Institut für Mathematik
Mathematische Statistik

Direct and Inverse Problems in Machine Learning : Kernel Methods and Spectral Regularization

Dissertation
zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Mathematische Statistik"

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Nicole Mücke

Potsdam, den 19. Juni 2017

Published online at the
Institutional Repository of the University of Potsdam:
URN urn:nbn:de:kobv:517-opus4-403479
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-403479

To my father

Acknowledgements

There are some people who have accompanied me on my way to my thesis and I would like to take this opportunity to say thanks.

First and foremost I want to thank my thesis adviser Gilles Blanchard for introducing me to the fascinating field of machine learning (with a strong emphasis on rigorous mathematical analysis) and for guiding me to the point where I could make an own proper contribution to the field. Without his guidance and encouragement this thesis would not exist.

Next I thank the referees Lorenzo Rosasco and Ingo Steinwart for reading my thesis very carefully and commenting helpfully and in detail. This is a major encouragement for my future work.

Not to be left unmentioned is our group in Potsdam (Franziska, Oleksandr, Maurilio, Alexandra), which has contributed with lots of good advice to the success of my work.

I also acknowledge valuable discussions with Peter Mathé on the subject of adaptivity.

I am indebted to my family (Markus, Rafael, Philippe Julien and Michael) for understanding and support, in particular at those not infrequent times, when writing and working on this thesis tended to become difficult. I would also like to say thanks to Felix, Irene, Iris and Kora at this point for demonstrating me the power of our family. For being a part, I am grateful. Thanks for all your unfailing love.

For many valuable conversations I would like to thank Angelika Kaddik.

Ramona, David and Jasmin: Thanks for being friends!

Contents

Abstract in deutscher Sprache **5**

Abstract **6**

1 Introduction **7**

1.1 Direct and Inverse Learning: Generalities 7

1.1.1 Minimax error in classical nonparametrics 8

1.1.2 Minimax error in a distribution-free context 9

1.2 Overview of the results of this thesis and comparison to related work 10

2 First results on the Direct and Inverse Learning Problem: Regular case **17**

2.1 Introduction and review of related work 17

2.2 Notation and Preliminaries 20

2.2.1 Inverse Problems induced by Carleman Operators 20

2.2.2 Discretization by random sampling 24

2.2.3 Statistical model, noise assumption, and prior classes 25

2.2.4 Effective Dimension 26

2.2.5 Equivalence with classical kernel learning setting 27

2.2.6 Regularization 28

2.3 Main results: upper and lower bounds on convergence rates 30

2.4 Discussion 32

2.5 Proof of Upper Rate 33

2.6 Proof of Lower Rate . . . 39

3 Minimax Rates beyond the regular Case **45**

3.1 Setting . . . 46

3.2 Main results . . . 47

3.3 Discussion . . . 48

3.4 Proofs . . . 49

3.4.1 Preliminaries . . . 49

3.4.2 Proof of upper rate . . . 52

3.4.3 Proof of minimax lower rate . . . 53

4 Distributed Learning **57**

4.1 Distributed Learning Algorithm . . . 57

4.2 Main Results . . . 58

4.3 Numerical Studies . . . 59

4.4 Discussion . . . 66

4.5 Proofs . . . 68

5 Adaptivity **77**

5.1 Introduction . . . 77

5.2 Empirical Effective Dimension . . . 80

5.3 Balancing Principle . . . 83

5.4 Applications . . . 90

5.5 Discussion . . . 97

5.6 Proofs . . . 99

6 Future Research **105**

6.1 Asymptotics of Effective Dimension . . . 105

6.2 Non-Linear Inverse Problems . . . 109

6.3 LocalNysation: Combining Localized Kernel Regression and Nyström Subsampling 113

Appendices **115**

A Tools **117**

A.1 Concentration Inequalities 117

A.2 A New Useful Inequality 121

A.3 Auxiliary Technical Lemmata 123

A.4 Some Operator Perturbation Inequalities 124

A.5 General Reduction Scheme 126

B **129**

LocalNysation: Combining Localized Kernel Regression and Nyström Subsampling **129**

B.1 Introduction and Motivation 130

B.1.1 Kernel Regression 130

B.1.2 Upper bounds on rates of convergence and optimality 131

B.1.3 Large Scale Problems: Localization and Subsampling 132

B.1.4 Contribution 133

B.2 The Partitioning Approach 134

B.2.1 Error Bounds 135

B.2.2 Improved Error Bound 136

B.3 KRR Nyström Subsampling 137

B.4 LocalNysation 138

B.5 Conclusion 139

B.6 Operators and norms 140

B.7 Proofs of Section B.2 141

B.8 Proofs of Section B.3 145

B.9 Proofs of Section B.4 148

B.10 Probabilistic Inequalities 150

B.11 Miscellanea 151

Bibliography **152**

Eigenständigkeitserklärung **159**

Abstract in deutscher Sprache

In dieser Arbeit analysieren wir ein zufälliges und verrauschtes inverses Regressionsmodell im *random design*. Wir konstruiueren aus gegebenen Daten eine Schätzung der unbekannten Funktion, von der wir annehmen, dass sie in einem Hilbertraum mit reproduzierendem Kern liegt.

