_1 \otimes H_2), \varepsilon\big) &= \mathcal{H}\big(M \circ I[X_2] \circ Q, \varepsilon\big)\\ &= \mathcal{H}\big(M \circ I[X_2], \varepsilon\big)\\ &\le \mathcal{H}\big(I[X_2], \varepsilon/\|M\|\big)\ .\end{aligned}$$

Since M is the multiple of an isometric mapping, we get from (C.4) the claimed lower bound and hence the proof is finished. □

9.2 Covering Number Bounds

In this section we provide log-covering number bounds for the ℓ_∞-embedding $I_{\boldsymbol{\sigma}}[X]$ of the Gaussian RKHS $H_{\boldsymbol{\sigma}}(X)$ on the infinite-dimensional domain

$$X := \prod_{i\ge1}[-r_i, r_i] \subseteq \ell_2$$

with some positive sequence $r = (r_i)_{i\ge1}$. To this end, we use the fact that the Gaussian kernel on X equals the product kernel of the Gaussian kernel on the infinite-dimensional set $X_1 := \prod_{i>d}[-r_i, r_i]$ and the Gaussian kernel on the finite-dimensional set $X_2 := \prod_{i=1}^{d}[-r_i, r_i]$. On the infinite-dimensional set X_1 we use a trivial bound and on the finite-dimensional set X_2 we use Lemma 8.1.2. Here we benefit from the explicit constants provided by Lemma 8.1.2. However, let us start with some preparatory lemmas for the isotropic Gaussian kernel. To this end, recall the notation introduced in Section 7.1.

9.2.1 Lemma (Operator Norm on a Subspace) *Let $I \subseteq \mathbb{N}$ be non-empty, $X \subseteq \ell_2(I)$, $x_0 \in X$, $R(X) \coloneqq \sup_{x\in X} \|x - x_0\|_{\ell_2(I)}$, $\sigma > 0$, and $U \coloneqq \operatorname{span}\{k_\sigma(x_0, \cdot\,)\} \subseteq H_\sigma(X)$. Then, for the orthogonal projection*

$$P(U^\perp)\colon H_\sigma(X) \to H_\sigma(X)$$

onto $U^\perp$, the following operator norm bound is satisfied

$$\big\|I_\sigma[X] \circ P(U^\perp)\big\|^2 = 1 - \exp\!\big(-2\sigma^2 R^2(X)\big) \le 2\sigma^2 R^2(X) \ .$$

This result generalizes the bound (8.1) used in the proof of Lemma 8.1.1 from finite index sets I to infinite index sets in the case $N = 1$.

Proof. Note that $k_\sigma(x_0, \cdot\,) \in H_\sigma(X)$ has the norm

$$\|k_\sigma(x_0, \cdot\,)\|^2_{H_\sigma(X)} = k_\sigma(x_0, x_0) = 1 \ .$$

Consequently, the orthogonal projection $P(U)\colon H_\sigma(X) \to H_\sigma(X)$ onto U is given by

$$P(U)f = \big\langle f, k_\sigma(x_0, \cdot\,)\big\rangle_{H_\sigma(X)} k_\sigma(x_0, \cdot\,) \ .$$

Since $P(U^\perp) = \operatorname{Id} - P(U)$ holds true, we find

$$\big\|I_\sigma[X] \circ P(U^\perp)\big\|^2 = \sup_{f \in B_{H_\sigma(X)}} \sup_{x \in X} \big| f(x) - \big\langle f, k_\sigma(x_0, \cdot\,)\big\rangle_{H_\sigma(X)} k_\sigma(x_0, x)\big|^2 \ .$$

Interchanging the suprema and using the reproducing property $f(x) = \langle f, k_\sigma(x, \cdot\,)\rangle_{H_\sigma(X)}$ yields

$$\begin{aligned} &\big\|I_\sigma[X] \circ P(U^\perp)\big\|^2 \\ = \ &\sup_{x\in X} \sup_{f \in B_{H_\sigma(X)}} \big|\big\langle f, k_\sigma(x, \cdot\,) - k_\sigma(x_0, \cdot\,) k_\sigma(x_0, x)\big\rangle_{H_\sigma(X)}\big|^2_{H_\sigma(X)} \\ = \ &\sup_{x\in X} \big\|k_\sigma(x, \cdot\,) - k_\sigma(x_0, \cdot\,) k_\sigma(x_0, x)\big\|^2_{H_\sigma(X)} \ . \end{aligned}$$

For every $x \in X$ we have $\langle k_\sigma(x_0, \cdot\,), k_\sigma(x, \cdot\,)\rangle_{H_\sigma(X)} = k_\sigma(x, x_0)$ and hence

$\|k_\sigma(x,\,\cdot\,)\|^2_{H_\sigma(X)} = k_\sigma(x,x) = 1$. As a result, we find

$$\begin{aligned}&\Big\|k_\sigma(x,\,\cdot\,) - k_\sigma(x_0,\,\cdot\,)k_\sigma(x_0,x)\Big\|^2_{H_\sigma(X)}\\ &= k_\sigma(x,x) - 2k_\sigma^2(x,x_0) + k_\sigma(x_0,x_0)k_\sigma^2(x_0,x)\\ &= 1 - k_\sigma^2(x,x_0)\ \ .\end{aligned}$$

Taking the supremum over $x \in X$ gives the claimed equality and an application of $1 - e^{-t} \leq t$, for $t = 2\sigma^2 R^2(X)$ yields the claimed inequality. □

The next lemma uses the fact that Gaussian kernels are product kernels to split the estimation of $\mathcal{H}(I_\sigma[X_1 \times X_2], \varepsilon)$ into two parts, namely $\mathcal{H}(I_\sigma[X_1], \varepsilon)$, which is bounded by a trivial bound, and $\mathcal{H}(I_\sigma[X_2], \varepsilon)$.

9.2.2 Lemma (Splitting) *Let $I_1, I_2 \subseteq \mathbb{N}$ be disjoint non-empty index sets, $X_1 \subseteq \ell_2(I_1)$, $X_2 \subseteq \ell_2(I_2)$ be non-empty subsets such that X_1 is precompact, and $\sigma > 0$. Then, for $I := I_1 \uplus I_2$, the following log-covering number bound is satisfied, for $\varepsilon, \delta > 0$,*

$$\mathcal{H}\Big(I_\sigma[X_1 \times X_2], \varepsilon + \delta\sqrt{2}\Big) \leq \mathcal{N}(\sigma X_1, \delta) \cdot \mathcal{H}\big(I_\sigma[X_2]I_,\varepsilon\big)\ \ .$$

Note that Lemma 9.2.2 holds for infinite index sets I_1, I_2 and that the covering numbers $\mathcal{N}(\sigma X_1, \delta)$ are with respect to the $\ell_2(I_1)$-norm.

Proof. Let $\varepsilon, \varepsilon_0, \delta > 0$ be fixed. First, we apply (8.12) and Lemma 8.3.3 to exchange X_1 by the closed ball $\delta B_{\ell_2(I_1)}$ of radius δ, namely

$$\begin{aligned}\mathcal{H}\Big(I_\sigma[X_1 \times X_2], \varepsilon + \varepsilon_0\Big) &= \mathcal{H}\Big(I_1\big[\sigma(X_1 \times X_2)\big], \varepsilon + \varepsilon_0\Big)\\ &\leq \mathcal{N}(\sigma X_1, \delta) \cdot \mathcal{H}\Big(I_1\big[\delta B_{\ell_2(I_1)} \times \sigma X_2\big], \varepsilon + \varepsilon_0\Big)\ \ .\end{aligned}$$

In the following we use the abbreviation $\tilde{X} := \delta B_{\ell_2(I_1)} \times \sigma X_2$. Using this notation it remains to prove $\mathcal{H}(I_1[\tilde{X}], \varepsilon + \varepsilon_0) \leq \mathcal{H}(I_\sigma[X_2], \varepsilon)$.

Since the Gaussian kernel k_1 on $\tilde{X}$ is the product kernel of two Gaussian kernels, namely the Gaussian kernel on $\delta B_{\ell_2(I_1)}$ and the Gaussian kernel on

$\sigma X_2 \subseteq \ell_2(I_2)$, we have $H_1(\tilde{X}) = H_1(\delta B_{\ell_2(I_1)}) \otimes H_1(\sigma X_2)$ according to the definition of the tensor product in (9.2). Now, let $U_1 \coloneqq \operatorname{span}\{k_1(0,\,\cdot\,)\} \subseteq H_1(\delta B_{\ell_2(I_1)})$ be a one-dimensional subspace and

$$P\big(U_1 \otimes H_1(\sigma X_2)\big), P\big(U_1^\perp \otimes H_1(\sigma X_2)\big) \colon H_1(\tilde{X}) \to H_1(\tilde{X})$$

the orthogonal projections onto $U_1 \otimes H_1(\sigma X_2)$ and $U_1^\perp \otimes H_1(\sigma X_2)$, respectively. Using Lemma 9.1.3, $\|I_1[\sigma X_2]\| = 1$, and Lemma 9.2.1 we get

$$\begin{aligned} \big\|I_1[\tilde{X}] \circ P\big(U_1^\perp \otimes H_1(\sigma X_2)\big)\big\| &= \big\|I_1[\delta B_{\ell_2(I_1)}] \circ P(U_1^\perp)\big\| \cdot \big\|I_1[\sigma X_2]\big\| \\ &\leq \delta\sqrt{2} =: \varepsilon_0 \ . \end{aligned} \tag{9.5}$$

Since the orthogonal complement of $U_1 \otimes H_1(\sigma X_2)$ in $H_1(\delta B_{\ell_2(I_1)}) \otimes H_1(\sigma X_2)$ is

$$\big(U_1 \otimes H_1(\sigma X_2)\big)^\perp = U_1^\perp \otimes H_1(\sigma X_2) \ ,$$

we have $\mathrm{Id}_{H_1(\tilde{X})} = P\big(U_1 \otimes H_1(\sigma X_2)\big) + P\big(U_1^\perp \otimes H_1(\sigma X_2)\big)$ and a combination of (C.3) with (9.5) gives us

$$\mathcal{H}\big(I_1[\tilde{X}], \varepsilon + \varepsilon_0\big) \leq \mathcal{H}\Big(I_1[\tilde{X}] \circ P\big(U_1 \otimes H_1(\sigma X_2)\big), \varepsilon\Big) \ .$$

Since $\|k_1(0,\,\cdot\,)\|_{\ell_\infty(\delta B_{\ell_2(I_1)})} = 1$ holds, an application of Lemma 9.1.4 yields

$$\mathcal{H}\big(I_1[\tilde{X}], \varepsilon + \varepsilon_0\big) \leq \mathcal{H}\big(I_1[\sigma X_2], \varepsilon\big) \ .$$

Finally, (8.12) gives $\mathcal{H}(I_1[\sigma X_2], \varepsilon) = \mathcal{H}(I_\sigma[X_2], \varepsilon)$ and hence the assertion is proven. □

In the case $|I_2| < \infty$ we can combine Lemma 9.2.2 with Lemma 8.1.2 for $d = |I_2|$. To this end, we recall the definition of the function

$$h_\sigma(y) = 2e\sigma^2 \exp\Big(W_0\Big(\frac{y}{e\sigma^2}\Big)\Big)$$

for $\sigma > 0$ and $y \geq -\sigma^2$ from Lemma 7.2.6. This approach leads to the log-covering number bound in the following lemma.

9.2.3 Lemma (Log-Covering Number Bound) *Let $I_1, I_2 \subseteq \mathbb{N}$ be disjoint non-empty index sets with $|I_2| < \infty$, $X_1 \subseteq \ell_2(I_1)$, $X_2 \subseteq \ell_2(I_2)$ be non-empty precompact subsets, and $\sigma > 0$. Then, for $X \coloneqq X_1 \times X_2$ and $I \coloneqq I_1 \uplus I_2$, the following log-covering number bound is satisfied, for $\delta_1, \delta_2 > 0$ and $0 < \varepsilon \le 1$,*

$$\mathcal{H}\big(I_\sigma[X], \varepsilon + \delta_1\sqrt{2}\big) \le \mathcal{N}(\sigma X_1, \delta_1)\mathcal{N}(\sigma X_2, \delta_2) \cdot \binom{(h_{\delta_2} \circ \log)(4/\varepsilon) + |I_2|}{|I_2|} \log(4/\varepsilon) \ .$$

Recall that the covering numbers $\mathcal{N}(\sigma X_1, \delta_1)$ and $\mathcal{N}(\sigma X_2, \delta_2)$ are with respect to the $\ell_2(I_1)$- and $\ell_2(I_2)$-norm, respectively.

Proof. After an application of Lemma 9.2.2 with $\delta = \delta_1$ it remains to bound $\mathcal{H}(I_\sigma[X_2], \varepsilon)$. With the help of Lemma 8.3.3, for $X_1 = X_2$ and $X_2 = \{0\}$, we can exchange X_2 by $\delta_2 B_{\ell_2(I_2)}$, namely

$$\mathcal{H}\big(I_\sigma[X_2], \varepsilon\big) \le \mathcal{N}(X_2, \delta_2) \cdot \mathcal{H}\big(I_{\delta_2}[B_{\ell_2(I_2)}], \varepsilon\big) \ .$$

Finally, an application of Lemma 8.1.2 on $\mathcal{H}(I_{\delta_2}[B_{\ell_2(I_2)}], \varepsilon)$ gives the assertion. □

In the following we consider the index set $I = \mathbb{N}$ and domains X of the form $X = \prod_{i\ge1}[-r_i, r_i]$ with some positive sequence $r = (r_i)_{i\ge1} \in \ell_2$. Then, for $d \ge 1$, we split the index set $I = I_1 \uplus I_2$ into $I_1 \coloneqq d + \mathbb{N}$ and $I_2 \coloneqq [d] = \{1, \dots, d\}$ and analogously, we split the domain $X = X_1 \times X_2$ into $X_1 \coloneqq \prod_{i>d}[-r_i, r_i]$ and $X_2 \coloneqq \prod_{i=1}^{d}[-r_i, r_i]$. This allows us to apply Lemma 9.2.3. Consequently, we have to choose the parameters $d \ge 1$, $\delta_1, \delta_2 > 0$, and $\varepsilon > 0$, depending on the sequence r, to get a reasonable log-covering number bound.

To this end, we use the following regularity assumption for the sequence $r = (r_i)_{i\ge1}$: For some $\beta > 0$ the supremum

$$c \coloneqq \sup_{k\le n} \frac{r_n e^{n\beta}}{r_k e^{k\beta}} < \infty \tag{EXP}$$

is finite. Note that this supremum is taken over all tuples $(k, n) \in \mathbb{N}^2$ with $k \leq n$. Moreover, (EXP) is independent of the scaling of r, i.e. if r satisfies (EXP) then $\sigma r = (\sigma r_i)_{i\geq 1}$ satisfies (EXP) for the same $c \geq 1$ and $\beta > 0$. If the sequence r is non-increasing then some characterizations of (EXP) can be found in Lemma 10.2.1 of Part III. We call this condition (EXP) because it implies an exponential decay and the sequence $r_i = ae^{-i\beta}$, for $i \geq 1$, is probably the most important example satisfying this condition with (optimal) constant $c = 1$. The following lemma provides a covering number bound for the infinite-dimensional set X_1 if r satisfies (EXP).

9.2.4 Lemma *Let $d \geq 1$, $r = (r_i)_{i\geq 1} \in \ell_2$ be a sequence with $r_i > 0$ for all $i \geq 1$, and $X \coloneqq \prod_{i\geq 1}[-r_i, r_i] \subseteq \ell_2$. If r satisfies* (EXP) *with $\beta > 0$ and $c \geq 1$ then the covering numbers of $X_1 \coloneqq \prod_{i>d}[-r_i, r_i]$ satisfy, for $\delta_1 > 0$,*

$$\mathcal{N}(X_1, \delta_1) \leq \mathcal{N}\big(X, \delta_1 e^{d\beta}/c\big) \ .$$

Proof. The condition in (EXP) yields $r_{i+d} \leq c r_i e^{-d\beta}$ for all $i \geq 1$. If we shift the index then we find

$$X_1 = \prod_{i\geq 1}[-r_{i+d}, r_{i+d}] \subseteq ce^{-d\beta} \prod_{i\geq 1}[-r_i, r_i] = ce^{-d\beta} X$$

and hence $\mathcal{N}(X_1, \delta_1) \leq \mathcal{N}(ce^{-d\beta} X, \delta_1)$. Finally, the scaling property of the covering numbers, see Lemma C.6, gives the assertion. □

Now, we are ready to prove our log-covering number bound for Gaussian kernels on the infinite-dimensional domain $X = \prod_{i\geq 1}[-r_i, r_i]$.

9.2.5 Lemma (Log-Covering Number Bound for (EXP) Sequences) *Let $\sigma > 0$, $r = (r_i)_{i\geq 1} \in \ell_2$ with $r_i > 0$ for all $i \geq 1$, and $X \coloneqq \prod_{i\geq 1}[-r_i, r_i] \subseteq \ell_2$. If r satisfies* (EXP) *for $\beta > 0$ and $c \geq 1$ then, for $0 < \varepsilon \leq 1$ and $y \coloneqq \log((4 + \sqrt{2})/\varepsilon)$, the following log-covering number bound is satisfied*

$$\mathcal{H}\big(I_\sigma[X], \varepsilon\big) \leq \mathcal{N}^2(\sigma X, 1) \cdot \exp\!\left(\frac{2y\big(1 + W_0(y/e)\big)}{W_0(y/e)}\big(1 + o(1)\big)\right) \ ,$$

where $o(1)$ denotes a function converging to 0 for $\varepsilon \to 0^+$.

Note that the kernel width σ influences only the constant and *not* the asymptotic behavior of the upper bound for $\varepsilon \to 0^+$.