Ein erstes Hauptergebnis dieser Arbeit betrifft obere Schranken an die Konvergenzraten. Wir legen sog. *source conditions* fest, definiert über geeignete Kugeln im Wertebereich von (reellen) Potenzen des normierten Kern-Kovarianzoperators. Das führt zu einer Einschränkung der Klasse der Verteilungen in einem statistischen Modell, in dem die spektrale Asymptotik des von der Randverteilung abhängigen Kovarianzoperators eingeschränkt wird.

In diesem Kontext zeigen wir obere und entsprechende untere Schranken für die Konvergenzraten für eine sehr allgemeine Klasse spektraler Regularisierungsmethoden und etablieren damit die sog. *Minimax-Optimalität* dieser Raten. Da selbst bei optimalen Konvergenzraten Kernmethoden, angewandt auf große Datenmengen, noch unbefriedigend viel Zeit verschlingen und hohen Speicherbedarf aufweisen, untersuchen wir einen Zugang zur Zeitersparnis und zur Reduktion des Speicherbedarfs detaillierter. Wir studieren das sog. *distributed learning* und beweisen für unsere Klasse allgemeiner spektraler Regularisierungen ein neues Resultat, allerdings immer noch unter der Annahme einer bekannten *a priori* Regularität der Zielfunktion, ausgedrückt durch die Fixierung einer source condition. Das große Problem bei der Behandlung realer Daten ist das der *Adaptivität*, d.h. die Angabe eines Verfahrens, das ohne eine solche *a priori* Voraussetzung einen in einem gewissen Sinn optimalen Schätzer aus den Daten konstruiert. Das behandeln wir vermöge einer Variante des *balancing principle*.

Abstract

We analyse an inverse noisy regression model under random design with the aim of estimating the unknown target function based on a given set of data, drawn according to some unknwon probability distribution. Our estimators are all constructed by kernel methods, which depend on a Reproducing Kernel Hilbert Space structure using spectral regularization methods.

A first main result establishes upper and lower bounds for the rate of convergence under a given *source condition* assumption, restricting the class of admissible distributions. But since kernel methods scale poorly when massive datasets are involved, we study one example for saving computation time and memory requirements in more detail. We show that Parallelizing spectral algorithms also leads to minimax optimal rates of convergence provided the number of machines is chosen appropriately.

We emphasize that so far all estimators depend on the assumed *a-priori* smoothness of the target function and on the eigenvalue decay of the kernel covariance operator, which are in general unknown. To obtain good purely data driven estimators constitutes the problem of *adaptivity* which we handle for the single machine problem via a version of the *Lepskii principle*.

Chapter 1

Introduction

1.1 Direct and Inverse Learning: Generalities

Let A be a known linear operator from a Hilbert space $\mathcal{H}_1$ to a linear space $\mathcal{H}_2$ of real-valued functions on some input space $\mathcal{X}$. In this thesis, we consider a random and noisy observation scheme of the form

$$Y_i := g(X_i) + \varepsilon_i \,, \quad g = Af \,, \quad i = 1, \dots, n \,, \tag{1.1.1}$$

at i.i.d. data points $X_1, \dots, X_n$, drawn according to a probability distribution ν on $\mathcal{X}$, where ε_i are independent centered noise variables. Here $\mathcal{X}$ is taken as a standard Borel space. For simplicity, we take the output space $\mathcal{Y}$ as the set of real numbers, but this could be generalized to any separable Hilbert space. More precisely, we assume that the observed data $(X_i, Y_i)_{1 \le i \le n} \in (\mathcal{X} \times \mathcal{Y})^n$ are i.i.d. observations, with $\mathbb{E}[Y_i | X_i] = g(X_i)$, so that the distribution of ε_i may depend on X_i, while satisfying $\mathbb{E}[\varepsilon_i | X_i] = 0$. This is also commonly called a *statistical learning* setting, in the sense that the data (X_i, Y_i) are generated by some external random source and the *learner* aims to infer from the data some reconstruction $\widehat{f}_n$ of f, without having influence on the underlying sampling distribution ν. For this reason, we call the model (1.1.1) an *inverse statistical learning problem*. The special case $A = I$ is just non-parametric regression under random design, which we also call the direct problem. Thus, introducing a general A gives a unified approach to the direct and inverse problem.

In the statistical learning context, the relevant notion of convergence and associated reconstruction rates to recover f concern the limit $n \to \infty$. More specifically, let $\widehat{f}_n$ be an estimator of f based on the observed data $(X_i, Y_i)_{1 \le i \le n}$. The usual notion of estimation error in the statistical learning context is the averaged squared loss for the prediction of $g(X)$ at a new independent sample point X:

$$\mathbb{E}_{X \sim \nu}[(g(X) - A\widehat{f}_n(X))^2] = \left\| A(f - \widehat{f}_n) \right\|^2_{L^2(\nu)} . \tag{1.1.2}$$

In this work, we are interested as well in the *inverse* reconstruction problem, that is, the reconstruction error for f itself, i.e.,

$$\left\| f - \widehat{f}_n \right\|^2_{\mathcal{H}_1} .$$

Estimates in $L^2(\nu)$-norm are standard in the learning context, while estimates in $\mathcal{H}_1$-norm are standard

for inverse problems, and our results will present convergence results for a family of norms interpolating between these two. We emphasize that $\left\|A(f-\widehat{f}_n)\right\|^2_{L^2(\nu)}$ as well as $\left\|f-\widehat{f}_n\right\|^2_{\mathcal{H}_1}$ are random variables, depending on the observations. Thus the error rates above can be estimated either in expectation or in probability.

In this thesis we will present convergence rates for these different criteria, as n tends to infinity, in expectation for moments of all orders (in Cpapter 5, where we treat the problem of adaptivity via a version of the balancing principle, we also consider estimates in probability). Our analysis always is in the context of kernel methods: For the direct problem (i.e. $A = 1$), a reproducing kernel Hilbert space (RKHS) structure is imposed on $\mathcal{H}_1$ by choosing a positive semi-definite kernel, while for the inverse problem an RKHS structure is imposed on $\mathrm{Im}(A)$ by the operator A (assuming the evaluation functionals $f \mapsto Af(x)$ being uniformly bounded).

As a general rule, sequences of estimators are produced throughout this thesis by general spectral regularization methods: The function $g_0(t) := t^{-1}$ is replaced, by use of an additional positive so called spectral parameter λ, by a modified real function $g_\lambda(t)$, which in some sense converges to $g_0(t)$ as $\lambda \downarrow 0$. One then introduces the covariance operator

$$B : \mathcal{H}_K \ni g \mapsto \int g(x)K(x,.)\, d\nu(x) \in \mathcal{H}_K \ ,$$

which simply is the integral operator restricted to the RKHS $\mathcal{H}_K$ and induced by the real symmetric kernel $K(x,y)$ and the sampling measure ν. Then, $g_\lambda(B)$ is self-adjoint, non-negative and trace class and defines a spectral regularization operator of the ill-defined inverse B^{-1}. Note that zero is necessarily contained in the spectrum of the trace class operator B. One then chooses the regularization parameter $\lambda = \lambda_n$ depending on the sample size, and this induces our sequence of estimators.

The basic idea for choosing the regularization parameter is the following: One writes the overall error as a sum of the approximation error (also called bias) and the sample error (also called variance), the decomposition depending on λ. Predominance of the approximation error (large bias) gives underfitting, predominance of the sample error (large variance) gives overfitting. The regularization parameter λ_n will be chosen to make these conceptually very different types of error equal in magnitude, thus striking a good compromise between under- and overfitting. Since this depends on the sample size, the procedure will lead to convergence rates as a function of n. For more details on this well established procedure of spectral regularization and a precise definition of the classes treated in this thesis we refer to Section 2.2.

1.1.1 Minimax error in classical nonparametrics

When upper bounds or convergence rates for a specific method are obtained, it is natural to ask to what extent they can be considered optimal. The classical yardstick is the notion of minimax error over a set $\mathcal{M}$ of candidates (hypotheses) for the data generating distribution ρ:

$$\mathcal{R}(\mathcal{M}, n) := \inf_{\widehat{f}} \sup_{\rho \in \mathcal{M}} \mathbb{E}_{\mathcal{D}\sim\rho^{\otimes n}} \left[\|\widehat{f} - f_\rho\|^2_{2,\nu} \right], \tag{1.1.3}$$

where the inf operation is over all estimators, and we added an index ρ to f_ρ to emphasize its dependence on the data generating distribution.

In the nonparametric statistics literature, it is commonly assumed that $\mathcal{X}$ is some compact set of $\mathbb{R}^d$, the sampling distribution ν has an upper bounded density with respect to the Lebesgue measure and the type of regularity considered for the target function is a Sobolev-type regularity, i.e., the target function f_ρ has a squared-integrable r-th derivative. This is equivalent to saying that f_ρ belongs to a Sobolev ellipsoid of radius R,

$$f_\rho \in \left\{ f : \sum_{i \geq 1} i^{-\frac{2r}{d}} f_i^2 \leq R^2 \right\}, \tag{1.1.4}$$

where f_i denotes the coefficients of f in a (multidimensional) trigonometric function basis. Minimax rates in such context are known to be of the order $O(n^{-\frac{2r}{2r+d}})$ and can be attained for a variety of classical procedures [83, 87].

1.1.2 Minimax error in a distribution-free context

In the statistical learning context, the above assumptions are unsatisfying. The first reason is that learning using kernels is often applied to non-standard spaces, for instance graphs, text strings, or even probability distributions (see, e.g., [22]). There is often no "canonical" notion of regularity of a function on such spaces, nor a canonical reference measure which would take the role of the Lebesgue measure in $\mathbb{R}^d$. The second reason is that learning theory focuses on a distribution-free approach, i.e., avoiding specific assumptions on the generating distribution. By contrast, it is a very strong assumption to posit that the sampling distribution ν is dominated by some reference measure (be it Lebesgue or otherwise), especially for non-standard spaces, or in $\mathbb{R}^d$ if the dimension d is large. In the latter case, the convergence rate $O(n^{-\frac{2r}{2r+d}})$ becomes very slow (the curse of high dimensionality), yet it is often noticed in practice that many kernel-based methods perform well. The reason is that for high-dimensional data, more often than not the sampling distribution ν is actually concentrated on some lower-dimensional structure, so that the assumption of ν having bounded density in $\mathbb{R}^d$ is violated: convergence rates could then be much faster. For these reasons, it has been proposed to consider regularity classes for the target function having a form similar to (1.1.4), but reflecting implicitly the geometry corresponding to the choice of the kernel and to the sampling distribution. More precisely, denote by B_ν the (uncentered) covariance operator of the kernel feature mapping $\Phi(X)$ and by $(\mu_{\nu,i}, \psi_{\nu,i})_{i \geq 1}$ an eigendecomposition of B_ν. For $r, R > 0$, introduce the class