Proof. For $d \geq 1$ we define $X_1 := \prod_{i=d+1}^{\infty}[-r_i, r_i] \subseteq \ell_2(d+\mathbb{N})$ and $X_2 := \prod_{i=1}^{d}[-r_i, r_i] \subseteq \ell_2^d$. An application of Lemma 9.2.3 with $\varepsilon = 4\varepsilon/(4+\sqrt{2})$ and $\delta_1 = \varepsilon/(4+\sqrt{2}) = e^{-y}$ together with $\log(4/\varepsilon) \leq y$ gives us

$$\mathcal{H}\bigl(I_\sigma[X], \varepsilon\bigr) \leq \mathcal{N}(\sigma X_1, e^{-y})\mathcal{N}(\sigma X_2, \delta_2)\binom{h_{\delta_2}(y)+d}{d} y$$

for $\delta_2 > 0$, $d \geq 1$, and $0 < \varepsilon \leq 1$. Since r satisfies (EXP), Lemma 9.2.4 yields

$$\mathcal{N}(\sigma X_1, e^{-y}) \leq \mathcal{N}\bigl(\sigma X, e^{-y+d\beta}/c\bigr) \ .$$

The choice $d := \bigl\lceil \bigl(\log(c\delta_2)+y\bigr)/\beta \bigr\rceil$ implies

$$\frac{e^{-y}e^{d\beta}}{c} \geq \frac{e^{-y}e^{\log(c\delta_2)+y}}{c} = \delta_2 \ .$$

and hence we find $\mathcal{N}(\sigma X_1, e^{-y}) \leq \mathcal{N}(\sigma X, \delta_2)$. Together with $\mathcal{N}(\sigma X_2, \delta_2) \leq \mathcal{N}(\sigma X, \delta_2)$ we get the upper bound

$$\mathcal{H}\bigl(I_\sigma[X], \varepsilon\bigr) \leq \mathcal{N}^2(\sigma X, \delta_2)\binom{h_{\delta_2}(y)+d}{d} y \ .$$

Now, we choose $\delta_2 = 1$ and use Lemma 7.2.1 as well as Stirling's formula to get

$$\begin{aligned}\binom{h_1(y)+d}{d} &= \frac{\Gamma(h_1(y)+d+1)}{\Gamma(h_1(y)+1)\Gamma(d+1)} \\ &\leq \frac{e^{1/24}}{\sqrt{2\pi}}\bigl(1/h_1(y)+1/d\bigr)^{1/2}\left(1+\frac{h_1(y)}{d}\right)^{d}\left(1+\frac{d}{h_1(y)}\right)^{h_1(y)} \ .\end{aligned}$$

It remains to investigate the asymptotic behavior of all the factors depending on ε (or y). To this end, we give the function in the exponent of our desired bound the name

$$f(y) := \frac{2y\log\bigl(1+W_0(y/e)\bigr)}{W_0(y/e)} \ .$$

Recall from Lemma 7.2.6 that

$$h_1(y) = \frac{2y}{W_0(y/e)} \quad .$$

First, using $W_0(y/e) \sim \log(y)$ from Lemma 7.2.3 we find $h_1(y) \to \infty$ and $d \sim y/\beta \to \infty$ for $\varepsilon \to 0^+$. As a result,

$$\frac{e^{1/24}}{\sqrt{2\pi}} \bigl(1/h_1(y) + 1/d\bigr)^{1/2}$$

is bounded and hence gets absorbed in $\exp(f(y) \cdot o(1))$. Second, the factor y satisfies $\log(y) = o(f(y))$ and hence it gets absorbed in $\exp(f(y) \cdot o(1))$, too. Third, using $1 + t \leq e^t$, for $t = h_1(y)/d$, yields

$$\log\left(1 + \frac{h_1(y)}{d}\right)^d \leq \log \circ e^{h_1(y)} = h_1(y) = o\bigl(f(y)\bigr) \quad .$$

and hence the third factor gets absorbed in $\exp\bigl(f(y) \cdot o(1)\bigr)$. Finally, the logarithm of the last factor behaves like

$$\begin{aligned} h_1(y) \log\left(1 + \frac{d}{h_1(y)}\right) &\sim \frac{2y}{W_0(y/e)} \log\left(1 + \frac{W_0(y/e)}{2\beta}\right) \\ &\sim \frac{2y}{W_0(y/e)} \log\bigl(1 + W_0(y/e)\bigr) = f(y) \quad . \end{aligned}$$

As a result, the last factor is of the type $\exp\bigl(f(y)\bigl(1 + o(1)\bigr)\bigr)$. Combining all estimates we get the desired bound. □

The final theorem presents a simplified, but more convenient, version of Lemma 9.2.5 for anisotropic Gaussian kernels. Except for a square, this theorem recovers the log-covering number bound in (8.15) for Gaussian RKHSs on finite-dimensional domains.

9.2.6 Theorem *Let $\boldsymbol{\sigma} = (\sigma_i)_{i\geq 1}$, $r = (r_i)_{i\geq 1}$ be sequences with $\sigma_i, r_i > 0$ for $i \geq 1$ and $X := \prod_{i\geq 1}[-r_i, r_i] \subseteq \ell_2$. If there is a real number $\beta > 0$ such that the sequence $\boldsymbol{\sigma} r := (\sigma_i r_i)_{i\geq 1}$ satisfies* (EXP) *with $c \geq 1$ then for every*

$p > 0$ there is a constant $C_{\beta,c,p} > 0$ such that the following log-covering number bound is satisfied, for $0 < \varepsilon \leq 1$,

$$\mathcal{H}\big(I_\sigma[X], \varepsilon\big) \leq C_{\beta,c,p} \cdot \mathcal{N}^2(D_{\boldsymbol{\sigma}} X, 1) \cdot \varepsilon^{-p} \ .$$

Proof. Let $p > 0$ be fixed and $y \coloneqq \log((4 + \sqrt{2})/\varepsilon)$ as in Lemma 9.2.5. Using $\mathcal{H}(I_{\boldsymbol{\sigma}}[X], \varepsilon) = \mathcal{H}(I_1[D_{\boldsymbol{\sigma}} X], \varepsilon)$ from (8.12) and Lemma 9.2.5 it is enough to show that

$$\varepsilon^p \exp\bigg(\frac{2y \log\big(1 + W_0(y/e)\big)}{W_0(y/e)}\big(1 + o(1)\big)\bigg)$$

is bounded (or even converges to 0) for $\varepsilon \to 0^+$. However, $\varepsilon^p = (4+\sqrt{2})^p e^{-py}$ and

$$\frac{\log\big(1 + W_0(y/e)\big)}{W_0(y/e)} \to 0$$

for $\varepsilon \to 0^+$ already proves this claim. □

To the best of our knowledge, there is no bound in the literature which is directly comparable to our bound. In [75, Corollary 3.5] is a bound for general 1-Hölder continuous operators, which includes the operator $I_{\boldsymbol{\sigma}}[X]$. To be more precise, they use the condition

$$e_n(D_{\boldsymbol{\sigma}} X) \preccurlyeq n^{-1/p} \tag{9.6}$$

for some $0 < p < \infty$ on the dyadic entropy numbers of $D_{\boldsymbol{\sigma}} X$ to provide the bound

$$e_n\big(I_{\boldsymbol{\sigma}}[X]\big) \preccurlyeq \begin{cases} n^{-1/p}, & 2 < p < \infty \\ n^{-1/2} \log^{1/2-1/p}(n), & 0 < p < 2 \ . \end{cases} \tag{9.7}$$

To compare this with our findings we translate Theorem 9.2.6 into a dyadic entropy number bound. Using (C.2) Theorem 9.2.6 reads as follows: If the sequence $\boldsymbol{\sigma} r$ satisfies (EXP) then we have

$$e_n\big(I_{\boldsymbol{\sigma}}[X]\big) \preccurlyeq n^{-1/p}$$

for all $p > 0$. First, we compare the assumptions. The condition (EXP) implies $\sigma_i r_i \preccurlyeq e^{-\beta i}$ and hence analogously to the proof of [58, Proposition 6], for real sequence spaces, we find

$$e_n(D_{\boldsymbol{\sigma}} X) = e_n\big(D_{\boldsymbol{\sigma} r} B_{\ell_\infty}\big) \preccurlyeq n^{1/4} \cdot \exp\Big(-\sqrt{2\beta \log(2) n}\Big) \ .$$

As a result, (EXP) is stronger than the condition in (9.6) used by [75, Corollary 3.5]. But our bound on $e_n(I_{\boldsymbol{\sigma}}[X])$ is stronger than the bound in (9.7) provided by [75, Corollary 3.5]. More precisely, the polynomial decay of the bound in (9.7) saturates at the polynomial order of 1/2, however, our bound decreases faster than any polynomial.

Part III

Diagonal Operators

In many applications discretization techniques are used to reduce the often difficult problem of estimating entropy numbers in function spaces to easier estimation problems in sequence spaces. For instance, the problem of quantifying the compactness of Sobolev embeddings can be reduced to diagonal operators in sequence spaces via wavelet or Fourier bases, see e.g. [58, 21] and references therein. In this part we therefore derive new entropy number bounds for diagonal operators $D_\sigma \colon \ell_p \to \ell_q$, where $p \neq q$. In the case $p < q$ we prove the optimality for fast decaying diagonal sequences, which include exponentially decreasing sequences. In the case $p > q$ we show optimality under weaker assumptions than previously used in the literature. The content of this part is mainly taken from the article:

> S. Fischer. Some new bounds on the entropy numbers of diagonal operators. *J. Approx. Theory*, 251:105343, 2020.

Chapter 10

Introduction and Preparation

In the first section of this chapter we introduce the notation and provide some preparatory material. In Section 10.2 we give a brief introduction to regularity conditions for sequences. Such regularity conditions allow us to prove the optimality of our entropy number bounds.

10.1 Definitions and Basic Properties

Since we reduce the investigation of diagonal operators on sequence spaces to the case of diagonal operators on $\mathbb{R}^k$ we use a unifying notation. To this end, we consider sequences over an index set $I \subseteq \mathbb{N}$ and recall, for $0 < p \leq \infty$, the definition of the sequence space $\ell_p(I) \coloneqq \bigl\{x = (x_i)_{i\in I} \in \mathbb{R}^I : \|x\|_{\ell_p(I)} < \infty\bigr\}$ with the (quasi-)norm

$$\|x\|_{\ell_p(I)} \coloneqq \begin{cases} \bigl(\sum_{i\in I} |x_i|^p\bigr)^{1/p}, & 0 < p < \infty \\ \sup_{i\in I} |x_i|, & p = \infty \end{cases}$$

and the closed unit ball $B_{\ell_p(I)}$. With this notation we have $\ell_p = \ell_p(\mathbb{N})$ and $\ell_p^k = \ell_p([k])$ for $k \geq 1$. It is well-known that $\ell_p(I)$ is a quasi-Banach space for all $0 < p \leq \infty$ and that $\ell_p(I)$ is a Banach space if and only if $1 \leq p \leq \infty$. Moreover, the quasi-triangle constant is $\kappa_p \coloneqq \kappa_{\ell_p(I)} = \max\bigl\{2^{1/p-1}, 1\bigr\}$.

In the following we fix some $0 < p, q \leq \infty$, a sequence $\sigma = (\sigma_i)_{i\in I} \in \mathbb{R}^I$, and the diagonal operator $D_\sigma \colon \ell_p(I) \to \ell_q(I)$ given by $D_\sigma(x_i)_{i\in I} \coloneqq (\sigma_i x_i)_{i\in I}$. As a consequence of Hölder's inequality the operator norm of D_σ

is given by

$$\|D_\sigma\| = \begin{cases} \|\sigma\|_{\ell_r(I)}, & p > q,\ 1/q = 1/p + 1/r \\ \|\sigma\|_{\ell_\infty(I)}, & p \leq q\ . \end{cases} \tag{10.1}$$

Consequently, D_σ is well-defined and bounded if and only if $\sigma \in \ell_r(I)$ in the case $p > q$ and $\sigma \in \ell_\infty(I)$ in the case $p \leq q$. Moreover, we define the auxiliary operators

$$\begin{aligned} D^k_{p,q} &\colon \ell^k_p \to \ell^k_q,\ (x_n)^k_{n=1} \mapsto (\sigma_1 x_1, \ldots, \sigma_k x_k)\ , \\ P^k_p &\colon \ell_p \to \ell^k_p,\ (x_n)_{n\geq 1} \mapsto (x_1, \ldots, x_k)\ , \\ I^k_p &\colon \ell^k_p \to \ell_p,\ (x_n)^k_{n=1} \mapsto (x_1, \ldots, x_k, 0, 0, \ldots)\ . \end{aligned} \tag{10.2}$$

Note that these operators satisfy $D^k_{p,q} = P^k_q D_\sigma I^k_p$ and $\|I^k_p\| = \|P^k_p\| = 1$.

The goal of this part is the investigation of the asymptotic behavior of the entropy numbers of D_σ. For $n \geq 1$, the *entropy number* $\varepsilon_n(D_\sigma)$ is defined as the infimum over all $r > 0$ that allows to cover $D_\sigma B_{\ell_p}$ with n translates of rB_{ℓ_q}. Note that we even need covering and packing numbers for the proofs. For the definitions and basic properties of these metric entropy quantities see Appendix C.

In the case $p = q$ the asymptotic behavior of the entropy numbers $\varepsilon_n(D_\sigma)$ is well-known for *all* diagonal sequences σ, see e.g. Gordon et al. [39, Proposition 1.7] for the Banach space case $1 \leq p \leq \infty$ but, modulo the constant, the result remains valid for all $0 < p \leq \infty$. Therefore, we focus on the case $p \neq q$, where—as far as we know—are only partial answers available, see e.g. [57, 58, 18].

Recall that for real sequences $(x_n)_{n\geq 1}$ and $(y_n)_{n\geq 1}$ we write $x_n \preccurlyeq y_n$ if there is a constant $c > 0$ with $x_n \leq c y_n$ for all $n \geq 1$ and the (weak) asymptotic equivalence $x_n \asymp y_n$ means $x_n \preccurlyeq y_n$ as well as $x_n \succcurlyeq y_n$. Using this notion, we declare an upper bound $(x_n)_{n\geq 1}$ on the entropy numbers to be *optimal* if there is a corresponding lower bound $(y_n)_{n\geq 1}$, which is asymptotically equivalent $x_n \asymp y_n$. In our bounds the ratio of the volumes of the unit balls $B_{\ell^k_p}$ and $B_{\ell^k_q}$ play an important role. To this end, recall that λ^k denotes the k-dimensional Lebesgue measure and that a combination of

[70, Equation (1.17)] with Stirling's formula yields, for $k \to \infty$,

$$\left(\frac{\lambda^k(B_{\ell_p^k})}{\lambda^k(B_{\ell_q^k})}\right)^{1/k} \asymp k^{1/q-1/p} \quad . \tag{10.3}$$

Finally, the following lemma establishes a comparison between covering and entropy numbers.

10.1.1 Lemma *Let $0 < p, q \leq \infty$, $(a_k)_{k\geq 1}$ be a positive sequence and $D_\sigma \colon \ell_p \to \ell_q$ be a diagonal operator with $\|D_\sigma\| < \infty$. If the covering number estimate*

$$\mathcal{N}(D_\sigma, \varepsilon) \leq \sup_{k\geq 1} a_k \left(\frac{1}{\varepsilon}\right)^k \tag{10.4}$$

holds true for all $0 < \varepsilon < \varepsilon_1(D_\sigma)$ then for $n \geq 1$ the n-th entropy number satisfies

$$\varepsilon_n(D_\sigma) \leq \sup_{k\geq 1} \left(\frac{a_k}{n}\right)^{1/k} \quad .$$

Note that Point (ii) and (iii) in Lemma C.9 in the language of entropy numbers read $\|D_\sigma\|/\kappa_q \leq \varepsilon_1(D_\sigma) \leq \|D_\sigma\|$. Consequently, in Lemma 10.1.1 it is sufficient to check (10.4) for all $0 < \varepsilon < \|D_\sigma\|$.

Proof. Let $n \geq 1$ be a natural number. If $\varepsilon_n(D_\sigma) = 0$ holds true then $D_\sigma = 0$ is the zero operator and there is nothing to prove. In the following we assume $\varepsilon_n(D_\sigma) > 0$ and choose $0 < \varepsilon < \varepsilon_n(D_\sigma)$. By the contrapositive of Point (i) in Lemma C.1 we have $n < \mathcal{N}(D_\sigma, \varepsilon)$. Moreover, by our assumption for every $\delta > 0$ there is a $k_\delta \geq 1$ with

$$n \leq \mathcal{N}(D_\sigma, \varepsilon) \leq (1+\delta)\, a_{k_\delta} \left(\frac{1}{\varepsilon}\right)^{k_\delta} \quad .$$

This implies

$$\varepsilon \leq \left(\frac{(1+\delta)\, a_{k_\delta}}{n}\right)^{1/k_\delta} \leq (1+\delta) \left(\frac{a_{k_\delta}}{n}\right)^{1/k_\delta} \leq (1+\delta) \sup_{k\geq 1} \left(\frac{a_k}{n}\right)^{1/k} \quad .$$

Letting $\delta \searrow 0$ and $\varepsilon \nearrow \varepsilon_n(D_\sigma)$ we get the assertion. □

10.2 Conditions on Sequences

In this section we introduce the regularity conditions on the diagonal sequence $\sigma = (\sigma_k)_{k\geq 1}$ that we need to prove the optimality of our entropy number bounds. Furthermore, we collect some characterizations of these conditions that allows us to compare our results with the existing literature. Most of the following results are consequences of the general theory of $\mathcal{O}$-regular varying functions/sequences, but for convenience we include the proofs or give detailed references. Throughout this section, all suprema $\sup_{k\leq n}$ and infima $\inf_{k\leq n}$ are taken over all tuples $(n,k) \in \mathbb{N}^2$ with $k \leq n$.

Let $\sigma = (\sigma_k)_{k\geq 1}$ be a positive and non-increasing sequence with $\sigma_k \searrow 0$ for $k \to \infty$. Then we define the *r-tail sequence*, for $r > 0$ and $n \geq 1$, by

$$\tau_n := \left(\sum_{k=n}^{\infty} \sigma_k^r\right)^{1/r} \tag{10.5}$$

and we define the *s-partial sum sequence*, for $s > 0$ and $n \geq 1$, by

$$\upsilon_n := \left(\sum_{k=1}^{n} \sigma_k^{-s}\right)^{1/s} . \tag{10.6}$$

Note that the r-tail sequence is well-defined if and only if $\sigma \in \ell_r$. Now, we are ready to formulate our regularity conditions on σ: We say that $\sigma = (\sigma_k)_{k\geq 1}$ satisfies

(EXP) if there is some $b > 1$ with $\sup_{k\leq n} \frac{\sigma_n b^n}{\sigma_k b^k} < \infty$.

(ALP) if $\sigma \in \ell_r$ and $\tau_n \preccurlyeq \sigma_n n^{1/r}$.

(AMP) if $\sigma \in \ell_r$ and $\tau_n \succcurlyeq \sigma_n n^{1/r}$.

Note that we already used the Condition (EXP) in Section 9.2 of Part II. Moreover, (EXP) implies an *exponential decay* $\sigma_n \preccurlyeq b^{-n}$. This is the reason why we call this condition (EXP). The abbreviation (ALP) stands for *at least polynomial decay* since this condition implies $\sigma_n \preccurlyeq n^{-\alpha}$ for some $\alpha > 1/r$ according to Point (i) of Lemma 10.2.3 below. In Lemma 10.2.3 below we see that (AMP) implies $\tau_n \succcurlyeq n^{-\alpha}$ for some $\alpha > 0$ and hence we called this condition *at most polynomial decay.* Our first lemma provides characterizations of the Condition (EXP).

10.2.1 Lemma ((EXP) Sequences) *Let $r, s > 0$, $\sigma = (\sigma_k)_{k\geq 1}$ with $\sigma_k > 0$ for $k \geq 1$ and $\sigma_k \searrow 0$ for $k \to \infty$, $\tau = (\tau_n)_{n\geq 1}$ be the r-tail sequence given by* (10.5)*, and $\upsilon = (\upsilon_n)_{n\geq 1}$ be the s-partial sum sequence given by* (10.6)*. Then the following statements are equivalent:*

(i) σ satisfies (EXP).

(ii) There is an $n_0 \geq 1$ and an $0 < a < 1$ with $\sigma_{k+n_0} \leq a\,\sigma_k$ for all $k \geq 1$.

(iii) $\sigma_n \asymp 1/\upsilon_n$.

(iv) $\sigma \in \ell_r$ and $\sigma_n \asymp \tau_n$.

Note that Condition (i) and (ii) are independent of $r > 0$ and $s > 0$. Consequently, if σ satisfies Condition (iii) or (iv) for some $s > 0$ or $r > 0$ then σ satisfies both conditions for all $r, s > 0$. Furthermore, from Point (iv) we get (EXP) $\subseteq$ (ALP) and (EXP) $\cap$ (AMP) $= \emptyset$.

Proof. (i)⇒(iii) For $c := \sup_{k\leq n} \frac{\sigma_n b^n}{\sigma_k b^k} < \infty$ we get

$$\upsilon_n^s \sigma_n^s = \sum_{k=1}^{n} \Big(\frac{\sigma_n}{\sigma_k}\Big)^s \leq c^s \sum_{k=1}^{n} b^{-s(n-k)} = c^s \sum_{k=0}^{n-1} b^{-sk} \leq \frac{(bc)^s}{b^s - 1}$$

for all $n \geq 1$. Moreover, $\upsilon_n \sigma_n \geq 1$ always holds. By considering $(\tau_k/\sigma_k)^r$ we can analogously prove (i)⇒(iv).

(iii)⇒(ii) Let $c > 0$ be a constant with $\upsilon_n \sigma_n \leq c$ for all $n \geq 1$. Because of the monotonicity of σ we get for $k, n_0 \geq 1$

$$c^s \geq \upsilon_{k+n_0}^s \sigma_{k+n_0}^s = \sum_{i=1}^{k+n_0} \Big(\frac{\sigma_{k+n_0}}{\sigma_i}\Big)^s \geq \sum_{i=k}^{k+n_0} \Big(\frac{\sigma_{k+n_0}}{\sigma_i}\Big)^s \geq \Big(\frac{\sigma_{k+n_0}}{\sigma_k}\Big)^s (n_0{+}1)\ .$$

Choosing $n_0 := \lceil c^s \rceil$ yields, for $k \geq 1$,

$$\frac{\sigma_{k+n_0}}{\sigma_k} \leq \frac{c}{(n_0+1)^{1/s}} \leq \frac{c}{(c^s+1)^{1/s}} < 1\ .$$

(iv)⇒(ii) Let $c > 0$ be a constant with $\tau_k \leq c\sigma_k$ for all $k \geq 1$. Because of the monotonicity of σ we get for $k, n_0 \geq 1$

$$c^r \geq \frac{\tau_k^r}{\sigma_k^r} = \sum_{n=k}^{\infty} \Big(\frac{\sigma_n}{\sigma_k}\Big)^r \geq \sum_{n=k}^{k+n_0} \Big(\frac{\sigma_n}{\sigma_k}\Big)^r \geq \Big(\frac{\sigma_{k+n_0}}{\sigma_k}\Big)^r (n_0 + 1) \ .$$

Hence Statement (ii) follows along the same line as (iii)⇒(ii).

(ii)⇒(i) For $k \leq n$ there is a unique $m \geq 0$ with $k + mn_0 \leq n < k + (m+1)n_0$. Using the monotonicity of σ and Assumption (ii) m-times we get

$$\sigma_n \leq \sigma_{k+mn_0} \leq \sigma_k a^m \leq \frac{\sigma_k}{a} a^{\frac{n-k}{n_0}} = \frac{\sigma_k}{a} b^{k-n}$$

with $b = a^{-1/n_0} > 1$. Hence the supremum is bounded by a^{-1}. □

Another important condition is the *doubling condition* $\sigma_{2n} \asymp \sigma_n$. The following lemma provides equivalent characterizations for this condition.

10.2.2 Lemma (Doubling Condition) *Let $\sigma = (\sigma_k)_{k\geq 1}$ with $\sigma_k > 0$ for $k \geq 1$ and $\sigma_k \searrow 0$ for $k \to \infty$. Then the following statements are equivalent:*

(i) $\sigma_n \asymp \sigma_{2n}$.

(ii) For all $\lambda > 0$ the function $f(x) := \sigma_{\lfloor x \rfloor + 1}$ satisfies $f(x) \asymp f(\lambda x)$.

(iii) $\inf_{k \leq n} \frac{\sigma_n n^\alpha}{\sigma_k k^\alpha} > 0$ *for some $\alpha > 0$.*

(iv) $\sigma_n \asymp (\sigma_1 \cdot \ldots \cdot \sigma_n)^{1/n}$.

Note that the symbol "$\asymp$" in Point (ii) means that for all $\lambda > 0$ there are constants $c_1, c_2 > 0$, depending on $\lambda > 0$, with $c_1 f(x) \leq f(\lambda x) \leq c_2 f(x)$ for all $x > 0$. Moreover, Point (iii) implies $\sigma_n \succcurlyeq n^{-\alpha}$ and hence σ decreases at most polynomially. Finally, the Point (iv) has already been used in Lemma 5.2.3 of Part I as an assumption. Consequently, this lemma provides equivalent assumptions for Lemma 5.2.3 .

Proof. (i)⇔(iii) This has already been pointed out by Kühn [57, p. 482] and is a direct consequence of the monotonicity of σ.

(i)⇔(ii) Statement (ii), for $\lambda = 2$ and $x = n - 1/2$, directly implies (i). For the converse implication we first show that

$$\lfloor nx \rfloor + 1 \leq n\big(\lfloor x \rfloor + 1\big) \tag{10.7}$$

holds for all $n \geq 1$ and all $x > 0$. To this end, let $0 \leq r < 1$ with $x = \lfloor x \rfloor + r$. Since the strict inequality $nx = n\lfloor x \rfloor + nr < n\lfloor x \rfloor + n$ holds true and the right hand side is an integer, we find $\lfloor nx \rfloor \leq n\lfloor x \rfloor + n - 1$ which is equivalent to (10.7). Now, to the implication (i)⇒(ii). Let $c > 0$ be the doubling constant of σ, i.e. $\sigma_{2n} \geq c\sigma_n$ for all $n \geq 1$. Using the monotonicity of σ, the inequality in (10.7), and (i) we find

$$f(2x) = \sigma_{\lfloor 2x \rfloor + 1} \geq \sigma_{2(\lfloor x \rfloor + 1)} \geq c\sigma_{\lfloor x \rfloor + 1} = cf(x) \ . \tag{10.8}$$

Finally, for fixed $\lambda \geq 1$ we choose an $m \geq 1$ with $2^m \geq \lambda$. The monotonicity of f and an m-fold application of (10.8) yields (ii). The case $0 < \lambda < 1$ can be easily deduced from the case $\lambda > 1$.

(iii)⇒(iv) Because of the monotonicity of σ we always have $(\sigma_1 \cdot \ldots \cdot \sigma_n)^{1/n} \geq \sigma_n$. For $c := \inf_{k \leq n} \frac{\sigma_n n^\alpha}{\sigma_k k^\alpha} > 0$ we have $\sigma_k \leq c^{-1}\sigma_n n^\alpha k^{-\alpha}$ for all $k \leq n$. Since Stirling's formula yields $(n!)^{1/n} \asymp n$, we get

$$(\sigma_1 \cdot \ldots \cdot \sigma_n)^{1/n} \leq c^{-1}\sigma_n \frac{n^\alpha}{(n!)^{\alpha/n}} \asymp \sigma_n \ .$$

(iv)⇒(i) Let $c > 0$ with $\sigma_n \leq (\sigma_1 \cdot \ldots \cdot \sigma_n)^{1/n} \leq c\sigma_n$ for all $n \geq 1$. Then

$$c\sigma_{2n} \geq (\sigma_1 \cdot \ldots \cdot \sigma_{2n})^{\frac{1}{2n}} = (\sigma_1 \cdot \ldots \cdot \sigma_n)^{\frac{1}{2n}} (\sigma_{n+1} \cdot \ldots \cdot \sigma_{2n})^{\frac{1}{2n}} \geq \sqrt{\sigma_n \sigma_{2n}} \ .$$

is satisfied for all $n \geq 1$. Hence we have $c^2\sigma_{2n} \geq \sigma_n \geq \sigma_{2n}$ for all $n \geq 1$. □

The final lemma provides some characterizations for various conditions on the r-tail sequence.

10.2.3 Lemma (Tail Sequence) *Let $r > 0$, $\sigma = (\sigma_k)_{k \geq 1}$ with $\sigma_k > 0$ for $k \geq 1$ and $\sigma_k \searrow 0$ for $k \to \infty$ and $\tau = (\tau_n)_{n \geq 1}$ be the r-tail sequence given by* (10.5). *Then the following statements hold true:*

(i) The following statements are equivalent:

(a) $\sup_{k\le n}\frac{\sigma_n n^\alpha}{\sigma_k k^\alpha}<\infty$ *for some* $\alpha>1/r$.

(b) σ *satisfies (ALP).*

(ii) The following statements are equivalent:

(c) $\sigma\in\ell_r$ *and* $\tau_n\asymp\tau_{2n}$.

(d) σ *satisfies (AMP).*

(iii) If $\sigma\in\ell_r$ *and* $\sigma_n\asymp\sigma_{2n}$ *are satisfied then the* r*-tail sequence satisfies* $\tau_n\asymp\tau_{2n}$.

(iv) If (a) is satisfied then $\sigma_n\asymp\sigma_{2n}$ *is satisfied if and only if* $\tau_n\asymp\tau_{2n}$ *is satisfied.*

If we combine Point (ii) with Lemma 10.2.2, we see that Condition (AMP) implies $\tau_n\succcurlyeq n^{-\alpha}$ for some $\alpha>0$.

Proof. (a)⇒(b) For $c:=\sup_{k\le n}\frac{\sigma_n n^\alpha}{\sigma_k k^\alpha}<\infty$ we get

$$\frac{\tau_k^r}{k\sigma_k^r}=\frac{1}{k}\sum_{n=k}^{\infty}\left(\frac{\sigma_n}{\sigma_k}\right)^r\le c^r k^{\alpha r-1}\sum_{n=k}^{\infty}n^{-\alpha r}$$

for all $k\ge1$. Estimating the remaining sum using integrals we get the assertion, namely

$$k^{\alpha r-1}\sum_{n=k}^{\infty}n^{-\alpha r}\le k^{\alpha r-1}\left(k^{-\alpha r}+\int_k^\infty t^{-\alpha r}\,\mathrm{d}t\right)\le\frac{\alpha r}{\alpha r-1}\quad.$$

(b)⇒(a) This is a consequence of Bingham et al. [8, Theorem 2.6.3] to the positive and measurable function $f(x):=x\sigma^r_{\lfloor x\rfloor}$ for $x\ge1$. To this end, we recall the definition of *almost decreasing* functions from [8, Section 2.2.1] and the *Matuszewska index* $\alpha(f)$ of f, defined in [8, Section 2.1.2]. Moreover, we have

$$\alpha(f)=\inf\big\{\alpha\in\mathbb{R}:\ x^{-\alpha}f(x)\text{ is almost decreasing}\big\}$$

according to [8, Theorem 2.2.2]. Since $x^{-1}f(x)$ is decreasing, we have $\alpha(f) \leq 1 < \infty$ and hence f is of *bounded increase*, i.e. $f \in \mathrm{BI}$, see [8, p. 71] for a definition. Consequently, [8, Theorem 2.6.3 (d)] is applicable to the function f. For $\tilde{f}(x) := \int_x^\infty f(t)/t \,\mathrm{d}t$ we have

$$\frac{f(x)}{\tilde{f}(x)} = \frac{x\sigma^r_{\lfloor x\rfloor}}{\tau^r_{\lfloor x\rfloor} - (x-\lfloor x\rfloor)\sigma^r_{\lfloor x\rfloor}} \geq \frac{x\sigma^r_{\lfloor x\rfloor}}{\tau^r_{\lfloor x\rfloor}} \geq \frac{\lfloor x\rfloor\sigma^r_{\lfloor x\rfloor}}{\tau^r_{\lfloor x\rfloor}} \geq c^{-r}$$

for all $x \geq 1$, where $c > 0$ is a constant satisfying $\tau_n \leq c\sigma_n n^{1/r}$ for all $n \geq 1$. Therefore, $\liminf_{x\to\infty} f(x)/\tilde{f}(x) > 0$ and [8, Theorem 2.6.3 (d)] yields $\alpha(f) < 0$. Consequently, there is a $\alpha_0 < 0$ such that $x^{-\alpha_0}f(x)$ is almost decreasing. The definition of almost decreasing gives us the assertion with $\alpha = \frac{1-\alpha_0}{r} > 1/r$.

(c)⇒(d) This is from [58, first equation on p. 45]. (d)⇒(c) The following idea is from [13, proof of Theorem 4]. According to our assumption the sequence

$$\rho_n := n\left(1 - \frac{\tau^r_{n+1}}{\tau^r_n}\right) = n\frac{\tau^r_n - \tau^r_{n+1}}{\tau^r_n} = \frac{n\sigma^r_n}{\tau^r_n}$$

is positive and bounded. Building a telescope product we get

$$\frac{\tau^r_n}{\tau^r_1} = \prod_{k=1}^{n-1}\frac{\tau^r_{k+1}}{\tau^r_k} = \prod_{k=1}^{n-1}\left(1 - \frac{\rho_k}{k}\right) .$$

Since $0 < 1 - \frac{\rho_k}{k} < 1$ holds true, this gives $\tau^r_n = \exp\circ\log(\tau^r_n) = \exp\bigl(\gamma_n - \sum_{k=1}^{n-1}\rho_k/k\bigr)$ with

$$\gamma_n := \log\tau^r_1 + \sum_{k=1}^{n-1}\left(\log\left(1 - \frac{\rho_k}{k}\right) + \frac{\rho_k}{k}\right) .$$

Below we will prove that $(\gamma_n)_{n\geq 1}$ converges and hence the assertion is a consequence of this representation of τ^r_n according to [26, Theorem 2]. Now, to the convergence of $(\gamma_n)_{n\geq 1}$. Since $(\rho_k)_{k\geq 1}$ is bounded, the sequence $a_k := \rho_k/k$ is square summable. Without loss of generality we assume that there is a $0 < q < 1$ with $a_n < q$ for all $n \geq 1$. Using the Taylor series of the logarithm we get

$$\log(1-a_k)+a_k=-\sum_{\ell=1}^{\infty}\frac{a_k^\ell}{\ell}+a_k=-\sum_{\ell=2}^{\infty}\frac{a_k^\ell}{\ell}\quad .$$

Additionally, for $\ell \geq 2$, we have the estimate $\sum_{k=1}^{\infty} a_k^\ell \leq \|a\|_{\ell_2}^2 q^{\ell-2}$. Together we get the absolute convergence of the series

$$\sum_{k=1}^{\infty}\big|\log(1-a_k)+a_k\big|=\sum_{k=1}^{\infty}\sum_{\ell=2}^{\infty}\frac{a_k^\ell}{\ell}=\sum_{\ell=2}^{\infty}\frac{1}{\ell}\sum_{k=1}^{\infty}a_k^\ell\leq\frac{\|a\|_{\ell_2}^2}{q^2}\sum_{\ell=2}^{\infty}\frac{q^\ell}{\ell}<\infty\quad .$$

(iii) According to our assumption there is a constant $c>0$ with $\sigma_{2n}\geq c\sigma_n$ for all $n\geq 1$. Then the assertion follows by

$$\tau_{2n}^r\geq\sum_{k=n}^{\infty}\sigma_{2k}^r\geq c^r\sum_{k=n}^{\infty}\sigma_k^r=c^r\tau_n^r\quad .$$

(iv) It is enough to prove the converse implication of Point (iii). Since we assume (a), we have (b) and (d), i.e. $\tau_n \asymp \sigma_n n^{1/r}$. Consequently, $\sigma_{2n}\asymp\tau_{2n}(2n)^{-1/r}\asymp\tau_n n^{-1/r}\asymp\sigma_n$ is satisfied. □

Chapter 11

Entropy Number Bounds

In this chapter, besides establishing some new upper bounds on the entropy numbers, we prove their optimality under some regularity conditions on the diagonal sequence. Finally, we compare our findings with the existing literature.

11.1 Upper Bounds

In this section we prove upper bounds on the entropy numbers of diagonal operators on sequence spaces. All these bounds are based on the following bound for diagonal operators between finite-dimensional spaces.

11.1.1 Lemma *Let $0 < p, q \leq \infty$, $k \geq 1$, and $\sigma_1, \ldots, \sigma_k > 0$. Then the diagonal operator $D_\sigma \colon \ell_p^k \to \ell_q^k$ satisfies, for $\varepsilon > 0$,*

$$\mathcal{P}(D_\sigma, \varepsilon) \leq (2\kappa_p)^k \, \frac{\lambda^k(B_{\ell_p^k})}{\lambda^k(B_{\ell_q^k})} \left(\| \operatorname{Id}_{q,p}^k \| + \kappa_q \frac{\sigma_1}{\varepsilon} \right) \cdot \ldots \cdot \left(\| \operatorname{Id}_{q,p}^k \| + \kappa_q \frac{\sigma_k}{\varepsilon} \right) ,$$

where $\operatorname{Id}_{q,p}^k \colon \ell_q^k \to \ell_p^k$ denotes the identity operator and λ^k the k-dimensional Lebesgue measure.

In the case $p = q$ this bound originates from Oloff [68, Hilfsatz 2]. Furthermore, note that the proof of Kolmogorov and Tikhomirov [51, Theorem XVI] contains the case $p = q = 2$ and $\sigma_n = n^{-\alpha}$.

Proof. For the first step of the proof we proceed analogously to the proof of Lemma C.7. To be more precise, for $\varepsilon > 0$ we choose an ε-packing $N \subseteq D_\sigma B_{\ell_p^k}$ with $n := |N| = \mathcal{P}(D_\sigma, \varepsilon)$. Then we get from (C.1) for $M = D_\sigma B_{\ell_p^k}$ and $B = B_{\ell_q^k}$

$$\begin{aligned}\biguplus_{x\in N} \big(x + \varepsilon/\kappa_q \cdot B_{\ell_q^k}\big) &\subseteq D_\sigma B_{\ell_p^k} + \varepsilon/\kappa_q \cdot B_{\ell_q^k} \\ &\subseteq D_\sigma B_{\ell_p^k} + \varepsilon/\kappa_q \cdot \|\operatorname{Id}_{q,p}^k\| \cdot B_{\ell_p^k} \quad ,\end{aligned} \tag{11.1}$$

where we used the definition of the operator norm in the last step.

Before we continue to estimate (11.1) we prove the following auxiliary result: For a second diagonal operator $D_\omega \colon \ell_p^k \to \ell_q^k$ with $\omega_i > 0$ for all $i = 1, \ldots, k$ we have

$$D_\sigma B_{\ell_p^k} + D_\omega B_{\ell_p^k} \subseteq 2\kappa_p D_{\sigma+\omega} B_{\ell_p^k} \quad . \tag{11.2}$$

Since $D_{\sigma+\omega}$ is invertible, (11.2) is equivalent to $D_{\sigma+\omega}^{-1}(D_\sigma B_{\ell_p^k} + D_\omega B_{\ell_p^k}) \subseteq 2\kappa_p B_{\ell_p^k}$. Now, to show (11.2) we fix $x, y \in B_{\ell_p^k}$ and observe

$$\begin{aligned}\|D_{\sigma+\omega}^{-1}(D_\sigma x + D_\omega y)\|_{\ell_p^k} &\le \kappa_p \|D_{\sigma+\omega}^{-1} D_\sigma x\|_{\ell_p^k} + \kappa_p \|D_{\sigma+\omega}^{-1} D_\omega y\|_{\ell_p^k} \\ &\le \kappa_p \|D_{\sigma+\omega}^{-1} D_\sigma\| + \kappa_p \|D_{\sigma+\omega}^{-1} D_\omega\| \quad .\end{aligned}$$

Since $D_{\sigma+\omega}^{-1} D_\sigma$ is an operator from ℓ_p^k to ℓ_p^k, the operator norm is given by $\|D_{\sigma+\omega}^{-1} D_\sigma\| = \max_{i=1,\ldots,k} \frac{\sigma_i}{\sigma_i+\omega_i} \le 1$. Analogously we have $\|D_{\sigma+\omega}^{-1} D_\omega\| = \max_{i=1,\ldots,k} \frac{\omega_i}{\sigma_i+\omega_i} \le 1$ and therefore (11.2) is proven.