$$\Omega_\nu(r, R) := \left\{ f \in \mathcal{H}_1 : \sum_{i \geq 1} \mu_{\nu,i}^r f_i^2 \leq R^2 \right\} = B_\nu^r B(\mathcal{H}_1, R), \tag{1.1.5}$$

where $B(\mathcal{H}_1, R)$ is the ball of radius R in $\mathcal{H}_1$, and $f_i := \langle f, \psi_{\nu,i} \rangle$ are the coefficients of f in the eigenbasis. To be explicit, we here have indicated the dependence on the marginal distribution ν by a subscript; we shall also allow ourselves the liberty to drop this subscript for reasons of brevity. Clearly, (1.1.5) has a form similar to (1.1.4), but in a basis and scaling that reflects the properties of the distribution of $\Phi(X)$. If the target function f_ρ is well approximated in this basis in the sense that its coefficients decay fast enough in comparison to the eigenvalues, it is considered as regular in this geometry (higher regularity corresponds to higher values of r and/or lower values of R). This type of regularity class, also called a *source condition*, has been considered in a statistical learning context by [23]. The authors in [24] have established upper bounds for the performance of kernel ridge regression $\widehat{f}_\lambda$ over such classes. This has been extended to other types of kernel regularization methods by [17, 20, 29]. These bounds rely on tools

introduced in the seminal work of [94] and depend in particular on the notion of *effective dimension* [1] of the data with respect to the regularization parameter λ, defined as

$$\mathcal{N}(\lambda) := \mathrm{Tr}\left[(B_\nu + \lambda)^{-1} B_\nu\right] = \sum_{i \geq 1} \frac{\mu_{\nu,i}}{\mu_{\nu,i} + \lambda}\,. \tag{1.1.6}$$

As before, the next question of importance is whether such upper bounds can be proved to be minimax optimal over the classs $\Omega_\nu(r, R)$, assuming the regularization parameter λ is tuned appropriately. This question has been answered positively when a polynomial decay of the eigenvalues, $\mu_{v,i} \asymp i^{-b}$, is assumed ($\asymp$ stands for upper and lower bounded up to a constant). In this case $\mathcal{N}(\lambda)$ can be evaluated, and for an appropriate choice of λ, the upper bound can be matched by a corresponding lower bound (see Chapter 2).

1.2 Overview of the results of this thesis and comparison to related work

In this section we present a short, informal overview of the results of this thesis. Chapter 2 contains generalities of relevance throughout the entire thesis as well as first results on minimax optimality in the regular case. We start to show that, under appropriate assumptions, we can endow $\mathrm{Im}(A)$ with an appropriate reproducing kernel Hilbert space (RKHS) structure $\mathcal{H}_K$ with reproducing kernel K, such that A is a partial isometry from $\mathcal{H}_1$ to $\mathcal{H}_K$. Through this partial isometry, the initial problem (1.1.1) can be formally reduced to the problem of estimating the function $g \in \mathcal{H}_K$ by some $\widehat{g}$; control of the error $(g - \widehat{g})$ in $L^2(\nu)$-norm corresponds to the direct (prediction) problem, while control of this difference in $\mathcal{H}_K$-norm is equivalent to the inverse (reconstruction) problem. In particular, the kernel K completely encapsulates the information about the operator A. This equivalence also allows a direct comparison to previous existing results for convergence rates of statistical learning using an RKHS formalism (see the next chapter). Letting B be the covariance operator introduced above, the rates of convergence presented in this work will be governed by a *source condition* assumption on f of the form $\|B^{-r} f\| \leq R$ for some constants $r, R > 0$ as well as by the *ill-posedness* of the problem as measured by an assumed decay of the eigenvalues of B. At first, in the so called regular case of our Chapter 2, we will assume that this is precisely given by a power law specified by an exponent $b > 1$. Our main upper bound result establishes that for a broad class of estimators defined via spectral regularization methods, for $s \in [0, \frac{1}{2}]$ one has in the sense of p-th moment expectation that

$$\|B^s (g - \widehat{g}_{\lambda_n})\|_{\mathcal{H}_K} \lesssim R \left(\frac{\sigma^2}{R^2 n}\right)^{\frac{(r+s)}{2r+1+1/b}}\,,$$

[1] $\mathcal{N}(\lambda)$ describes the capacity of the hypothesis space or some kind of effective volume, somewhat related to the phase space volume in the classical Weyl estimates for the number of eigenvalues of an elliptic operator P (see e.g. [45] and [31]), the covariance B_ν being analog to the inverse P^{-1}. In the special case considered in our Chapter 3 we shall amplify by relating the effective dimension to a counting function for eigenvalues (by a rough estimate, see Lemma 3.4.2 for an upper and Lemma 3.4.3 for a lower bound). The term *effective dimension* then simply refers (roughly) to the dimension of the associated eigenspace, which coincides with the number of eigenvalues counted with multiplicity. In general, however, even for finite rank operators or matrices, the rank of the operator might be larger than the effective dimension, since very small eigenvalues different from zero contribute to the rank with weight 1, but with smaller weight, depending on λ, to the effective dimension.

for an appropriate choice of the regularization parameter λ_n. (Note that $s = 0$ corresponds to the reconstruction error, and $s = \frac{1}{2}$ to the prediction error, i.e., $L^2(\nu)$ norm.) Here, σ^2 denotes noise variance (and we remark that classical Bernstein moment conditions are assumed to hold for the noise). The symbol $\lesssim$ means that the inequality holds up to a multiplicative constant that can depend on various parameters entering in the assumptions of the result, but not on n, σ, nor R. An important assumption is that the inequality $q \geq r + s$ should hold, where q is the *qualification* of the regularization method, a quantity defined in the classical theory of inverse problems (see Section 2.2.6 for a precise definition).

This result is complemented by a minimax lower bound which matches the above rate not only in the exponent in n, but also in the precise behavior of the multiplicative constant as a function of R and the noise variance σ^2. The obtained lower bounds come in two flavors, which we call *weak* and *strong* asymptotic lower bounds, respectively (see Section 2.3).

Concerning related work, we remark that the analysis of inverse problems, discretized via (noisy) observations at a finite number of points, has a long history, which we will not attempt to cover in detail here. The introduction of reproducing kernel Hilbert space-based methods was a crucial step forward in the end of the 1970s. Early references have focused, mostly, on spline methods on $[0, 1]^d$; on observation point designs either deterministic regular, or random with a sampling probability comparable to Lebesgue; and on assumed regularity of the target function in terms of usual differentiability properties. We refer to [88] and references therein for a general overview. An early reference establishing convergence rates in a random design setting for (possibly nonlinear) inverse problems in a setup similar to those delineated above and a Tikhonov-type regularization method is [69]. Analysis of the convergence of fairly general regularization schemes for statistical inverse problems under a Hilbertian noise model was established in [10]. While these authors make the argument that this model can cover random sampling, to compute the regularized estimator they propose it must be assumed that the sampling distribution ν is known to the user. In this thesis we consider the more challenging setting where this distribution is unknown (and investigate if one can attain the same convergence rates).

We henceforth focus our attention on the more recent thread of literature concerning the *statistical learning* setting, whose results are more directly comparable to ours. In this setting, the emphasis is on general input spaces, and "distribution-free" results, which is to say, random sampling whose distribution ν is unknown, quite arbitrary and out of the control of the user. The use of reproducing kernel methods has enjoyed a wide popularity in this context since the 1990s, mainly for the direct learning problem. The connections between (the direct problem of) statistical learning using reproducing kernel methods, and inverse problem methodology, were first noted and studied in [26, 27, 38]. In particular, in [38] it was proposed to use general form regularization methods from the inverse problem literature for kernel-based statistical learning. There is a vast recent literature relating learning to regularization techniques for inverse problems (see [66], [89], [40] to mention just a few), confirming the strong conceptual analogy of certain learning algorithms with regularization algorithms. For example, Tikhonov regularization is known as regularized least-squares algorithm or ridge regression, while Landweber iteration is related to L^2-boosting or gradient descent, see, e.g. [93] and [18].

Concerning the history of upper rates of convergence in an RKHS setting, covering number techniques were used in [23] to obtain (non-asymptotic) upper rates. In [27], [78], [79] these techniques were replaced by estimates on integral operators via concentration inequalities, and this is the path we follow in this thesis. For a more detailed presentation we refer to Chapter 2. In some sense, the crucial step of all these

results is in effectively computing the effective dimension introduced above, assuming power law decay of the eigenvalues. The first comprehensive result in this direction was established by [24] ; our paper [17] gives a sharp estimate of the convergence rate in this case including the dependence on the parameters R and noise variance σ , namely $O\Big(R^2\big(\frac{\sigma^2}{R^2 n}\big)^{\frac{r+s}{2r+1+\frac{1}{b}}}\Big)$. Our presentation in Chapter 2 essentially follows that paper. But, while the assumption of polynomially decaying eigenvalues yields explicit minimax rates of convergence and ensures that kernel methods can achieve those optimal rates, it is unsatisfying from a distribution-free point of view. Remember: The stucture of the eigenvalues reflects the covariance of the feature mapping $\Phi(X)$; for complex data, there is no strong reason to expect that their decay should be strictly polynomial.
As a first step in treating decay behavior of eigenvalues as general as possible we present some results beyond the regular case in Chapter 3. We show that kernel methods are also able of achieving minimax optimal rates in this more general case, for target function classes of the form (1.1.5).
We remark that, while direct and inverse kernel-based methods for solving non-parametric (direct or inverse) regression problems are attractive because they attain asymptotically minimax optimal rates of convergence, these methods scale poorly when massive datasets are involved. Large training sets give rise to large computational and storage costs. For example, computing a kernel ridge regression estimate needs inversion of a $n \times n$- matrix, with n the sample size. This requires $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory, which becomes prohibitive for large sample sizes. For this reason, various methods have been developed for saving computation time and memory requirements. Among them are e.g. *low-rank approximation* of the kernel matrix, *early-stopping* and *distributed learning* . We shall give a quick overview.