Combining (11.1) with (11.2) and applying the Lebesgue measure gives

$$\begin{aligned}n(\varepsilon/\kappa_q)^k \lambda^k(B_{\ell_q^k}) &= \lambda^k\Big(\biguplus_{x\in N} \big(x + \varepsilon/\kappa_q\, B_{\ell_q^k}\big)\Big) \\ &\le \lambda^k\big(2\kappa_p D_{\sigma + \|\operatorname{Id}_{q,p}^k\| \cdot \varepsilon/\kappa_q} B_{\ell_p^k}\big) \\ &= (2\kappa_p)^k \cdot \lambda^k(B_{\ell_p^k}) \cdot \prod_{i=1}^k \big(\sigma_i + \|\operatorname{Id}_{q,p}^k\| \cdot \varepsilon/\kappa_q\big) \quad .\end{aligned}$$

Solving this inequality for n yields the assertion. □

In order to transfer the bound of Lemma 11.1.1 from finite- to infinite-dimensional diagonal operators we split the diagonal operator in to a finite-dimensional part and a remainder. To this end, we fix $0 < p, q \leq \infty$ and a sequence $\sigma = (\sigma_n)_{n\geq 1}$ such that the corresponding diagonal operator $D_\sigma \colon \ell_p \to \ell_q$ is well-defined and bounded. Using the auxiliary operators defined in (10.2) and properties of the covering numbers from Lemma C.9 we find, for $k \geq 1$,

$$\begin{aligned}\mathcal{N}\big(D_\sigma, \kappa_q\varepsilon\big) &= \mathcal{N}\Big(I_q^k D_{p,q}^k P_p^k + \big(D_\sigma - I_q^k D_{p,q}^k P_p^k\big), \kappa_q\varepsilon\Big)\\ &\leq \mathcal{N}\Big(I_q^k D_{p,q}^k P_p^k, \varepsilon/2\Big) \cdot \mathcal{N}\Big(D_\sigma - I_q^k D_{p,q}^k P_p^k, \varepsilon/2\Big)\\ &\leq \mathcal{N}\Big(D_{p,q}^k, \varepsilon/2\Big) \cdot \mathcal{N}\Big(D_\sigma - I_q^k D_{p,q}^k P_p^k, \varepsilon/2\Big) \ .\end{aligned}$$

In the following we will choose a suitable $k \geq 1$ with $\|D_\sigma - I_q^k D_{p,q}^k P_p^k\| \leq \varepsilon/2$. Since in this case we have $\mathcal{N}\big(D_\sigma - I_q^k D_{p,q}^k P_p^k, \varepsilon/2\big) = 1$, the estimate above reduces to

$$\mathcal{N}(D_\sigma, \kappa_q\varepsilon) \leq \mathcal{N}\big(D_{p,q}^k, \varepsilon/2\big) \ . \tag{11.3}$$

First, we apply (11.3) in the case $p < q$.

11.1.2 Lemma (Upper Bound for $p < q$) *Let $0 < p < q \leq \infty$ with $1/p = 1/q + 1/s$ and $\sigma = (\sigma_k)_{k\geq 1}$ with $\sigma_k > 0$ for $k \geq 1$ and $\sigma_k \searrow 0$ for $k \to \infty$. Then the diagonal operator $D_\sigma \colon \ell_p \to \ell_q$ satisfies, for $n \geq 1$,*

$$\begin{aligned}\varepsilon_n(D_\sigma) \leq 4\kappa_p\kappa_q \sup_{k\geq 1} \Bigg(\frac{\lambda^k(B_{\ell_p^k})}{\lambda^k(B_{\ell_q^k})} \cdot n^{-1} \cdot (2\kappa_q\sigma_1 + k^{1/s}\sigma_k) \cdot \ldots\\ \ldots \cdot (2\kappa_q\sigma_k + k^{1/s}\sigma_k)\Bigg)^{1/k} \ .\end{aligned}$$

Proof. For every $0 < \varepsilon/2 < \|D_\sigma\| = \sigma_1$, there is a $k \geq 1$ with $\sigma_{k+1} \leq \varepsilon/2 < \sigma_k$. Then (10.1) gives us $\big\|D_\sigma - I_q^k D_{p,q}^k P_p^k\big\| = \sigma_{k+1} \leq \varepsilon/2$. Using (11.3) with this k, Lemma 11.1.1, and $\|\operatorname{Id}_{q,p}^k\| = k^{1/s}$ we get

$$\mathcal{N}(D_\sigma, \kappa_q\varepsilon) \leq (2\kappa_p)^k \frac{\lambda^k(B_{\ell_p^k})}{\lambda^k(B_{\ell_q^k})} \Big(k^{1/s} + \frac{4\kappa_q\sigma_1}{\varepsilon}\Big) \cdot \ldots \cdot \Big(k^{1/s} + \frac{4\kappa_q\sigma_k}{\varepsilon}\Big) \ .$$

Using $k^{1/s} < 2\sigma_k k^{1/s}/\varepsilon$ and taking the supremum over $k \geq 1$ gives

$$\begin{aligned}\mathcal{N}(D_\sigma, \kappa_q\varepsilon) \leq \sup_{k\geq 1}\Bigg\{ & \frac{\lambda^k(B_{\ell_p^k})}{\lambda^k(B_{\ell_q^k})}\big(\sigma_k k^{1/s} + 2\kappa_q\sigma_1\big)\cdot \ldots \\ & \ldots \cdot \big(\sigma_k k^{1/s} + 2\kappa_q\sigma_k\big)\Big(\frac{4\kappa_p}{\varepsilon}\Big)^k\Bigg\} \ .\end{aligned}$$

Finally, Lemma 10.1.1 yields the assertion. □

Next, we apply (11.3) in the case $p > q$.

11.1.3 Lemma (Upper Bound for $p > q$) *Let $0 < q < p \leq \infty$ with $1/q = 1/p + 1/r$, $\sigma = (\sigma_k)_{k\geq 1} \in \ell_r$ with $\sigma_k > 0$ for $k \geq 1$ and $\sigma_k \searrow 0$ for $k \to \infty$, and τ the r-tail sequence defined by* (10.5). *Then the diagonal operator $D_\sigma\colon \ell_p \to \ell_q$ satisfies, for $n \geq 1$,*

$$\varepsilon_n(D_\sigma) \leq 4\kappa_p\kappa_q \sup_{k\geq 1}\bigg(\frac{(\tau_k + 2\kappa_p k^{1/r}\sigma_1)\cdot \ldots \cdot (\tau_k + 2\kappa_p k^{1/r}\sigma_k)}{n}\bigg)^{1/k} \ .$$

Proof. For every $0 < \varepsilon/2 < \|D_\sigma\| = \tau_1$, there is a $k \geq 1$ with $\tau_{k+1} \leq \varepsilon/2 < \tau_k$. Then (10.1) gives us $\|D_\sigma - I_q^k D_{p,q}^k P_p^k\| = \tau_{k+1} \leq \varepsilon/2$. Using (11.3) with this k, the decomposition $D_{p,q}^k = \mathrm{Id}_{p,q}^k \circ D_{p,p}^k$, and $\|\mathrm{Id}_{p,q}^k\| = k^{1/r}$ we get

$$\mathcal{N}(D_\sigma, \kappa_q\varepsilon) \leq \mathcal{N}(D_{p,p}^k, k^{-1/r}\varepsilon/2)\cdot \mathcal{N}(\mathrm{Id}_{p,q}^k, k^{1/r}) = \mathcal{N}(D_{p,p}^k, k^{-1/r}\varepsilon/2) \ .$$

Using Lemma 11.1.1 and $1 < 2\tau_k/\varepsilon$ gives

$$\begin{aligned}\mathcal{N}(D_\sigma, \kappa_q\varepsilon) &\leq (2\kappa_p)^k\Big(1 + \frac{4\kappa_p k^{1/r}\sigma_1}{\varepsilon}\Big)\cdot \ldots \cdot \Big(1 + \frac{4\kappa_p k^{1/r}\sigma_k}{\varepsilon}\Big) \\ &\leq \big(\tau_k + 2\kappa_p k^{1/r}\sigma_1\big)\cdot \ldots \cdot \big(\tau_k + 2\kappa_p k^{1/r}\sigma_k\big)\Big(\frac{4\kappa_p}{\varepsilon}\Big)^k \ .\end{aligned}$$

Finally, taking the supremum over k and using Lemma 10.1.1 gives the assertion. □

11.2 Optimality

In this section we investigate the optimality of the upper bounds of Section 11.1. To this end, we use the regularity conditions (EXP), (ALP), and (AMP) introduced in Section 10.2. First, we recall a well-known lower bound.

11.2.1 Lemma (Lower Bound) *Let $0 < p, q \leq \infty$ and $\sigma = (\sigma_k)_{k\geq 1}$ with $\sigma_k > 0$ for $k \geq 1$ and $\sigma_k \searrow 0$ for $k \to \infty$ such that the diagonal operator $D_\sigma \colon \ell_p \to \ell_q$ is bounded. Then for all $n \geq 1$ the n-th entropy number satisfies*

$$\varepsilon_n(D_\sigma) \geq \sup_{k\geq 1} \left(\frac{\lambda^k(B_{\ell_p^k})}{\lambda^k(B_{\ell_q^k})} \, \frac{\sigma_1 \cdot \ldots \cdot \sigma_k}{n} \right)^{1/k} .$$

Note that this lower bound holds without any additional assumption on σ.

Proof. Let $n, k \geq 1$ be fixed and choose $\varepsilon > \varepsilon_n(D_\sigma)$. Using Lemma C.1 we find $n \geq \mathcal{N}(D_\sigma, \varepsilon)$. Next, we use the auxiliary operators defined in (10.2) and the multiplicativity of the covering numbers from Lemma C.9 to get $\mathcal{N}(D^k_{p,q}, \varepsilon) = \mathcal{N}(P^k_q D_\sigma I^k_p, \varepsilon) \leq \mathcal{N}(D_\sigma, \varepsilon)$. An application of the lower bound in Lemma C.7 gives us

$$n \geq \mathcal{N}(D_\sigma, \varepsilon) \geq \mathcal{N}(D^k_{p,q}, \varepsilon) \geq \frac{\lambda^k(B_{\ell_p^k})}{\lambda^k(B_{\ell_q^k})} \cdot \sigma_1 \cdot \ldots \cdot \sigma_k \cdot \varepsilon^{-k} .$$

Rearranging this inequality for ε yields

$$\varepsilon \geq \left(\frac{\lambda^k(B_{\ell_p^k})}{\lambda^k(B_{\ell_q^k})} \, \frac{\sigma_1 \cdot \ldots \cdot \sigma_k}{n} \right)^{1/k} .$$

Letting $\varepsilon \searrow \varepsilon_n(D_\sigma)$ and taking the supremum over $k \geq 1$ we get the assertion. □

Using the lower bound of Lemma 11.2.1 we can prove the optimality of our upper bounds in Lemma 11.1.2 and Lemma 11.1.3. Let us start with the optimality of Lemma 11.1.2 if the Condition (EXP) is satisfied.

11.2.2 Theorem (Optimality for $p < q$) *Let $0 < p < q \leq \infty$ with $1/p = 1/q + 1/s$ and $\sigma = (\sigma_k)_{k\geq 1}$ be a sequence with $\sigma_k > 0$ for $k \geq 1$ and $\sigma_k \searrow 0$ for $k \to \infty$. If σ satisfies (EXP) then the upper bound in Lemma 11.1.2 is optimal. In this case the entropy numbers of the diagonal operator $D_\sigma \colon \ell_p \to \ell_q$ satisfy*

$$\varepsilon_n(D_\sigma) \asymp \sup_{k\geq 1} k^{-1/s} \left(\frac{\sigma_1 \cdot \ldots \cdot \sigma_k}{n} \right)^{1/k} .$$

Proof. Lemma 11.1.2 together with (10.3) gives us

$$\varepsilon_n(D_\sigma) \preccurlyeq \sup_{k\geq 1} k^{-1/s} \left(\frac{(\sigma_1 + k^{1/s}\sigma_k) \cdot \ldots \cdot (\sigma_k + k^{1/s}\sigma_k)}{n} \right)^{1/k} .$$

Rearranging the term in parentheses gives

$$\begin{aligned} &\left(\frac{(\sigma_1 + k^{1/s}\sigma_k) \cdot \ldots \cdot (\sigma_k + k^{1/s}\sigma_k)}{n} \right)^{1/k} \\ &= \left(\frac{\sigma_1 \cdot \ldots \cdot \sigma_k}{n} \right)^{1/k} \cdot \left(\left(1 + \frac{k^{1/s}\sigma_k}{\sigma_1}\right) \ldots \left(1 + \frac{k^{1/s}\sigma_k}{\sigma_k}\right) \right)^{1/k} . \end{aligned}$$

Applying that the geometric mean is bounded by the arithmetic mean as well as the triangle inequality in ℓ_s^k yields

$$\begin{aligned} \left(\left(1 + \frac{k^{1/s}\sigma_k}{\sigma_1}\right) \cdot \ldots \cdot \left(1 + \frac{k^{1/s}\sigma_k}{\sigma_k}\right) \right)^{1/k} &\leq \left(1/k \sum_{i=1}^{k} \left(1 + \frac{k^{1/s}\sigma_k}{\sigma_i}\right)^s \right)^{1/s} \\ &\leq \kappa_s + \kappa_s \sigma_k \left(\sum_{i=1}^{k} \sigma_i^{-s} \right)^{1/s} . \end{aligned}$$

Since σ satisfies (EXP), according Point (iii) of Lemma 10.2.1 we find $\sigma_k v_k \asymp 1$. As a result, the right hand side is bounded in k and we get "$\preccurlyeq$" for the claimed asymptotic equivalence. If we combine Lemma 11.2.1 with (10.3), we get the corresponding lower bound "$\succcurlyeq$". □

Next, we consider the optimality of Lemma 11.1.3.

11.2.3 Theorem (Optimality for $p > q$) *Let $0 < q < p \leq \infty$ with $1/q = 1/p + 1/r$ and $\sigma = (\sigma_k)_{k\geq 1} \in \ell_r$ be a sequence with $\sigma_k > 0$ for $k \geq 1$ and $\sigma_k \searrow 0$ for $k \to \infty$. If σ satisfies one of the following conditions then the upper bound in Lemma 11.1.3 is optimal:*

(i) Condition (ALP), i.e. $\tau_n \preccurlyeq \sigma_n n^{1/r}$. In this case the entropy numbers of the diagonal operator $D_\sigma \colon \ell_p \to \ell_q$ satisfy

$$\varepsilon_n(D_\sigma) \asymp \sup_{k\geq 1} k^{1/r} \Big(\frac{\sigma_1 \cdot \ldots \cdot \sigma_k}{n} \Big)^{1/k} \quad .$$

(ii) Condition (AMP), i.e. $\tau_n \succcurlyeq \sigma_n n^{1/r}$. In this case the entropy numbers of the diagonal operator $D_\sigma \colon \ell_p \to \ell_q$ satisfy

$$\varepsilon_n(D_\sigma) \asymp \tau_{\lfloor \log_2(n) \rfloor + 1} \quad .$$

Proof. (i) Lemma 11.1.3 gives us

$$\varepsilon_n(D_\sigma) \preccurlyeq \sup_{k\geq 1} \Big(\frac{(\tau_k + k^{1/r}\sigma_1) \cdot \ldots \cdot (\tau_k + k^{1/r}\sigma_k)}{n} \Big)^{1/k} \quad .$$

Rearranging the term in parentheses gives

$$\begin{aligned} &\Big(\frac{(\tau_k + k^{1/r}\sigma_1) \cdot \ldots \cdot (\tau_k + k^{1/r}\sigma_k)}{n} \Big)^{1/k} \\ &= \sup_{k\geq 1} k^{1/r} \Big(\frac{\sigma_1 \cdot \ldots \cdot \sigma_k}{n} \Big)^{1/k} \Big(\Big(\frac{\tau_k}{k^{1/r}\sigma_1} + 1 \Big) \cdot \ldots \cdot \Big(\frac{\tau_k}{k^{1/r}\sigma_k} + 1 \Big) \Big)^{1/k} \quad . \end{aligned}$$

According to (ALP) the last factor is bounded in k and hence we get "$\preccurlyeq$" for the claimed asymptotic equivalence. If we combine Lemma 11.2.1 with (10.3), we get the corresponding lower bound "$\succcurlyeq$".

(ii) Because of Point (ii) of Lemma 10.2.3 we have $\tau_n \asymp \tau_{2n}$. Hence Kühn [58, Theorem 1] yields $\varepsilon_n(D_\sigma) \asymp \tau_{\lfloor \log_2(n) \rfloor + 1}$ and it is enough to show that the upper bound in Lemma 11.1.3 is asymptotically bounded by $\tau_{\lfloor \log_2(n) \rfloor + 1}$. According to (AMP) and Point (iii) of Lemma 10.2.2 applied to $(\tau_n)_{n\geq 1}$

there are constants $c_1, c_2, \beta > 0$ with $\sigma_i \leq c_1 \tau_i i^{-1/r}$ and $\tau_i \leq c_2 \tau_k k^\beta i^{-\beta}$ for all $k \geq i$. Together we get, for $\alpha := 1/r + \beta$,

$$\tau_k + k^{1/r}\sigma_i \leq \tau_k + c_1 c_2 \tau_k \frac{k^{1/r+\beta}}{i^{1/r+\beta}} \leq \tau_k \frac{k^\alpha}{i^\alpha}(1 + c_1 c_2)$$

and all $k \geq i$. Plugging this into the bound of Lemma 11.1.3 we get

$$\varepsilon_n(D_\sigma) \preccurlyeq \sup_{k\geq 1}\left(\frac{(\tau_k + k^{1/r}\sigma_1)\cdot \ldots \cdot (\tau_k + k^{1/r}\sigma_k)}{n}\right)^{1/k} \preccurlyeq \sup_{k\geq 1} \frac{\tau_k}{n^{1/k}} \frac{k^\alpha}{(k!)^{\alpha/k}} \quad .$$

From Stirling's formula we know $(k!)^{1/k} \asymp k$ and hence we have

$$\varepsilon_n(D_\sigma) \preccurlyeq \sup_{k\geq 1} \frac{\tau_k}{n^{1/k}} \tag{11.4}$$

and it remains to show, that the right hand side behaves asymptotically like $\tau_{\lfloor \log_2(n)\rfloor+1}$. To this end, let $c > 0$ be the doubling constant of τ, i.e. $\tau_{2n} \geq c\tau_n$ for all $n \geq 1$. Without loss of generality we can assume $c < 1$ and define $\alpha := \frac{\log(2)}{2\log(1/c)} > 0$. For $k \leq \alpha \log_2(n)$ we have

$$n^{\frac{1}{2k}-\frac{1}{k}} = n^{-\frac{1}{2k}} \leq \exp\left(-\frac{\log(n)}{2\alpha\log_2(n)}\right) = c \leq \frac{\tau_{2k}}{\tau_k}$$

and this implies

$$\frac{\tau_k}{n^{\frac{1}{k}}} \leq \frac{\tau_{2k}}{n^{\frac{1}{2k}}} \quad . \tag{11.5}$$

A recursive application of this inequality enables us to restrict our supremum to $k > \alpha\log_2(n)$. Moreover, for such k we have

$$1 \geq n^{-1/k} = \exp\left(-\frac{\log(n)}{k}\right) \geq \exp\left(-\frac{\log(n)}{\alpha\log_2(n)}\right) = 2^{-1/\alpha} \quad . \tag{11.6}$$

If we combine (11.4), (11.5), and (11.6) then we get

$$\varepsilon_n(D_\sigma) \preccurlyeq \sup_{k\geq 1}\frac{\tau_k}{n^{1/k}} = \sup_{k>\alpha\log_2(n)} \frac{\tau_k}{n^{1/k}} \asymp \sup_{k>\alpha\log_2(n)} \tau_k = \tau_{\lfloor \alpha\log_2(n)\rfloor+1} \quad .$$

Finally, an application of Point (ii) of Lemma 10.2.2 yields the assertion. □

11.3 Comparison

In this section we compare our results with some bounds previously obtained in the literature. Since all previously established results on the entropy (or covering) numbers of diagonal operators, see e.g. [51, 64, 62, 68, 17, 56] and the references therein, are essentially contained in [57, 58, 18], we restrict our comparison to the latter three articles.

In the case $p < q$ the most general entropy bounds are derived by Kühn in [57]. Namely, he obtained optimal bounds under each of the following set of assumptions:

(i) *polynomial decay*: $\sup_{k \le n} \frac{\sigma_n n^\alpha}{\sigma_k k^\alpha} < \infty$ for some $\alpha > 0$ and $\sigma_n \asymp \sigma_{2n}$,

(ii) *fast logarithmic decay*: $\sup_{k \le n} \frac{\sigma_n}{\sigma_k} \big(\frac{1+\log n}{1+\log k}\big)^{1/s} < \infty$ and $\sigma_{n^2} \asymp \sigma_n$,

(iii) *slow logarithmic decay*: $\inf_{k \le n} \frac{\sigma_n}{\sigma_k} \big(\frac{1+\log n}{1+\log k}\big)^{1/s} > 0$.

Note that Scenario (i) and (ii) both exclude sequences that decrease too slow as well as sequences that decrease too fast. In contrast, (iii) only excludes sequences that decrease too fast. In comparison, our bound of Lemma 11.1.2 is optimal if the diagonal sequence decays at least exponentially in the sense of (EXP). Since all of the Scenarios (i)–(iii) imply $\sigma_n \asymp \sigma_{2n}$, we easily see that they all exclude (EXP), that is, (EXP) is not covered by [57].

In the case $p > q$, [57] provides optimal bounds for sequences σ satisfying

$$\sup_{k \le n} \frac{\sigma_n n^\alpha}{\sigma_k k^\alpha} < \infty$$

for some $\alpha > 1/r$ and $\sigma_n \asymp \sigma_{2n}$. According to Lemma 10.2.3 the combination of both assumptions is equivalent to the combination of (AMP) *and* (ALP), i.e. $\tau_n \asymp \sigma_n n^{1/r}$. In [58], Kühn generalizes the results of [57] by establishing optimal bounds under the Condition (AMP), only. Consequently, the upper bound of Lemma 11.1.3 is optimal for sequences considered in [58] and is additionally optimal for sequences satisfying (ALP) only.

Table 11.1 lists three types of sequences σ that are barely covered by the literature and our bounds bring new insights. Compared to [57, 58], another advantage of our results is that they actually provide bounds for *all* $p \neq q$

$\sigma_n \asymp$	$\tau_n \asymp$	(AMP)	(ALP)	(EXP)
$\exp\big(-a\log^{\lambda}(n)\big)$	$\sigma_n\, n^{1/r}\log^{(1-\lambda)/r}(n)$	no	yes if $\lambda > 1$	no
$\exp\big(-an^{\lambda}\big)$	$\sigma_n\, n^{(1-\lambda)_+/r}$	no	yes	yes if $\lambda \geq 1$
$\exp\big(-ae^{\lambda n}\big)$	σ_n	no	yes	yes

Table 11.1: Three types of diagonal sequences for which our bounds bring new insights into the asymptotic behavior the entropy numbers and which are hardly covered by the existing literature. For all examples we assume $a > 0$ and $\lambda > 0$. In addition, the conditions (AMP) and (ALP) are only considered in the case $p > q$, whereas (EXP) is actually independent of p and q. Note some subtleties of the first example: For $\lambda = 1$ it reduces to a plain polynomial decay, which is already well understood. Moreover, for $\lambda < 1$ the operator D_σ is not even bounded in the case $p > q$ and the asymptotic behavior can be found in [28, Example 14] in the case $p < q$.

and *all* sequences σ. However, in some cases the question of optimality is not answered yet.

There is another strand of research, see e.g. [17, 18], that describes the asymptotic behavior of the entropy numbers in terms of *(generalized) Lorentz spaces*. The most general result in this direction is [18, Corollary 1.2]:

$$\sigma \in \ell_{t,v,\varphi} \qquad \Longleftrightarrow \qquad \varepsilon_{2^{n-1}}(D_\sigma) \in \ell_{u,v,\varphi} \ ,$$

where $\ell_{u,v,\varphi}$ is a generalized Lorentz space with slowly varying function φ, see [18, Section 2] for a definition, and the parameters satisfy $1 \leq p, q \leq \infty$, $0 < t, v \leq \infty$, $1/t > (1/q - 1/p)_+$, and $1/u = 1/t - (1/q - 1/p)$. Note that the implication "$\Leftarrow$" is contained in Lemma 11.2.1 and "$\Rightarrow$" is contained in Lemma 11.1.3 in the case $p > q$ and $v = \infty$.

Finally, many results previously obtained in the literature are based on the operator ideal theory and a dyadic splitting of the diagonal operator, see e.g. [69] for an introduction to operator ideals and [17, 56, 18] for their application to entropy numbers of diagonal operators. This approach reduces the problem of bounding $\varepsilon_n(D_\sigma)$ to the estimation of entropy numbers of embeddings between finite-dimensional spaces. In order to bound the entropy numbers of these finite-dimensional embeddings advanced bounds with a good so-called *preasymptotic* behavior are needed. Such bounds can be found e.g. in [74, 30, 42, 55, 29] and are often based on sophisticated combinatoric arguments and interpolation theory. In contrast, our results are based on a single splitting of the diagonal operator and a simple bound for finite-dimensional diagonal operators. The latter bound has no good preasymptotic behavior but it is based on a plain volume argument.

Appendix

Appendix A

Poofs of the Oracle Inequalities

In this chapter we provide proofs of the oracle inequalities in Lemma 1.3.1 and Lemma 1.3.2. To this end, we use the fact that histograms are empirical risk minimizers (ERMs). To be more precise, for the (convex) output space $Y = [-M, M]$ with $M > 0$ and the hypothesis class

$$H(\mathcal{A}, Y) := \left\{ \sum_{k \in K} \mathbb{1}_{A_k} \cdot c_k : \ c_k \in Y \right\} \subseteq \mathcal{L}_0\big(X, \sigma(\mathcal{A})\big)$$

we get as direct consequences of (1.5), (1.18), and the projection property of the conditional expectation that $h_{D,\mathcal{A}} \in H(\mathcal{A}, Y)$ and

$$\mathcal{R}_{\mathrm{LS},D}(h_{D,\mathcal{A}}) = \mathcal{R}^*_{\mathrm{LS},D,H(\mathcal{A},Y)} = \inf_{f \in H(\mathcal{A},Y)} \mathcal{R}_{\mathrm{LS},D}(f) \tag{A.1}$$

are satisfied for all data sets $D \in (X \times Y)^n$. This means the histogram learning method is an ERM with respect to the LS loss and the hypothesis class $H(\mathcal{A}, Y)$. Moreover, since we use measurable and countable partitions $\mathcal{A}$ the set $\bigcup_{k \in K \setminus \mathcal{A}_\nu} A_k \subseteq X$ is a ν-zero set. As a result, the histogram $h_{D,\mathcal{A}}$ is even a LS ERM with respect to the possibly smaller hypothesis class $H(\mathcal{A}_\nu, Y)$ for P^n-almost all data sets $D \in (X \times Y)^n$. By abuse of notation we write $H(\mathcal{A}_\nu, Y)$ even if we mean $H\big((A_k)_{k \in \mathcal{A}_\nu}, Y\big)$. Note that (A.1) analogously remains true if we replace the empirical distribution D by P. Using these facts we are ready to prove the LS-risk oracle inequality.

Proof of Lemma 1.3.1. The proof is an application of a modified version of the general oracle inequality for ERMs in [76, Theorem 7.2], see also [66, Theorem E.2] for details. Since the histogram is P^n-almost surely an LS ERM over $H(\mathcal{A}_\nu, Y)$, the assumptions for this theorem are satisfied with $\vartheta = 1$, $B = 4M^2$, $V = 16M^2$, and $|L|_{M,1} = 4M$, see e.g. [76, Example 7.3]. Consequently, we find

$$\begin{aligned}\mathcal{R}_{\mathrm{LS},P}(h_{D,\mathcal{A}}) - \mathcal{R}^*_{\mathrm{LS},P} < 4 \cdot \big(&\mathcal{R}^*_{\mathrm{LS},P,H(\mathcal{A}_\nu,Y)} - \mathcal{R}^*_{\mathrm{LS},P}\big)\\ &+ 20 \cdot M\varepsilon\\ &+ 512 \cdot M^2 \cdot \frac{\tau + 1 + \log\big(\mathcal{N}(H(\mathcal{A}_\nu, Y), \varepsilon)\big)}{n}\end{aligned}$$

with probability P^n not less than $1 - e^{-\tau}$, where $\mathcal{N}\big(H(\mathcal{A}_\nu, Y), \varepsilon\big)$ denotes the covering numbers of $H(\mathcal{A}_\nu, Y)$ as subset of the set of bounded functions $\ell_\infty(X)$. For a definition of covering numbers and basic properties see e.g. Appendix C. Using $\mathcal{R}^*_{\mathrm{LS},P,H(\mathcal{A}_\nu,Y)} = \mathcal{R}_{\mathrm{LS},P}(h_{P,\mathcal{A}})$ from (A.1) (with P instead of D) and (1.5) we get

$$\mathcal{R}^*_{\mathrm{LS},P,H(\mathcal{A}_\nu,Y)} - \mathcal{R}^*_{\mathrm{LS},P} = \mathcal{R}_{\mathrm{LS},P}(h_{P,\mathcal{A}}) - \mathcal{R}^*_{\mathrm{LS},P} = \|h_{P,\mathcal{A}} - f^*_{\mathrm{LS},P}\|^2_{L_2(\nu)} \ .$$

Next, the hypothesis class $H(\mathcal{A}_\nu, Y)$ is isometrically isomorphic to the set $MB_{\ell_\infty^{|\mathcal{A}_\nu|}} \subseteq \ell_\infty^{|\mathcal{A}_\nu|}$ via $\sum_{k\in\mathcal{A}_\nu} c_k \cdot \mathbb{1}_{A_k} \mapsto (c_k)_{k\in\mathcal{A}_\nu}$ and hence Lemma C.7 and Lemma C.3 yield

$$\mathcal{N}\big(H(\mathcal{A}_\nu, Y), \varepsilon\big) = \mathcal{N}\big(MB_{\ell_\infty^{|\mathcal{A}_\nu|}}, \varepsilon\big) \le \big(1 + 2M/\varepsilon\big)^{|\mathcal{A}_\nu|} \le (3M/\varepsilon)^{|\mathcal{A}_\nu|} \ ,$$

where we used $\varepsilon \le M$ in the last step. Using $\tau \ge 1$, $\log(3M/\varepsilon) \ge 1$, and $|\mathcal{A}_\nu| \ge 1$ we find

$$\tau + 1 + \log\big(\mathcal{N}(H(\mathcal{A}_\nu, Y), \varepsilon)\big) \le 3\tau \log(3M/\varepsilon)|\mathcal{A}_\nu| \ .$$

Finally, combining all pieces we find the assertion. □

For the proof of the classification-risk oracle inequality we use an ERM property of the histogram again. To be more precise, for the classification loss it is well-known that the sign of the histogram sgn $\circ\, h_{D,\mathcal{A}}$ is an ERM with respect to the hypothesis class $H(\mathcal{A}, \{\pm 1\})$ and the classification loss.

Analogously to the LS loss case, $\operatorname{sgn} \circ h_{D,\mathcal{A}}$ is a classification loss ERM over the potentially smaller hypothesis class $H(\mathcal{A}_\nu, \{\pm 1\})$ for P^n-almost all data sets $D \in (X \times Y)^n$ and the population version $\operatorname{sgn} \circ h_{P,\mathcal{A}}$ minimizes $\mathcal{R}_{\mathrm{Class},P}$ over $H(\mathcal{A}, \{\pm 1\})$. Finally, note that the definition of the classification loss yields $\mathcal{R}_{\mathrm{Class},P}(f) = \mathcal{R}_{\mathrm{Class},P}(\operatorname{sgn} \circ f)$. Using these facts we are ready to prove the classification-risk oracle inequality.

Proof of Lemma 1.3.2. The proof is an application of the general oracle inequality for ERMs in [76, Theorem 7.2]. Since (the sign of) the histogram is P^n-almost surely a classification loss ERM over $H(\mathcal{A}_\nu, \{\pm 1\})$ and [10, Theorem 2.41] gives the *variance bound*

$$\mathbb{E}_P\big(L \circ f - L \circ f^*_{\mathrm{Class},P}\big)^2 \le c\big(\mathbb{E}_P L \circ f - L \circ f^*_{\mathrm{Class},P}\big)^{\frac{q}{q+1}}$$

for all $f \in \mathcal{L}_0(X, \mathcal{B})$ with some constant $c > 0$ depending on q and c_N, the assumptions of [76, Theorem 7.2] are satisfied with $\vartheta = q/(q+1)$, $B = 1$, and $V = c$. Consequently, we find

$$\begin{aligned}
&\mathcal{R}_{\mathrm{Class},P}(h_{D,\mathcal{A}}) - \mathcal{R}^*_{\mathrm{Class},P} \\
&\le 6 \cdot \big(\mathcal{R}^*_{\mathrm{Class},P,H(\mathcal{A}_\nu,\{\pm 1\})} - \mathcal{R}^*_{\mathrm{Class},P}\big) \\
&\quad + 4 \cdot \left(\frac{8V\big(\tau + \log\big(1 + \big|H(\mathcal{A}_\nu, \{\pm 1\})\big|\big)\big)}{n}\right)^{\frac{q+1}{q+2}}
\end{aligned} \tag{A.2}$$

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

Next, we consider the approximation error, i.e. the first summand in (A.2). Since $h_{P,\mathcal{A}}$ minimizes the classification loss over $H(\mathcal{A}_\nu, \{\pm 1\})$ and $2\eta - 1 = 0$ on $(X_+ \cup X_-)^c$, we can rewrite the approximation error using (1.10)

$$\begin{aligned}
&\mathcal{R}^*_{\mathrm{Class},P,H(\mathcal{A}_\nu,\{\pm 1\})} - \mathcal{R}^*_{\mathrm{Class},P} \\
&= \int_{X_+ \triangle \{h_{P,\mathcal{A}} \ge 0\}} |2\eta - 1| \, \mathrm{d}\nu \\
&= \sum_{k \ge 1} \int_X \mathbb{1}_{A_k \cap (X_+ \cup X_-) \cap (X_+ \triangle \{h_{P,\mathcal{A}} \ge 0\})} |2\eta - 1| \, \mathrm{d}\nu \ .
\end{aligned} \tag{A.3}$$

In order to proceed we fix $k \geq 1$ and distinguish four different cases. First, if $\nu\big(A_k \cap (X_+ \cup X_-)\big) = \nu(A_k \cap X_+) + \nu(A_k \cap X_-) = 0$ then the corresponding summand obviously vanishes. Second, assume that $\nu(A_k \cap X_+) = 0$ and $\nu(A_k \cap X_-) > 0$ hold true. Since $\nu(A_k \cap X_+) = 0$ holds true, the summand is equal to

$$\int_X \mathbb{1}_{A_k \cap X_- \cap (X_+ \triangle \{h_{P,\mathcal{A}} \geq 0\})} |2\eta - 1| \,\mathrm{d}\nu$$

We know that $h_{P,\mathcal{A}} = c_k$ is constant on A_k with $c_k := \int_{A_k} f^*_{\mathrm{LS},P} \,\mathrm{d}\nu / \nu(A_k)$. Since $\nu(A_k \cap X_+) = 0$, $\nu(A_k \cap X_-) > 0$, and $f^*_{\mathrm{LS},P} = 2\eta - 1 < 0$ on X_- hold true, we find

$$c_k = \frac{1}{\nu(A_k)} \int_{A_k} f^*_{\mathrm{LS},P} \,\mathrm{d}\nu = \frac{1}{\nu(A_k)} \int_{A_k \cap X_-} f^*_{\mathrm{LS},P} \,\mathrm{d}\nu < 0 \ .$$

As a result, we have $A_k \cap \{h_{P,\mathcal{A}} \geq 0\} = \emptyset$ and we get

$$\begin{aligned} & A_k \cap X_- \cap \big(X_+ \triangle \{h_{P,\mathcal{A}} \geq 0\}\big) \\ = & \big(A_k \cap X_- \cap X_+\big) \triangle \big(A_k \cap X_- \cap \{h_{P,\mathcal{A}} \geq 0\}\big) \\ = & \big(A_k \cap \emptyset\big) \triangle \big(X_- \cap \emptyset\big) \\ = & \emptyset \triangle \emptyset = \emptyset \ , \end{aligned}$$

where we used that the intersection $\cap$ is distributive over the symmetric difference $\triangle$. Consequently, the corresponding summand vanishes. Third, if $\nu(A_k \cap X_+) > 0$ and $\nu(A_k \cap X_-) = 0$ an analogous argument as in the second case applies and hence the corresponding summand vanishes. Forth, if $\nu(A_k \cap X_+) > 0$ and $\nu(A_k \cap X_-) > 0$ the corresponding summand eventually does not vanish. In this case we show $A_k \subseteq \{\Delta_d \leq \mathrm{diam}(\mathcal{A})\}$. To this end, note that we have $A_k \cap X_+ \neq \emptyset$ and $A_k \cap X_- \neq \emptyset$ and hence there are $x_+ \in A_k \cap X_+$ and $x_- \in A_k \cap X_-$. Then, for an arbitrary point $x \in A_k$ with $x \in X_+$ we have

$$\Delta_d(x) = \mathrm{dist}(x, X_-) \leq d(x, x_-) \leq \mathrm{diam}(A_k) \leq \mathrm{diam}(\mathcal{A}) \ .$$

In the case $x \in X_-$ an analogous argument yields $\Delta_d(x) \leq \mathrm{diam}(\mathcal{A})$.

Since $\Delta_d(x) = 0$ for all remaining points, we find $A_k \subseteq \big\{\Delta_d \leq \operatorname{diam}(\mathcal{A})\big\}$. Combining this with (A.3) we find the approximation error bound

$$\begin{aligned}\mathcal{R}^*_{\text{Class},P,H(\mathcal{A}_\nu,\{\pm 1\})} - \mathcal{R}^*_{\text{Class},P} &\leq \int_{\{\Delta_d \leq \operatorname{diam}(\mathcal{A})\}} |2\eta - 1| \,\mathrm{d}\nu \\ &\leq MN_d(r) \ .\end{aligned} \tag{A.4}$$

Here we used the assumption $\mathcal{B} \supseteq \mathcal{B}(X,d)$ which ensures the measurability of $\{\Delta_d \leq \operatorname{diam}(\mathcal{A})\}$.

Finally, we consider the estimation error, i.e. the second summand in (A.2). Since $|H(\mathcal{A}_\nu, \{\pm 1\})| = 2^{|\mathcal{A}_\nu|}$ and $\tau, |\mathcal{A}_\nu| \geq 1$ hold true, we find

$$\tau + \log\big(1 + \big|H(\mathcal{A}_\nu, \{\pm 1\})\big|\big) \leq \tau + (1 + |\mathcal{A}_\nu|)\log(2) \leq \tau|\mathcal{A}_\nu|\log(4e)$$

and hence combining this with (A.2) and (A.4) proves the assertion with $C := 4 \cdot (8c)^{\frac{q+1}{q+2}}$. □

Appendix B

Support of Measures

In this chapter we summarize some basic properties of the support of a measure. Recall, for a measure ν on the Borel σ-algebra $\mathcal{B}(X)$ of a topological Hausdorff space X the support $\operatorname{supp}\nu$ of ν is defined in (1.11) by

$$\operatorname{supp}\nu = \Bigg(\bigcup_{\substack{O\subseteq X \text{ open} \\ \nu(O)=0}} O \Bigg)^{c} \quad .$$

Moreover, for a Radon measure ν (locally finite and inner regular) the support is a set of full measure, i.e. $\nu((\operatorname{supp}\nu)^c) = 0$. Since most of the results of this chapter are true for general Radon measures, we do not assume that ν is a probability measure.