During the last years, a huge amount of research effort was devoted to finding low-rank approximations of the kernel matrix, both from an algorithmic and an inferential perspective (providing statistical guarantees). Important examples include Cholesky decomposition [3], Nyström sampling, see e.g. [90] , [2] , [77] , (randomized) sketches [92] , [1] , sparse greedy approximations [80] and others. The common feature of all these methods is to replace the theoretically optimal approximation obtained by a spectral decomposition (which requires time at least $\mathcal{O}(n^2)$) by a less ambitious suitable low rank approximation of the kernel matrix via column sampling, reducing run time to $\mathcal{O}(np^2)$ where p denotes the rank of the approximation. Clearly, the rules of the game are to choose p as small as possible while maintaining minimax optimality of convergence rates (hopefully in $L^2(\nu)-$ norm and RKHS$-$ norm) and to explicitely determine this p as a function of the sample size n (hopefully for a general class of spectral regularization methods), keeping track of the source condition and the rate of eigenvalue decay, entering the estimate via the effective dimension. This is usually done by solving computational-statistical trade-offs. Compared to that obviously very desirable standard there are yet only very partial results in the literature: Only KRR has been analyzed, excluding higher smoothness of the regression function (the case $r > 1$ in the source condition).

An alternative approach for reducing time complexity lies in early stopping of iterative regularization algorithms, e.g. gradient descent, see [93], [72] and conjugate gradient regression, see [12] , [13]. The number of iterations serves as regularization parameter. Optimal stopping is determined by some parameter selection rule, e.g. by solving a bias-variance trade-off (gradient descent) or by the discrepancy principle (conjugate gradient regression). Early stopping both reduces run time and provides regularization preventing overfitting. Similar to the results for low rank approximation, the early stopping index turns out to depend on *a priori* assumptions, reflected in the source condition and effective dimension. Time complexity is reduced to $\mathcal{O}(Tn^2)$ where T is the stopping index. Since early stopping still suffers

from high memory requirements, there is further research devoted to overcoming this issue by combining early stopping with subsampling methods, see e.g. [19].

Another standard tool for saving computation time and memory requirements is distributed learning (DL), and here this thesis will make a new contribution. In Chapter 4 we shall study the DL approach for the aforementioned statistical learning problem

$$Y_i := Af(X_j) + \varepsilon_i \, , j = 1 \, , \ldots , n \, , \tag{1.2.1}$$

at random i.i.d. data points $X_1, \ldots, X_n$ drawn according to a probability distribution ν on $\mathcal{X}$, where ε_j are independent centered noise variables. We uniformly partition the given data set $D = \{(X_1, Y_1), ..., (X_n, Y_n)\} \subset (\mathcal{X} \times \mathbb{R})^n$ into m disjoint equal-size subsets $D_1, ..., D_m$. On each subset D_j, we compute a local estimator $\hat{f}^{\lambda}_{D_j}$, using a spectral regularization method (with qualification $q \geq r + s$). The final estimator for the target function f_ρ is obtained by simple averaging: $\bar{f}^{\lambda}_D := \frac{1}{m} \sum_{j=1}^{m} \hat{f}^{\lambda}_{D_j}$.

Our aim is to extend our results from the non-distributed setting ($m = 1$) to distributed learning and to provide conditions for retaining minimax optimal rates. As before, our rates of convergence are governed by a source condition assumption on f_ρ of the form $||B^{-r} f_\rho||_{\mathcal{H}_1} \leq R$ for some constants $r, R > 0$ as well as by some capacity assumption $\mathcal{N}(\lambda) \lesssim \lambda^{-\frac{1}{b}}$, where in our case the rate $b > 1$ is induced from the assumed eigenvalue decay. We show, that for $s \in [0, \frac{1}{2}]$ in the sense of p-th moment expectation

$$\left\| B^s (f - \bar{f}^{\lambda_n}_D) \right\|_{\mathcal{H}_1} \lesssim R \left(\frac{\sigma^2}{R^2 n} \right)^{\frac{b(r+s)}{2br+b+1}} . \tag{1.2.2}$$

Basic problems are the choice of the regularization parameter on the subsamples and, most importantly, the proper choice of m, since it is well known that choosing m too large gives a suboptimal convergence rate in the limit $n \to \infty$, see, e.g., [91]. We show, that by choosing λ_n depending on the *global* sample size n, the number of subsample sets is allowed to grow at most polynomially with n, namely

$$m \leq n^{\alpha} \, , \qquad \alpha = \frac{2b \min(r, 1)}{2br + b + 1} \, .$$

Our approach to this problem is classical. Using a bias-variance decomposition and choosing the regularization parameter according to the total sample size n yields undersmoothing on each of the m individual samples and causes an inflation of variance, but m-fold averaging reduces the variance sufficiently to get minimax optimality. The bias estimate is then straightforward.

For the hard part we write the variance as a sum of independent random variables, allowing to successfully apply Rosenthal's inequality (in the Hilbert space case), see [70]. Comparable results mostly focus on KRR, corresponding to Tikhonov regularization. In [95] the authors derive minimax-optimal rates in 3 cases (finite rank kernels, sub - Gaussian decay of eigenvalues of the kernel and polynomial decay), provided m satisfies a certain upper bound, depending on the rate of decay of the eigenvalues and an additional crucial upper bound on the eigenfunctions ϕ_j of the integral operator associated to the kernel K (see Section 4.4). Proving such a condition often turns out to be a great hurdle. In fact, it is not understood for which kernels (and marginals ν) such an eigenfunction assumption is satisfied. It is therefore of great interest to investigate if and how m can be allowed to go to infinity as a function of n *without imposing any conditions on the eigenfunctions* of the kernel. Results in this direction have

been obtained in the recent paper [58], for KRR, which is a great improvement on the worst rate of [95]. The authors follow [95] in giving estimates only in $L^2(\nu)-$ norm and dub their approach *a second order decomposition*, which uses concentration inequalities and certain resolvent identities adapted to KRR. These results are extended (independently from our work) in [97] to general spectral regularization methods, but with a proof yet unpublished at the time of submission of this thesis. Our results cover the whole range of interpolation norms between RKHS norm and prediction error. Restricted to $L^2(\nu)-$ norm, they match those given in [97] and [58], but are more precise concerning the scaling of the noise variance σ^2 and the radius $R > 0$ from the source condition.