As a direct consequence of the definition of the support, we get $\operatorname{supp}\mu \subseteq \operatorname{supp}\nu$ for every absolute continuous Radon measure $\mu \ll \nu$. The first lemma considers sums of measures.

B.1 Lemma (Sums of Measures) *Let X be a topological Hausdorff space and $\nu_\pm$ are Radon measures on the Borel σ-algebra $\mathcal{B}(X)$. Then the support of $\nu := \nu_+ + \nu_-$ satisfies*

$$\operatorname{supp}\nu = \operatorname{supp}\nu_+ \cup \operatorname{supp}\nu_- \quad .$$

Proof. Taking the complement of the claimed identity we see that this identity is equivalent to

$$\bigcup_{\substack{O\subseteq X \text{ open} \\ \nu(O)=0}} O = \bigcup_{\substack{O_+\subseteq X \text{ open} \\ \nu_+(O_+)=0}} O_+ \cap \bigcup_{\substack{O_-\subseteq X \text{ open} \\ \nu_-(O_-)=0}} O_- \ .$$

"$\subseteq$" Since $\nu_\pm \ll \nu$ is absolute continuous, every ν-zero set is a ν_+-zero set and a ν_--zero set. This gives the inclusion "$\subseteq$".

"$\supseteq$" Let $x \in (\operatorname{supp}\nu_+)^c \cap (\operatorname{supp}\nu_-)^c$ be fixed. Then there are $O_\pm \subseteq X$ open sets with $x \in O_\pm$ and $\nu_\pm(O_\pm) = 0$. Consequently, $O := O_+ \cap O_-$ is a open set with $x \in O$ and $\nu(O) \leq \nu_+(O_+) + \nu_-(O_-) = 0$. This gives $x \in O \subseteq (\operatorname{supp}\nu)^c$ and hence the inclusion "$\supseteq$" is proven. □

The following lemma provides a representation for the support of a push-forward measure.

B.2 Lemma (Push-Forward Measure) *Let X and $\bar{X}$ be topological Hausdorff spaces, $s\colon X \to \bar{X}$ be a continuous function, ν be a Radon measure on $\mathcal{B}(X)$, and $\bar{\nu} := \nu \circ s^{-1}$ the push-forward measure of ν under s. Then $\bar{\nu}$ is a Radon measure with support*

$$\operatorname{supp}\bar{\nu} = \overline{s(\operatorname{supp}\nu)} \ .$$

If, in addition, $\operatorname{supp}\nu$ *is compact then* $\operatorname{supp}\bar{\nu} = s(\operatorname{supp}\nu)$ *holds true.*

Proof. That $\bar{\nu}$ is a Radon measure is stated in [31, Aufgabe VIII.1.10], see also [11, Theorem 9.1.1 (i)] for the case of a finite Radon measure.

"$\subseteq$" For this inclusion we prove that $\overline{s(\operatorname{supp}\nu)}^c$ is an (open) $\bar{\nu}$-zero set. Using $s^{-1}(\bar{A}^c) = (s^{-1}(\bar{A}))^c$ and $s^{-1}(s(A)) \supseteq A$, which hold for all $\bar{A} \subseteq \bar{X}$ and $A \subseteq X$, we find

$$s^{-1}\big(\overline{s(\operatorname{supp}\nu)}^c\big) \subseteq s^{-1}\big(s(\operatorname{supp}\nu)^c\big) = \big(s^{-1} \circ s(\operatorname{supp}\nu)\big)^c \subseteq (\operatorname{supp}\nu)^c \ .$$

This inclusion gives

$$\bar{\nu}\big(\overline{s(\operatorname{supp}\nu)}^c\big) = \nu\big(s^{-1}\big(\overline{s(\operatorname{supp}\nu)}^c\big)\big) \leq \nu\big((\operatorname{supp}\nu)^c\big) = 0$$

and hence the inclusion "$\subseteq$" is proven.

"$\supseteq$" We prove this inclusion using the following auxiliary result, which we will prove below,

$$s^{-1}(\operatorname{supp}\bar{\nu}) \supseteq \operatorname{supp}\nu \ . \tag{B.1}$$

Together with $\bar{A} \supseteq s\big(s^{-1}(\bar{A})\big)$, which holds for all $\bar{A} \subseteq \bar{X}$, we find

$$\operatorname{supp}\bar{\nu} \supseteq s\big(s^{-1}(\operatorname{supp}\bar{\nu})\big) \supseteq s(\operatorname{supp}\nu) \ .$$

Since $\operatorname{supp}\bar{\nu}$ is closed we get the inclusion "$\supseteq$".

Finally, it remains to prove (B.1). To this end, we show that the set $(s^{-1}(\operatorname{supp}\bar{\nu}))^c$ is an open ν-zero set. Since $s\colon X \to \bar{X}$ is continuous, the set $s^{-1}(\operatorname{supp}\bar{\nu}) \subseteq X$ is closed and hence its complement is open. Using $(s^{-1}(\bar{A}))^c = s^{-1}(\bar{A}^c)$, which holds for all $\bar{A} \subseteq \bar{X}$, we find

$$\nu\big((s^{-1}(\operatorname{supp}\bar{\nu}))^c\big) = \nu\big(s^{-1}(\operatorname{supp}\bar{\nu}^c)\big) = \bar{\nu}(\operatorname{supp}\bar{\nu}^c) = 0$$

and hence the inclusion in (B.1) is proven.

Finally, if we additionally assume that $\operatorname{supp}\nu$ is compact then as the image of a compact set under a continuous function the set $s(\operatorname{supp}\nu) \subseteq \bar{X}$ is compact. Since $\bar{X}$ is a Hausdorff spaces, the set $s(\operatorname{supp}\nu)$ is closed, see [32, Theorem 3.1.12]. As a result, we find $\operatorname{supp}\bar{\nu} = \overline{s(\operatorname{supp}\nu)} = s(\operatorname{supp}\nu)$. □

The final lemma relates the support of a measure on a product space with the supports of its one-dimensional push-forward measures.

B.3 Lemma (Infinite Products of Measures) *Let $X = \prod_{i\geq 1} X_i$ be the product of Polish spaces X_i for $i \geq 1$, ν be a Radon measure on $\mathcal{B}(X)$ with one-dimensional push-forward measures $\nu_i := \nu \circ \pi_i^{-1}$ on X_i for $i \geq 1$. Then the support of ν satisfies*

$$\operatorname{supp}\nu \subseteq \prod_{i\geq 1} \operatorname{supp}\nu_i \ .$$

If, in addition, ν_i are probability measures for $i \geq 1$ and $\nu = \bigotimes_{i\geq 1}\nu_i$ then equality holds.

Proof. Note that X is a Polish space and the Borel σ-algebra coincides with the product σ-algebra $\mathcal{B}(X) = \bigotimes_{i\geq 1} \mathcal{B}(X_i)$. See Section 3.1 for more details on such spaces. We start with two preparatory remarks without proving them. First, for a base $\mathfrak{B}$ of the product topology the complement of the support of ν is given by

$$(\operatorname{supp} \nu)^c = \bigcup_{\substack{O\subseteq X \text{ open}\\ \nu(O)=0}} O = \bigcup_{\substack{O\in\mathfrak{B}\\ \nu(O)=0}} O \ . \tag{B.2}$$

We will use (B.2) for the base $\mathfrak{B} := \big\{\big(\prod_{i=1}^N O_i\big) \times \prod_{i>N} X_i : \ N \geq 1, \ O_i \subseteq X_i \text{ open}\big\}$. Second, the complement of the right hand side is given by

$$\Big(\prod_{i\geq 1} \operatorname{supp} \nu_i\Big)^c = \bigcup_{i\geq 1}\Big((\operatorname{supp} \nu_i)^c \times \prod_{j\neq i} X_j\Big) \ . \tag{B.3}$$

After this two remarks we start with the actual proof. Since the right hand side of (B.3) is open, it is measurable and we have

$$\begin{aligned}\nu\Big(\Big(\prod_{i\geq 1} \operatorname{supp} \nu_i\Big)^c\Big) &\leq \sum_{i\geq 1} \nu\Big((\operatorname{supp} \nu_i)^c \times \prod_{j\neq i} X_j\Big)\\ &= \sum_{i\geq 1} \nu_i\big((\operatorname{supp} \nu_i)^c\big) = 0 \ .\end{aligned}$$

As a result, $(\prod_{i\geq 1} \operatorname{supp} \nu_i)^c$ is an open ν-zero set and consequently we get $(\prod_{i\geq 1} \operatorname{supp} \nu_i)^c \subseteq (\operatorname{supp} \nu)^c$. This proves the claimed inclusion.

For the converse inclusion we additionally assume that $\nu = \bigotimes_{i\geq 1} \nu_i$ is a product of probability measures ν_i. Let $O \in \mathfrak{B}$, i.e. $O = \prod_{i=1}^N O_i \times \prod_{i>N}$ for some $N \geq 1$ and some open sets $O_i \subseteq X_i$ for all $i = 1, \ldots, N$ with $\nu(O) = 0$. Since

$$0 = \nu(O) = \prod_{i=1}^N \nu_i(O_i)$$

holds true, there is some $i \in \{1, \ldots, N\}$ with $\nu_i(O_i) = 0$ and hence $O_i \subseteq$

$(\operatorname{supp}\nu_i)^c$. Together with (B.3) we find

$$O \subseteq (\operatorname{supp}\nu_i)^c \times \prod_{j\neq i} X_j \subseteq \left(\prod_{i\geq 1} \operatorname{supp}\nu_i\right)^c \ .$$

Since this holds true for all $O \in \mathfrak{B}$ with $\nu(O) = 0$, we get $(\operatorname{supp}\nu)^c \subseteq (\prod_{i\geq 1}\operatorname{supp}\nu_i)^c$ from (B.2). This proves the inclusion "$\supseteq$" for product measures. □

Appendix C

Metric Entropy

In this chapter we give a brief introduction to some *metric entropy quantities* such as entropy, covering, and packing numbers. For a more comprehensive introduction see e.g. [51, 19] and the references therein.

These metric entropy quantities are important tools for quantifying the compactness of sets and operators with various applications in different fields of mathematics, e.g. functional analysis (see e.g. [53, 19, 30] for operator ideals and eigenvalue distribution of compact operators), approximation theory (see e.g. [82, 30, 83] for embeddings of Sobolev or Besov spaces), probability theory (see e.g. [54, 60] for small deviations of Gaussian processes and [85] for empirical process theory), and statistical learning theory (see e.g. Appendix A and [73, 43, 24, 76] for the capacity of hypothesis spaces).

In the following let $M \subseteq X$ be a (non-empty) subset of a quasi-metric space (X, d), i.e. we only have the quasi-triangle inequality

$$d(x,y) \leq \kappa_X \big(d(x,z) + d(z,y)\big)$$

with some constant $\kappa_X \geq 1$ independent of $x, y, z \in X$. In this case we call κ_X the *quasi-triangle constant*. Unfortunately, we cannot limit ourselves to metric spaces, as we consider the quasi-Banach spaces $\ell_p(I)$ for $0 < p < 1$ in Part III.

For $\varepsilon > 0$, we call a subset $N \subseteq X$ *(external)* ε*-net* of M if the closed balls with centers in N and radius ε cover M, i.e.

$$M \subseteq \bigcup_{x\in N} B_X(x,\varepsilon) \ .$$

Or equivalently, for every $x \in M$ there is some $x' \in N$ with $d(x,x') \leq \varepsilon$. For the sake of completeness, N is called *internal ε-net* of M if $N \subseteq M$ is additionally satisfied. However, we only use external ε-nets in the following and hence we just call them ε-nets. Moreover, for $\varepsilon > 0$, we call a subset $N \subseteq M$ *ε-packing* of M if

$$d(x,x') > 2\varepsilon$$

is satisfied for all $x, x' \in N$ with $x \neq x'$. This means that for two distinct points $x, x' \in N$ in an ε-packing the closed balls with radius ε/κ_X and centers x and x', respectively, do not intersect. Using ε-nets and ε-packings we define the *entropy numbers* of M by

$$\varepsilon_n(M) := \inf\bigl\{0 < \varepsilon \leq \infty : \ \exists \ \varepsilon\text{-net } N \text{ of } M \text{ with } |N| \leq n\bigr\}$$

for $n \geq 1$, the *covering numbers* of M by

$$\mathcal{N}(M,\varepsilon) := \min\bigl\{1 \leq n \leq \infty : \ \exists \ \varepsilon\text{-net } N \text{ of } M \text{ with } |N| \leq n\bigr\}$$

for $\varepsilon > 0$, and the *packing numbers* of M by

$$\mathcal{P}(M,\varepsilon) := \max\bigl\{1 \leq n \leq \infty : \ \exists \ \varepsilon\text{-packing } N \text{ of } M \text{ with } |N| \geq n\bigr\}$$

for $\varepsilon > 0$. Here we use the convention that every non-empty subset $N \subseteq X$ is an ∞-*net* of M and that $N = M$ is an 0-*net* of M with $|N| = |M| \in [1,\infty]$ and hence the considered sets in the definitions of entropy and covering numbers are always non-empty. Analogously, every singleton $N \subseteq M$ is an ∞-*packing* of M and hence the considered set in the definition of packing numbers is always non-empty.

Attention, the definitions of the metric entropy quantities vary in the literature and we always have to check the precise definition. In this regard, we emphasize that we always use closed balls and external ε-nets.

For the sake of completeness, we mention that in the literature the numbers $\varphi_n(M) \coloneqq \sup\bigl\{0 < \varepsilon \leq \infty :$ there is an ε-packing N of M with $|N| \geq n+1\bigr\}$ for $n \geq 1$ can be found, see e.g. [19, p. 7]. The $\varphi_n(M)$-numbers are in the same way related to the packing numbers as the entropy numbers to the covering numbers. Moreover, the only difference in the definition of the entropy and covering numbers is the following: For entropy numbers the cardinality of the ε-nets is bounded and we minimize over the radius ε and for covering numbers the radius ε is fixed an we minimize over the cardinality n of the ε-nets. The following lemma is a direct consequence of this observation.

C.1 Lemma (Entropy vs. Covering Numbers) *For a subset $M \subseteq X$ of a quasi-metric space (X, d) the following statements are true:*

(i) For $\varepsilon > 0$ the bound $\varepsilon_n(M) \leq \varepsilon$ is satisfied for all $n \geq \mathcal{N}(M, \varepsilon)$.

(ii) For $n \geq 1$ the bound $\mathcal{N}(M, \varepsilon) \leq n$ is satisfied for all $\varepsilon > \varepsilon_n(M)$.

This lemma implies $\mathcal{N}(M, \varepsilon) < \infty$ for all $\varepsilon > 0$ if and only if $\varepsilon_n(M) \to 0$ for $n \to \infty$.

Proof. This statement is a direct consequence of the definition of entropy and covering numbers. □

Recall, $M \subseteq X$ is called *precompact* (or *totally bounded*) if for every $\varepsilon > 0$ there is a finite ε-net of M. Consequently, Lemma C.1 yields the equivalence of the following statements:

(i) M is precompact.

(ii) $\mathcal{N}(M, \varepsilon) < \infty$ for all $\varepsilon > 0$.

(iii) $\varepsilon_n(M) \to 0$ for $n \to \infty$.

In this sense the metric entropy quantities are quantitative refinements of the notion of precompact sets.

Another application of Lemma C.1 allows to translate bounds on covering numbers to bounds on entropy numbers and vice versa.

C.2 Corollary (Conversion of Bounds) *Let $M \subseteq X$ be a subset of a quasi-metric space (X, d) and $f\colon [a, \infty) \to (0, b]$ be a decreasing bijective function with $a, b > 0$. Then the following statements are true:*

(i) If $\varepsilon_n(M) \le f(n)$ is satisfied for all $n \ge a$ then

$$\mathcal{N}(M, \varepsilon) \le f^{-1}(\varepsilon) + 1$$

is satisfied for all $0 < \varepsilon \le b$.

(ii) If $\mathcal{N}(M, \varepsilon) \le f^{-1}(\varepsilon)$ is satisfied for all $0 < \varepsilon \le b$ then

$$\varepsilon_n(M) \le f(n)$$

is satisfied for all $n \ge a$.

This lemma suggests that entropy and covering numbers are *inverse* concepts. Especially, a polynomial bound of order $1/p$ on the entropy numbers, i.e. $f(t) = Ct^{-1/p}$, gives a polynomial bound of order p on the covering numbers, i.e. $f^{-1}(t) = (C/t)^p$, and vice versa. Note that the constant C of the entropy number bound does not influence the asymptotic behavior of the covering number bound. This is in general not true in the non-polynomial regime.

Proof. (i) For $0 < \varepsilon \le b$ we choose the minimal integer $n \ge a$ with $f(n) < \varepsilon$. Since f is bijective and decreasing there is some $0 < \delta \le 1$ with $f(n - \delta) = \varepsilon$. Consequently, we have $n = f^{-1}(\varepsilon) + \delta \le f^{-1}(\varepsilon) + 1$, $\varepsilon_n(M) \le f(n) < \varepsilon$, and Point (ii) of Lemma C.1 gives the assertion, namely

$$\mathcal{N}(M, \varepsilon) \le n \le f^{-1}(\varepsilon) + 1 \quad .$$

(ii) For $n \ge a$ and $\varepsilon := f(n)$ we have $\mathcal{N}(M, \varepsilon) \le f^{-1}(\varepsilon) = n$. Consequently, Point (i) of Lemma C.1 gives the assertion $\varepsilon_n(M) \le \varepsilon = f(n)$. □

Our general strategy for proving upper and lower bounds is as follows: If we can construct an ε-net of $M \subseteq X$ with cardinality $n \ge 1$ then we directly get the upper bounds $\varepsilon_n(M) \le \varepsilon$ and $\mathcal{N}(M, \varepsilon) \le n$. Conversely, if we can construct an ε-packing of M with cardinality $n \ge 1$ then we get the lower bound $\mathcal{P}(M, \varepsilon) \geq n$. The following lemma already uses this strategy to relate covering and packing numbers.

C.3 Lemma (Covering vs. Packing Numbers) *For a subset $M \subseteq X$ of a quasi-metric space (X, d), with quasi-triangle constant $\kappa_X \geq 1$, the following inequalities are satisfied, for $\varepsilon > 0$,*

$$\mathcal{P}(M, \varepsilon\kappa_X) \leq \mathcal{N}(M, \varepsilon) \leq \mathcal{P}(M, \varepsilon/2) \ \ .$$

Proof. In order to prove the first inequality we assume the opposite, i.e. we assume that $m := \mathcal{P}(M, \kappa_M \varepsilon) > \mathcal{N}(M, \varepsilon) =: n$ holds true. The definition of the packing numbers gives us an $\varepsilon\kappa_X$-packing N of M with cardinality $|N| = m$ and the definition of the covering numbers gives us an ε-net N' of M with cardinality $|N'| = n$. Since $|N| = m > n = |N'|$ and

$$N \subseteq M \subseteq \bigcup_{y \in N'} B_X(y, \varepsilon)$$

hold true, there are $y \in N'$ and $x, x' \in N$ with $x \neq x'$ and $x, x' \in B_X(y, \varepsilon)$. For these elements we get the contradiction

$$2\kappa_X \varepsilon < d(x, x') \leq \kappa_X \big(d(x, y) + d(y, x')\big) \leq 2\kappa_X \varepsilon \ \ .$$

To prove the second inequality we choose a maximal $\varepsilon/2$-packing N of M with cardinality $m := |N| = \mathcal{P}(M, \varepsilon)$ and show that N is an ε-net of M. To this end, we assume the opposite, i.e. there is some $x \in M$ with $d(x, x') > \varepsilon$ for all $x' \in N$. This means that $N \cup \{x\}$ is an $\varepsilon/2$-packing with cardinality $m + 1$. This contradicts the maximality of N and hence N is an ε-net of M. This proves the desired inequality. □

The next lemma provides basic properties of the metric entropy quantities.

C.4 Lemma (Basic Properties) *For a quasi-metric space (X, d), with quasi-triangle constant $\kappa_X \geq 1$, the following statements are true:*

(i) For fixed $M \subseteq X$ the sequence $(\varepsilon_n(M))_{n \geq 1}$ and the functions $\varepsilon \mapsto \mathcal{N}(M, \varepsilon)$, $\varepsilon \mapsto \mathcal{P}(M, \varepsilon)$ on $(0, \infty)$ are non-increasing.

(ii) For fixed $n \geq 1$ and $\varepsilon > 0$ the numbers $\varepsilon_n(M)$, $\mathcal{N}(M,\varepsilon)$, and $\mathcal{P}(M,\varepsilon)$ are non-decreasing in $M \subseteq X$.

(iii) For fixed $M \subseteq X$ the function $\varepsilon \mapsto \mathcal{P}(M,\varepsilon)$ is right-continuous.

(iv) $\varepsilon_n(M) = \varepsilon_n(\overline{M})$ and $\mathcal{N}(M,\varepsilon) = \mathcal{N}(\overline{M},\varepsilon)$ for $M \subseteq X$, $n \geq 1$ and $\varepsilon > 0$, respectively.

(v) $\mathcal{P}(M,\varepsilon) \leq \mathcal{P}(\overline{M},\varepsilon) \leq \mathcal{P}\big(M,\varepsilon/\min\{2\kappa_X,\kappa_X^2\}\big)$ for $M \subseteq X$ and $\varepsilon > 0$.

Point (iii)–(v) heavily depend on the fact that we use closed balls in the definitions of the metric entropy quantities. Note that in the case that X is even a metric space, i.e. $\kappa_X = 1$, Point (v) reduces to $\mathcal{P}(M,\varepsilon) = \mathcal{P}(\overline{M},\varepsilon)$.

Proof. (i) If $N \subseteq X$ is an ε-net of M then every $N' \supseteq N$ is an ε-net of M . This proves the monotonicity of $(\varepsilon_n(M))_{n\geq1}$. If $N \subseteq X$ is an ε-net of M then N is an ε'-net of M for every $\varepsilon' \geq \varepsilon$. This proves the monotonicity of $\mathcal{N}(M,\,\cdot\,)$. Finally, if $N \subseteq M$ is an ε-packing of M then N is an ε'-packing of M for every $\varepsilon' \leq \varepsilon$. This proves the monotonicity of $\mathcal{P}(M,\,\cdot\,)$.

(ii) If $N \subseteq X$ is an ε-net of M then N is an ε-net for every $M' \subseteq M$. This proves the monotonicity of $\varepsilon_n(\,\cdot\,)$ and $\mathcal{N}(\,\cdot\,,\varepsilon)$. Finally, if $N \subseteq M$ is an ε-packing of M then N is an ε-packing for every $M' \supseteq M$. This proves the monotonicity of $\mathcal{P}(\,\cdot\,,\varepsilon)$.

(iii) Let $\varepsilon > 0$ be fixed and choose an ε-packing $N \subseteq M$ with $n := |N| = \mathcal{P}(M,\varepsilon)$. Then, we define

$$\varepsilon_0 := \min_{\substack{x,x'\in N:\\ x\neq x'}} d(x,x')/2 - \varepsilon > 0 \ .$$

and show $\mathcal{P}(M,\delta) = n$ for all $\varepsilon < \delta < \varepsilon + \varepsilon_0$. Since $\delta > \varepsilon$ holds true, we get $\mathcal{P}(M,\delta) \leq n$ from Point (i). The choice of ε_0 ensures $2\delta < 2(\varepsilon+\varepsilon_0) = \min_{x,x'\in N:\ x\neq x'} d(x,x')$ and hence N is a δ-packing of M. This proves $\mathcal{P}(M,\delta) \geq n$.

(iv) Since we have $M \subseteq \overline{M}$, the inequality "$\leq$" follows from Point (ii) for entropy and covering numbers. In order to prove the inequality "$\geq$" we choose some finite ε-net $N \subseteq X$ of M. Note that in the case that there is no finite ε-net of M the entropy and covering numbers are infinite and there is nothing to prove. Using the fact that the closure interchanges with finite unions, see e.g. [32, Theorem 1.1.3], we get

$$\overline{M} \subseteq \overline{\bigcup_{x\in N} B_X(x,\varepsilon)} = \bigcup_{x\in N} \overline{B_X(x,\varepsilon)} = \bigcup_{x\in N} B_X(x,\varepsilon) \ .$$

As a result, N is an ε-net of $\overline{M}$ and hence the inequalities "$\geq$" are proven.

(v) Since we have $M \subseteq \overline{M}$, the first inequality follows from Point (ii). In order to prove the second inequality we consider the case $\kappa_X \leq 2$ and $\kappa_X \geq 2$ separately.

In the case $\kappa_X \leq 2$ we use Point (iii), i.e. there is some $\delta > 0$ with $n := \mathcal{P}(\overline{M},\varepsilon) = \mathcal{P}(\overline{M},\varepsilon+\delta)$. Consequently, we can choose an $(\varepsilon+\delta)$-packing $N \subseteq \overline{M}$ with $|N| = n$. Using the definition of the closure, for every $y \in N$ there is some $x_y \in M$ with $d(x,y) < \delta/\kappa_X^2$. With this choice we get, for $y, y' \in N$ with $y \neq y'$,

$$\begin{aligned} 2(\varepsilon+\delta) &< d(y,y') \\ &\leq \kappa_X^2\bigl(d(y,x_y) + d(x_y,x_{y'}) + d(x_{y'},y')\bigr) \\ &\leq 2\delta + \kappa_X^2 d(x_i,x_j) \ . \end{aligned}$$

As a result, $N' := \{x_y : \ y \in N\} \subseteq M$ is an ε/κ_X^2-packing of M with $|N'| = |N| = n$ and hence the second inequality is proven in the case $\kappa_X \leq 2$.

In the case $\kappa_X \leq 2$ a two-fold application of Lemma C.3 together with Point (v) gives us

$$\mathcal{P}\bigl(\overline{M},\varepsilon\bigr) \leq \mathcal{N}\bigl(\overline{M},\varepsilon/\kappa_X\bigr) = \mathcal{N}\bigl(M,\varepsilon/\kappa_X\bigr) \leq \mathcal{P}\bigl(M,\varepsilon/(2\kappa_X)\bigr)$$

and hence the assertion is proven. □

Next, we consider images of Hölder continuous mappings.

C.5 Lemma (Hölder Continuous Images) *Let $0 < \alpha \leq 1$, $f\colon X \to X'$ be an α-Hölder continuous mapping between quasi-metric spaces (X,d) and*

(X', d') with constant L, i.e.

$$d'\big(f(x), f(y)\big) \leq L \cdot d^\alpha(x,y)$$

for $x, y \in X$, and $M \subseteq X$ be a subset. Then the following inequality is satisfied, for $\varepsilon > 0$,

$$\mathcal{N}\big(f(M), L\varepsilon^\alpha\big) \leq \mathcal{N}(M, \varepsilon) \ .$$

Proof. Let $\varepsilon > 0$ be fixed and choose an ε-net $N \subseteq X$ of M with $|N| = \mathcal{N}(M,\varepsilon)$. Then we show that $f(N)$ forms an $L\varepsilon^\alpha$-net of $f(M)$. For an element $x' = f(x) \in f(M)$ with $x \in M$ there is some $y \in N$ with $d(x,y) \leq \varepsilon$. Consequently, we have $y' \coloneqq f(y) \in f(N)$ and $d'(x',y') = d'\big(f(x), f(y)\big) \leq Ld^\alpha(x,y) \leq L\varepsilon^\alpha$. This shows that $f(N)$ is an $L\varepsilon^\alpha$-net of $f(M)$ with at most $|N| = \mathcal{N}(M,\varepsilon)$ elements and hence the assertion is proven. □

If X is additionally a vector space and the quasi-metric is introduced by a quasi-norm $\|\cdot\|$, we can consider the multiple $s \cdot M = \{sx \in X : \ x \in M\}$ of the set $M \subseteq X$ and the *Minkowski sum* $M + M' = \{x + x' \in X : \ x \in M, \ x' \in M'\}$ of two sets $M, M' \subseteq X$. Using this notation we can write the closed ball as $B_X(x,r) = x + r \cdot B_X$, where B_X denotes the cosed unit ball with center 0. The following lemma presents basic bounds for these constructions.

C.6 Lemma (Covering Numbers on Vector Spaces) *Let $(X, \|\cdot\|)$ be a quasi-normed vector space with quasi-triangle constant $\kappa_X \geq 1$ and $M \subseteq X$ be a subset. Then the following statements are true, for $\varepsilon > 0$:*

(i) $\mathcal{N}\big(M + M', \kappa_X(\varepsilon + \varepsilon')\big) \leq \mathcal{N}(M,\varepsilon)\mathcal{N}(M',\varepsilon')$ for $M' \subseteq X$ and $\varepsilon' > 0$.

(ii) $\mathcal{N}(m' + M, \varepsilon) = \mathcal{N}(M,\varepsilon)$ for $m' \in X$.

(iii) $\mathcal{N}(s \cdot M, \varepsilon) = \mathcal{N}(M, \varepsilon/|s|)$ for $s \in \mathbb{R}\backslash\{0\}$.

Proof. (i) Let $\varepsilon, \varepsilon' > 0$ be fixed and choose an ε-net $N \subseteq X$ of M with $n \coloneqq |N| = \mathcal{N}(M,\varepsilon)$ as well as an ε'-net $N' \subseteq X$ of M' with $n' \coloneqq |N'| = \mathcal{N}(M',\varepsilon')$. Then we show that $N + N'$ forms a $\kappa_X(\varepsilon + \varepsilon')$-net of $M + M'$. For an element $y + y' \in M + M'$ there is some $x \in N$ and some $x' \in N'$ with

$\|x - y\| \leq \varepsilon$ and $\|x' - y'\| \leq \varepsilon'$, respectively. Consequently, $x + x' \in N + N'$ and

$$\big\|(x + x') - (y + y')\big\| \leq \kappa_X\big(\|x - y\| + \|x' - y'\|\big) \leq \kappa_X(\varepsilon + \varepsilon') \ .$$

This shows that $N + N'$ is a $\kappa_X(\varepsilon + \varepsilon')$-net of $M + M'$ with at most $n \cdot n'$ elements and hence the statement is proven.

(ii)+(iii) Both statements follow by a two-fold application of Lemma C.5 with the Lipschitz continuous mappings $x \mapsto m' + x$ and $x \mapsto sx$, respectively, and their (Lipschitz continuous) inverses. □

For the vector space $X = \mathbb{R}^d$ so-called volume arguments are one way to bound the covering and packing numbers. To this end, we denote the d-dimensional Lebesgue measure by λ^d. The following lemma presents some elementary bounds based on volume arguments.

C.7 Lemma (Volume Arguments) *Let $X = \mathbb{R}^d$ be equipped with some quasi-norm $\|\cdot\|$, $\kappa_X \geq 1$ be the quasi-triangle constant, and B be the closed unit ball. Then the following statements are true, for $M \subseteq \mathbb{R}^d$ and $\varepsilon > 0$:*

(i) If M is measurable then the following lower bound is satisfied

$$\mathcal{N}(M, \varepsilon) \geq \frac{\lambda^d(M)}{\lambda^d(B)} \cdot (1/\varepsilon)^d \ .$$

(ii) If there is some $R > 0$ and $a \in \mathbb{R}^d$ with $M \subseteq a + RB$ then the following upper bound is satisfied

$$\mathcal{P}(M, \varepsilon) \leq \kappa_X^d \cdot \big(1 + R\kappa_X/\varepsilon\big)^d \ .$$

(iii) If M is convex and there is some $\varepsilon_0 > 0$ and $a \in \mathbb{R}^d$ with $a + \varepsilon_0 B \subseteq M$ then the following upper bound is satisfied

$$\mathcal{P}(M, \varepsilon) \leq \frac{\lambda^d(M)}{\lambda^d(B)} \cdot \big(1/\varepsilon_0 + \kappa_X/\varepsilon\big)^d \ .$$

Point (i) can be easily transferred from covering numbers to entropy numbers. But if we want to state Point (ii) and (iii) for entropy numbers instead of packing numbers then the inequalities appearing in the proof have to be solved for ε instead of n. But this is not an easy task because these inequalities are of the form $p(1/\varepsilon) \geq n$ with a polynomial p of degree d. As a workaround one typically further estimate these inequalities until it is easier to solve them for ε. But the price one has to pay is that this procedure results in weaker bounds for the entropy numbers than for the packing numbers. This is the reason why we present the upper bounds for packing numbers and not for entropy numbers.

For particular sets there are further volume arguments to derive packing number bounds, see e.g. Lemma 11.1.1 of Part III for generalized ellipses.

Proof. (i) For a fixed $\varepsilon > 0$ we choose an ε-net $N \subseteq \mathbb{R}^d$ with $n := |N| = \mathcal{N}(M, \varepsilon)$, i.e.

$$M \subseteq \bigcup_{x \in N} (x + \varepsilon B) \enspace .$$

Since both sides are measurable, we can apply the Lebesgue measure. Using the translation invariance of the Lebesgue measure we find

$$\lambda^d(M) \leq \sum_{x \in N} \lambda^d(x + \varepsilon B) = n\varepsilon^d \lambda^d(B) \enspace .$$

Since B has a non-empty interior, we have $\lambda^d(B) > 0$ and hence we can solve this inequality for n. This proves the statement.

(ii)+(iii) For both statements we fix some $\varepsilon > 0$ and choose an ε-packing $N \subseteq X$ of M with $n := |N| = \mathcal{P}(X, \varepsilon)$. Then the balls $x + \varepsilon/\kappa_X \cdot B$ for $x \in N$ are disjoint subsets of $M + \varepsilon/\kappa_X \cdot B$, i.e.

$$\biguplus_{x \in N} \big(x + \varepsilon/\kappa_X \cdot B\big) \subseteq M + \varepsilon/\kappa_X \cdot B \enspace . \tag{C.1}$$

Under the assumptions of Point (ii) we can continue the inclusion of (C.1) as follows

$$M + \varepsilon/\kappa_X \cdot B \subseteq a + R \cdot B + \varepsilon/\kappa_X \cdot B \subseteq a + \kappa_X \cdot (R + \varepsilon/\kappa_X) \cdot B \enspace .$$

Since the right hand side is measurable, we can apply the Lebesgue measure

$$\begin{aligned} n \cdot (\varepsilon/\kappa_X)^d \cdot \lambda^d(B) &= \lambda^d\Big(\biguplus_{i=1}^{n} \big(x_i + \varepsilon/\kappa_X \cdot B\big)\Big) \\ &\leq \lambda^d\big(a + \kappa_X \cdot (R + \varepsilon/\kappa_X) \cdot B\big) \\ &= \kappa_X^d \cdot (R + \varepsilon/\kappa_X)^d \cdot \lambda^d(B) \ \ . \end{aligned}$$

Solving this inequality for n proves Point (ii).

Under the assumptions of Point (iii) we continue the inclusion of (C.1) slightly different. From $B \subseteq (M - a)/\varepsilon_0$ we get

$$\begin{aligned} M + \varepsilon/\kappa_X \cdot B &\subseteq M + \varepsilon/\kappa_X \cdot \frac{M - a}{\varepsilon_0} \\ &= \Big(1 + \frac{\varepsilon}{\kappa_X \varepsilon_0}\Big) \cdot M - \frac{\varepsilon}{\kappa_X \varepsilon_0} \cdot a \ \ , \end{aligned}$$

where we used the identity $s_1 M + s_2 M = (s_1 + s_2) M$, which holds for all convex sets M and $s_1, s_2 > 0$, in the last step. Using the convexity of M again gives us the measurability of M, see e.g. [31, Satz II.7.7], and hence we can apply the Lebesgue measure

$$\begin{aligned} n \cdot (\varepsilon/\kappa_X)^d \cdot \lambda^d(B) &= \lambda^d\Big(\biguplus_{i=1}^{n} \big(x_i + \varepsilon/\kappa_X \cdot B\big)\Big) \\ &\leq \Big(1 + \frac{\varepsilon}{\kappa_X \varepsilon_0}\Big)^d \cdot \lambda^d(M) \ \ . \end{aligned}$$

Solving this inequality for n proves Point (iii). □

For a bounded set $M \subseteq \mathbb{R}^d$ with $\lambda^d(M) > 0$ Lemma C.7 gives the polynomial behavior $\mathcal{N}(M, \varepsilon) \asymp \mathcal{P}(M, \varepsilon) \asymp (1/\varepsilon)^d$ for $\varepsilon \to 0^+$. Consequently, the polynomial order is related to the dimension of a set. This observation can be used as the definition of a generalized or fractal dimension, the so-called *box-counting dimension.* For infinite-dimensional sets the covering numbers typically grow faster than polynomially for $\varepsilon \to 0^+$. In such cases it is more convenient to consider the *log-covering numbers*

$$\mathcal{H}(M,\varepsilon) \coloneqq \log \mathcal{N}(M,\varepsilon)$$

or the *dyadic entropy numbers*

$$e_n(M) \coloneqq \varepsilon_{2^{n-1}}(M) \ .$$

Analogously to entropy and covering numbers, bounds on dyadic entropy numbers can be converted into log-covering numbers bounds and vice versa. The following lemma is a direct translation of Corollary C.2 from entropy to dyadic entropy numbers and from covering to log-covering numbers.

C.8 Lemma (Conversion of Bounds) *Let $M \subseteq X$ be a subset of a quasi-metric space (X,d) and $F\colon [A,\infty) \to (0,B]$ be a decreasing bijective function with $A,B > 0$. Then the following statements are true:*

(i) If $e_n(M) \leq F(n)$ is satisfied for all $n \geq A$ then

$$\mathcal{H}(M,\varepsilon) \leq \log(2) \cdot \bigl(F^{-1}(\varepsilon) + 1\bigr)$$

is satisfied for all $0 < \varepsilon \leq B$.

(ii) If $\mathcal{H}(M,\varepsilon) \leq F^{-1}(\varepsilon)$ is satisfied for all $0 < \varepsilon \leq B$ then

$$e_n(M) \leq F\bigl(\log(2) \cdot (n-1)\bigr)$$

is satisfied for all $n \geq A/\log(2) + 1$.

In the polynomial regime the conversion is particularly pleasing. To be more precise, ignoring the constants we get, for $p > 0$,

$$\mathcal{H}(M,\varepsilon) \preccurlyeq \varepsilon^{-p} \qquad \Longleftrightarrow \qquad e_n(M) \preccurlyeq n^{-1/p} \ , \tag{C.2}$$

see also [76, Lemma 6.21 and Exercise 6.8].