This basically sums up the contributions of this thesis to minimax optimality. Our final result concerns adaptivity via a version of the balancing principle. Here, our approach and our results are somewhat different from the bulk of this thesis (e.g. we only derive estimates in large probability and we do not push our estimates, as elsewhere, to estimates in expectation), and our results are possibly not yet in final form. Clearly, the problem of adaptivity is both of great theoretical and practical interest: On one side, an appropriate choice of the regularization parameter is essential for spectral regularization to work well, while on the other side in any statistical problem any *a priori* choice of the regularization parameter is bad since it should dependent on the unknown source conditions describing the given set of data.

There is a number of (sometimes very different) approaches to address this problem in the context of learning, some of them depending on data-splitting (e.g. cross-validation). An attractive approach avoiding data-splitting is the balancing principle which originated in the seminal paper of Lepskii [56] and has been elaborated in quite a number of papers, see, e.g., [56], [57], [41], [8], [63] and references therein. In the context of Learning Theory the first comprehensive version is the paper [25]. We basically follow this approach, adapting it to the case of fast (minimax optimal) rates studied in the rest of this thesis. As a technical complication, this leads to a conceptually important loss of uniformity (with respect to the regularization parameter) in the constants appearing in the basic probabilistic error estimates needed for balancing and finally results in an additional $\log \log n$ term describing the data-driven estimator obtained from balancing, somewhat spoiling the minimax optimal convergence rate. For slow rates, the slightly different approach of [25] - based on an additive instead of a multiplicative error decomposition - gives uniformity which our approach does not achieve, even if specialized to the case of slow rates. For more details we refer to our discussion in Section 5.5.

Crucial for our approach is a two-sided estimate of the effective dimension in terms of its empirical approximation. This in particular allows to control the spectral structure of the covariance operator through the given input data. A further very convenient result is the fact that balancing in $L^2(\nu)-$ norm (which is easiest) automatically gives good balancing in all other (stronger) interpolation norms. An analogous result is open for other approaches to data dependent choices of the regularization parameter, e.g. for hold-out (see our Discussion in Section 2.4). We think that all of our contributions to this subject are important and valid steps in a future and possibly more comprehensive solution of this important problem, but at present it does not yet seem to be in final form.

The outline of the thesis is as follows: Chapter 2 covers the general introduction to the class of models considered in this thesis and gives first results on minimax-optimality for the single machine problem in the regular case. Chapter 3 presents some results beyond the regular case, relaxing the conditions on the asymptotic behaviour of the eigenvalues of the covariance operator, but basically keeping the same notion of source conditions. Chapter 4 covers the case of distributed learning, while Chapter 5 contains a

discussion of adaptivity via the balancing principle. In Chapter 6 we present some thoughts about future research, based on this thesis. The Appendix A collects (known) background information which is useful for our discussion in Chapters 2 -5. It is put in the Appendix in order to avoid disturbing the flow of our arguments, but, for the sake of the reader, we do not simply refer to the original literature but present things in a form adapted to our line of argument in the main text. In Appendix B we add a first result on a combination of a localized approach with subsampling methods.

Chapter 2

First results on the Direct and Inverse Learning Problem: Regular case

2.1 Introduction and review of related work

As mentioned in Chapter 1, the general introduction of this thesis, we consider a random and noisy observation scheme of the form

$$Y_i := g(X_i) + \varepsilon_i \,, \quad g = Af \,, \quad i = 1, \ldots, n \,, \tag{2.1.1}$$

at i.i.d. data points $X_1, \ldots, X_n$ drawn according to a probability distribution ν on $\mathcal{X}$, where ε_i are independent centered noise variables. Here A is a known linear operator from a Hilbert space $\mathcal{H}_1$ to a linear space $\mathcal{H}_2$ and we start with the assumption that the map $(f, x) \mapsto (Af)(x)$ is continuous in f and measurable in x, which implies that A can be seen as a Carleman operator from $\mathcal{H}_1$ to $L^2(\nu)$. This point of view goes back to [26], where this more general setting of the random discretization of an inverse problem defined by a Carleman operator has been considered.

Moreover, we observe that $\mathrm{Im}(A)$ can be endowed with a RKHS structure $\mathcal{H}_K$ such that A is a partial isometry from $\mathcal{H}_1$ onto $\mathcal{H}_K$. While we do not expect this result to be considered as a true novelty, it was not explicitly mentioned in [26] and in our opinion it helps to clarify the equivalence between inverse statistical learning and direct learning with reproducing kernels. In particular, it makes possible a direct comparison between our results and previous results for the direct (kernel) learning problem.