Proof. (i) For $k \geq 2^A$ there is a (unique) $n \geq A$ with $2^n > k \geq 2^{n-1}$ and hence we have

$$\varepsilon_k(M) \le \varepsilon_{2^{n-1}}(M) = e_n(M) \le F(n) \le F\biggl(\frac{\log(k)}{\log(2)}\biggr) \enspace ,$$

where we used $n > \log(k)/\log(2)$ and the monotonicity of F in the last step. Consequently, we can apply Point (i) of Corollary C.2 with $f\colon [2^A, \infty) \to (0, B]$ given by $f(t) \coloneqq F\bigl(\log(t)/\log(2)\bigr)$. Since $f^{-1}(t) = \exp\bigl(\log(2)F^{-1}(t)\bigr)$ we find the assertion, namely, for $0 < \varepsilon \le B$,

$$\mathcal{N}(M, \varepsilon) \le f^{-1}(\varepsilon) + 1 \le \exp\bigl(\log(2) \cdot (F^{-1}(\varepsilon) + 1)\bigr) \enspace .$$

(ii) Using Point (ii) of Corollary C.2 for $f^{-1}\colon (0, B] \to [e^A, \infty)$ given by $f^{-1}(t) \coloneqq \exp\bigl(F^{-1}(t)\bigr)$ yields

$$\varepsilon_k(M) \le f(k) = F\bigl(\log(k)\bigr)$$

for all $k \ge e^A$. Since $k = 2^{n-1} \ge e^A$ is satisfied for $n \ge A/\log(2) + 1$ we get the assertion. □

As final part of this chapter we transfer the metric entropy quantities from sets to operators. To this end, let U and V be quasi-Banach spaces with quasi-norms $\|\cdot\|_U$ and $\|\cdot\|_V$. Moreover, we denote the closed unit balls by B_U and B_V, respectively. Then for a bounded (linear) operator $R\colon U \to V$ we define

$$\varepsilon_n(R) \coloneqq \varepsilon_n(RB_U) \qquad \text{and} \qquad \mathcal{N}(R, \varepsilon) \coloneqq \mathcal{N}(RB_U, \varepsilon)$$

for $n \ge 1$ and $\varepsilon > 0$, respectively. Analogously, the notion of packing numbers, log-covering numbers, and dyadic entropy numbers transfers from sets to operators. These metric entropy quantities allow a quantitative description of the compactness of an operator. The following lemma summarizes some basic properties of the covering numbers for operators.

C.9 Lemma (Covering Numbers for Operators) *Let U, V, and W be quasi-Banach spaces and $R, S\colon U \to V$, and $T\colon V \to W$ bounded (linear) operators. Then the following statements are true, for $\varepsilon, \delta > 0$:*

(i) $\mathcal{N}(R,\varepsilon) = \mathcal{N}(R\mathring{B}_U,\varepsilon)$.

(ii) $\mathcal{N}(R,\varepsilon) = 1$ *for all* $\varepsilon \geq \|R\|$.

(iii) $\|R\| \leq \kappa_V \varepsilon$ *for all* $\varepsilon > 0$ *with* $\mathcal{N}(R,\varepsilon) = 1$.

(iv) $\mathcal{N}\big(R+S, \kappa_V(\varepsilon+\delta)\big) \leq \mathcal{N}(R,\varepsilon)\mathcal{N}(S,\delta)$.

(v) $\mathcal{N}(TR,\varepsilon\delta) \leq \mathcal{N}(R,\varepsilon)\mathcal{N}(T,\delta)$.

(vi) $\mathcal{N}(R,\varepsilon) \leq \kappa_V^d \cdot \big(1+2\|R\|\kappa_V/\varepsilon\big)^d$ *if* $d := \operatorname{rank} R < \infty$.

Note that κ_V denotes the quasi-triangle constant of the quasi-Banach space V. As an important consequence of Point (ii), (iv), and (v) we find, for $\varepsilon > 0$,

$$\begin{aligned} \mathcal{N}\big(R+S, \kappa_V(\varepsilon+\|S\|)\big) &\leq \mathcal{N}(R,\varepsilon) \qquad \text{and} \\ \mathcal{N}\big(TR, \varepsilon\|T\|\big) &\leq \mathcal{N}(R,\varepsilon) \ , \end{aligned} \tag{C.3}$$

respectively.

Proof. (i) Since $R\mathring{B}_U \subseteq RB_U$ holds true, Point (ii) of Lemma C.4 yields the inequality "$\geq$". Using Point (iv) of Lemma C.4 we get

$$\mathcal{N}(R\mathring{B}_U,\varepsilon) = \mathcal{N}\big(\overline{R\mathring{B}_U},\varepsilon\big)$$

and the continuity of R implies

$$\overline{R\mathring{B}_U} \supseteq R\overline{\mathring{B}_U} = RB_U \ ,$$

see e.g. [32, Theorem 1.4.1 (v)]. Together we get the inequality "$\leq$".

(ii) For $\varepsilon \geq \|R\|$ the definition of the operator norm implies $RB_U \subseteq \|R\|B_V$. As a result, $N = \{0\}$ is an ε-net of RB_U and hence $\mathcal{N}(R,\varepsilon) = 1$.

(iii) For $\varepsilon > 0$ with $\mathcal{N}(R,\varepsilon) = 1$ there is some ε-net $N = \{v\}$ of RB_U consisting of one element. Then for $u \in B_U$ there are $v_+, v_- \in B_V$ with $Ru = v + \varepsilon v_+$ and $R(-u) = v + \varepsilon v_-$. This yields

$$2\|Ru\|_V = \big\|Ru - R(-u)\big\|_V = \varepsilon\|v_+ - v_-\|_V \leq 2\kappa_V\varepsilon$$

and hence $\|Ru\|_V \leq \kappa_V\varepsilon$ is proven for all $u \in B_U$.

(iv) This is a direct consequence of $(R+S)B_U \subseteq RB_U + SB_U$ and Point (i) of Lemma C.6.

(v) Let $N_V \subseteq V$ be an ε-net of RB_U with $|N_V| = \mathcal{N}(R,\varepsilon)$ and $N_W \subseteq W$ be a δ-net of TB_V with $|N_W| = \mathcal{N}(T,\delta)$. Since N_V is an ε-net, we have

$$RB_U \subseteq \bigcup_{v\in N_V} v + \varepsilon B_V \quad .$$

Applying T and using the δ-net property of N_W yields

$$\begin{aligned} TRB_U &\subseteq \bigcup_{v\in N_V} Tv + \varepsilon TB_V \\ &\subseteq \bigcup_{v\in N_V} Tv + \varepsilon\Bigg(\bigcup_{w\in N_W} w + \delta B_W \Bigg) \\ &= \bigcup_{\substack{v\in N_V,\\ w\in N_W}} (Tv + \varepsilon w) + \varepsilon\delta B_W \quad . \end{aligned}$$

As a result, $TN_V + \varepsilon N_W \subseteq W$ is an $\varepsilon\delta$-net of TRB_U with at most $|N_V|\cdot|N_W|$ elements. This proves the desired inequality.

(vi) Since $\operatorname{ran} R \subseteq V$ is a finite-dimensional space of dimension d, choosing a basis of $\operatorname{ran} R$ we can construct an isomorphism $T\colon \operatorname{ran} R \to \mathbb{R}^d$. Using the quasi-norm $\|x\| \coloneqq \|T^{-1}x\|_V$ for $x \in \mathbb{R}^d$ on $\mathbb{R}^d$ the mapping T becomes even an isometric isomorphism. Moreover, this definition implies $\kappa_{\mathbb{R}^d} = \kappa_V$. From Point (v) we get

$$\mathcal{N}(R,\varepsilon) = \mathcal{N}\big(T^{-1}TR,\varepsilon\big) \leq \mathcal{N}(TR,\varepsilon)\cdot\mathcal{N}\big(T^{-1},1\big) \quad .$$

Since T^{-1} is also isometric, we find $\|T^{-1}\| = 1$ and hence the second factor equals $\mathcal{N}(T^{-1},1) = 1$ according to Point (ii). For the first factor we have $TRB_U \subseteq \|TR\| B_{\mathbb{R}^d}$ and hence Point (ii) of Lemma C.7 implies

$$\mathcal{N}(TR,\varepsilon) \leq \mathcal{P}\big(TR,\varepsilon/2\big) \leq \kappa_V^d \cdot \big(1 + 2\|TR\|\kappa_V/\varepsilon\big)^d \quad .$$

Finally, using $\|TR\| = \|R\|$ and combining both inequalities gives the assertion. □

Recall that an operator $Q\colon X \to U$ between quasi-Banach spaces is called *metric surjection* if it satisfies $Q\mathring{B}_X = \mathring{B}_U$. For such an operator we directly get from Point (i) of Lemma C.9

$$\mathcal{N}(RQ, \varepsilon) = \mathcal{N}(R, \varepsilon)$$

for all $\varepsilon > 0$. Moreover, for an isometric operator $T\colon V \to W$ between quasi-Banach spaces, i.e. $\|Tv\|_W = \|v\|_V$ for $v \in V$, the definition of packing numbers directly gives

$$\mathcal{P}(TR, \varepsilon) = \mathcal{P}(R, \varepsilon)$$

for all $\varepsilon > 0$. In this case for the covering numbers only a weaker relation holds true. To be more precise, (C.3) with $\|T\| = 1$ and Lemma C.3 yield

$$\begin{aligned} \mathcal{N}(R, \varepsilon) \geq \mathcal{N}(TR, \varepsilon) &\geq \mathcal{P}(TR, \varepsilon\kappa_W) \\ &= \mathcal{P}(R, \varepsilon\kappa_W) \geq \mathcal{N}(R, 2\varepsilon\kappa_W) \ . \end{aligned} \tag{C.4}$$

Note that such an operator T is called *metric injection* in [19, Equation (1.3.5)].

Finally, recall that more details about metric entropy quantities can be found, for example, in [51, 19] and the references therein.

Bibliography

[1] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*, Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] M. Aoshima and K. Yata. High-dimensional quadratic classifiers in non-sparse settings. *Methodol. Comput. Appl. Probab.*, 21:663–682, 2018.

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

[4] H. Bauer. *Probability Theory*, De Gruyter, Berlin, 1996.

[5] H. Bauer. *Measure and Integration Theory*, De Gruyter, Berlin, 2001.

[6] R. Bellman. *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[7] P. J. Bickel and E. Levina. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10, 2004.

[8] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*, Cambridge University Press, Cambridge, 1989.

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

[10] I. K. Blaschzyk. *Improved Classification Rates for Localized Algorithms under Margin Conditions.* PhD thesis, Universität Stuttgart, 2020.

[11] V. I. Bogachev. *Measure Theory Volume 2*, Springer, Berlin, 2007.

[12] V. I. Bogachev. *Differentiable Measures and the Malliavin Calculus*, American Mathematical Society, Providence, 2010.

[13] R. Bojanic and E. Seneta. A unified theory of regularly varying sequences. *Math. Z.*, 134:91–106, 1973.

[14] R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2:107–144, 2005.

[15] T. Cai and X. Shen. *High-Dimensional Data Analysis*, World Scientific Publishing, Singapore, 2010.

[16] T. T. Cai and P. Hall. Prediction in functional linear regression. *Ann. Statist.*, 34:2159–2179, 2006.

[17] B. Carl. Entropy numbers of diagonal operators with an application to eigenvalue problems. *J. Approx. Theory*, 32:135–150, 1981.

[18] B. Carl and P. Rudolph. Entropy numbers of operators factoring through general diagonal operators. *Rev. Mat. Complut.*, 27:623–639, 2014.

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

[20] I. Chatzigeorgiou. Bounds on the Lambert function and their application to the outage analysis of user cooperation. *IEEE Commun. Lett.*, 17:1505–1508, 2013.

[21] F. Cobos and T. Kühn. Approximation and entropy numbers in Besov spaces of generalized smoothness. *J. Approx. Theory*, 160:56–70, 2009.

[22] D. L. Cohn. *Measure Theory*, Birkhäuser, New York, second edition, 2013.

[23] J. B. Conway. *A Course in Functional Analysis*, Springer, New York, second edition, 1990.

[24] F. Cucker and D.-X. Zhou. *Learning Theory*, Cambridge University Press, Cambridge, 2007.

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

[26] D. Djurčić and A. Torgašev. Representation theorems for the sequences of the classes CR_c and ER_c. *Siberian Math. J.*, 45:855–859, 2004.

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

[28] D. E. Edmunds and Y. Netrusov. Entropy numbers and interpolation. *Math. Ann.*, 351:963–977, 2010.

[29] D. E. Edmunds and Y. Netrusov. Schütt's theorem for vector-valued sequence spaces. *J. Approx. Theory*, 178:13–21, 2014.

[30] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*, Cambridge University Press, Cambridge, 1996.

[31] J. Elstrodt. *Maß- und Integrationstheorie*, Springer, Berlin, seventh edition, 2011.

[32] R. Engelking. *General Topology*, Heldermann, Berlin, second edition, 1989.

[33] J. Fan and Y. Fan. High-dimensional classification using features annealed independence rules. *Ann. Statist.*, 36, 2008.

[34] M. Farooq and I. Steinwart. Learning rates for kernel-based expectile regression. *Mach. Learn.*, 108:203–227, 2018.

[35] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis*, Springer, New York, 2006.

[36] F. Ferraty, A. Laksaci, A. Tadj, and P. Vieu. Rate of uniform consistency for nonparametric estimates with functional variables. *J. Statist. Plann. Inference*, 140:335–352, 2010.

[37] S. Fischer. Some new bounds on the entropy numbers of diagonal operators. *J. Approx. Theory*, 251:105343, 2020.

[38] S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *J. Mach. Learn. Res.*, 21:1–38, 2020.

[39] Y. Gordon, H. König, and C. Schütt. Geometric and probabilistic estimates for entropy and approximation numbers of operators. *J. Approx. Theory*, 49:219–239, 1987.

[40] E. Greenshtein. Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.*, 34, 2006.

[41] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10, 2004.

[42] O. Guédon and A. E. Litvak. *Euclidean projections of a p-convex body*, pages 95–108. Springer, Berlin, 2000.

[43] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*, Springer, New York, 2002.

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

[45] H. Hang. *Statistical Learning of Kernel Based Methods for Non-iid Observations.* PhD thesis, Universität Stuttgart, 2015.

[46] A. Ishii. A classifier under the strongly spiked eigenvalue model in high-dimension, low-sample-size context. *Comm. Statist. Theory Methods*, 49:1561–1577, 2020.

[47] B. Jiang, Z. Chen, and C. Leng. Dynamic linear discriminant analysis in high dimensional space. *Bernoulli*, 26, 2020.

[48] A. Klenke. *Probability Theory*, Springer, London, 2014.

[49] I. Koch, K. Naito, and H. Tanaka. Kernel naive Bayes discrimination for high-dimensional pattern recognition. *Aust. N. Z. J. Stat.*, 61: 401–428, 2019.

[50] M. Kohler, A. Krzyżak, and H. Walk. Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *J. Multivariate Anal.*, 97:311–323, 2006.

[51] A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *Uspekhi Mat. Nauk*, 17, 1961.

[52] D. Kong, K. Xue, F. Yao, and H. H. Zhang. Partially functional linear regression in high dimensions. *Biometrika*, 103:147–159, 2016.

[53] H. König. *Eigenvalue Distribution of Compact Operators*, Birkhäuser, Basel, 1986.

[54] J. Kuelbs and W. V. Li. Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.*, 116:133–157, 1993.

[55] T. Kühn. A lower estimate for entropy numbers. *J. Approx. Theory*, 110:120–124, 2001.

[56] T. Kühn. Entropy numbers of diagonal operators of logarithmic type. *Georgian Math. J.*, 8:307–318, 2001.

[57] T. Kühn. Entropy numbers of general diagonal operators. *Rev. Mat. Complut.*, 18:479–491, 2005.

[58] T. Kühn. Entropy numbers in sequence spaces with an application to weighted function spaces. *J. Approx. Theory*, 153:40–52, 2008.

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

[60] W. V. Li and W. Linde. Approximation, metric entropy and small ball estimates for Gaussian measures. *Ann. Probab.*, 27:1556–1578, 1999.

[61] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27, 1999.

[62] M. B. Marcus. The ε-entropy of some compact subsets of ℓ^p. *J. Approx. Theory*, 10:304–312, 1974.

[63] R. Meise and D. Vogt. *Introduction to Functional Analysis*, Clarendon Press, Oxford, 1997.

[64] B. S. Mitjagin. Approximate dimension and bases in nuclear spaces. *Russian Math. Surveys*, 16:59–127, 1961.

[65] N. Mücke. Stochastic gradient descent meets distribution regression. In *Proceedings of The 24^{th} International Conference on Artificial Intelligence and Statistics*, pages 2143–2151. PMLR, San Diego, 2021.

[66] N. Mücke and I. Steinwart. Empirical risk minimization in the interpolating regime with application to neural network learning. *arxiv e-prints*, 1905.10686v2, 2021.

[67] E. Novak and H. Woźniakowski. *Tractability of Multivariate Problems. Vol. I: Linear Information*, European Mathematical Society, Zürich, 2008.

[68] R. Oloff. Entropieeigenschaften von Diagonaloperatoren. *Math. Nachr.*, 86:157–165, 1978.

[69] A. Pietsch. *Operator Ideals*, North-Holland Publishing, Amsterdam, 1980.

[70] G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge University Press, Cambridge, 1989.

[71] R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[72] W. Rudin. *Principles of Mathematical Analysis*, McGraw-Hill, New York, third edition, 1976.

[73] B. Schölkopf and A. J. Smola. *Learning with Kernels*, MIT Press, Cambridge, 2001.

[74] C. Schütt. Entropy numbers of diagonal operators between symmetric Banach spaces. *J. Approx. Theory*, 40:121–128, 1984.

[75] I. Steinwart. *Entropy of $C(K)$-valued Operators and Some Applications.* PhD thesis, Universität Jena, 2000.

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

[77] I. Steinwart and S. Fischer. A closer look at covering number bounds for Gaussian kernels. *J. Complexity*, 62:101513, 2021.

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

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

[80] Z. Szabo, A. Gretton, B. Poczos, and B. Sriperumbudur. Two-stage sampled learning theory on distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 948–957. PMLR, San Diego, 2015.

[81] Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *J. Mach. Learn. Res.*, 17:1–40, 2016.

[82] H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland Publishing, Amsterdam, 1978.

[83] H. Triebel. *Theory of Function Spaces. III*, Birkhäuser, Basel, 2006.

[84] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Ann. Statist.*, 37:2655–2675, 2009.

[85] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*, Springer, New York, 1996.

[86] F. Yao and H.-G. Müller. Functional quadratic regression. *Biometrika*, 97:49–64, 2010.

[87] D.-X. Zhou. The covering number in learning theory. *J. Complexity*, 18:739–767, 2002.

[88] D.-X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Trans. Inf. Theory*, 49:1743–1752, 2003.

Abstract

This thesis is divided into three parts. In the first part we introduce a framework that allows us to investigate learning scenarios with restricted access to the data. We use this framework to model high-dimensional learning scenarios as an infinite-dimensional one in which the learning algorithm has only access to some finite-dimensional projections of the data. Finally, we provide a prototypical example of such an infinite-dimensional classification problem in which histograms can achieve polynomial learning rates.

In the second part we present some individual results that might by useful for the investigation of kernel-based learning methods using Gaussian kernels in high- or infinite-dimensional learning problems. To be more precise, we present log-covering number bounds for Gaussian reproducing kernel Hilbert spaces on general bounded subsets of the Euclidean space. Unlike previous results in this direction we focus on small explicit constants and their dependence on crucial parameters such as the kernel width as well as the size and dimension of the underlying space. Afterwards, we generalize these bounds to Gaussian kernels defined on special infinite-dimensional compact subsets of the sequence space ℓ_2. More precisely, the considered domains are given by the image of the unit ℓ_∞-ball under some diagonal operator.

In the third part we contribute some new insights to the compactness properties of diagonal operators from ℓ_p to ℓ_q for $p \neq q$.