We shall now briefly review previous results which are directly comparable to ours: Smale and Zhou [79], Bauer et al. [4], Yao et al. [93], Caponnetto and De Vito [24] and Caponnetto [20]. For convenience, we try to condense the most essential points in Table 2.1. Compared with our more general setting, all of these previous references only consider the special case $A = I$, but assume from the onset that $\mathcal{H}_1$ is a RKHS with given kernel. Thus, in the first column of Table 2.1, A is the identity and $g = f$, and

	$\left\|A(\widehat{f}_n - f)\right\|_{L^2(\nu)}$	$\left\|\widehat{f}_n - f\right\|_{\mathcal{H}_K}$	Assumptions (q: qualification)	Method
Smale-Zhou [79]	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r+1}{2r+2}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+1}}$	$r \leq \frac{1}{2} (= q - \frac{1}{2})$	Tikhonov
Bauer et al. [4]	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r+1}{2r+2}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+1}}$	$r \leq q - \frac{1}{2}$	General
Yao et al. [93]	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r+1}{2r+3}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+\frac{5}{2}}}$	$(q = \infty)$	Landweber iteration
Caponnetto-De Vito [24]	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{(2r+1)}{2r+1+\frac{1}{b}}}$	N/A	$r \leq \frac{1}{2} (= q - \frac{1}{2})$	Tikhonov
Caponnetto [20] Caponnetto-Yao [21]	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{(2r+1)}{2r+1+\frac{1}{b}}}$	N/A	$r \leq q - \frac{1}{2}$ +unlabeled data if $2r + \frac{1}{b} < 1$	General
This thesis	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{(2r+1)}{2r+1+\frac{1}{b}}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r}{2r+1+\frac{1}{b}}}$	$r \leq q - \frac{1}{2}$	General

Table 2.1: Comparison to upper rates available from earlier literature (for their applicability to the inverse learning setting considered in the present paper, see Section 2.2.5).

in the second column $\mathcal{H}_1 = \mathcal{H}_K$. The more complicated form given in Table 2.1 is the reinterpretation in our setting (see Section 2.2). The first three references ([79], [4], [93]) do not analyze lower bounds and their upper bounds do not take into account the behavior of the eigenvalues of the integral operator L corresponding to the assumed RKHS structure. But all three derive estimates on the error both in $L^2(\nu)$-norm and in RKHS-norm. Only [4] considers a general class of spectral regularization methods.

The last two papers [24] and [20] obtain fast upper rates (depending on the eigenvalues of L) which are minimax optimal. The estimates, however, are only given in $L^2(\nu)$-norm. Furthermore, only [20] goes beyond Tikhonov regularization to handle a general class of spectral regularization methods. A closer look at Table 2.1 reveals that in treating general spectral regularization methods, the results of [20] require for certain parameter configurations ($r < 1/2 - 1/2b$) the availability of additional unlabeled data from the sampling distribution ν. This appears somewhat suboptimal, since this does not reproduce the previously obtained result for Tikhonov in [24] which does not require unlabeled data.

After completion of this work, we became aware of the independent work of Dicker *et al.* [29], which has overlap with our results since they consider general regularization methods for regression using a reproducing kernel. We briefly compare the present contribution to that work: Because we are motivated by an inverse problem point of view, we derive convergence results for the reconstruction error (RKHS $\mathcal{H}_1-$ norm), while [29] only considers prediction error estimates ($L^2(\nu)-$ norm). We also establish the optimal dependence of the convergence rate in the important secondary parameters σ^2 (noise variance) and R (source condition or Sobolev radius). The authors in [29] give an upper bound depending explicitly on those parameters, but with suboptimal dependence. In this sense even for the $L^2(\nu)-$ norm alone

our contribution brings something novel, including a matching minimax lower bound for all considered norms with the correct dependence in σ^2 and R. Finally, our estimates hold for normalized moments of any order of the considered norms, and we also establish exponential deviation bounds, while [29] only considers expected squared $L^2(\nu)-$ norm (though under hypotheses weaker as ours concerning the noise).

We conclude this review by mentioning the recent work [60], which also concerns inverse statistical learning (see also [61]), albeit in a quite different setting. In that work, the main focus is on classification (Y only can take finitely many values or "classes"), and the inverse problem is that the sampling distribution for X is transformed via a linear operator A. The method analyzed there is empirical risk minimization using a modified loss which implicitly includes an estimation of the original class-conditional distributions from the transformed ones. In the present paper, we consider an (inverse) regression setting with a continuous output variable, the nature of the inverse problem is different since the transformation is applied to the regression function, and we also use a different methodological approach.

The main question addressed in this chapter is that of minimax optimal rates of convergence as n grows to infinity - first under the assumption of strictly polynomial eigenvalue decay of the covariance operator $\bar{B} = \bar{B}_\nu$ which is adjoint to the integral operator L defining the RKHS structure, see Section 2.2. Our contribution is to improve on and extend the existing results presented above, aiming to present a complete picture. We consider a unified approach which allows to simultaneously treat the direct and the inverse learning problem, derive upper bounds (non-asymptotic and asymptotic) as well as lower bounds, both for the $L^2(\nu)-$ and for the $\mathcal{H}_1-$ norm (as well as intermediate norms) for a general class of regularization methods, without requiring additional unlabeled data. In this generality, this is new. In addition, we present a refined analysis of (both strong and weak) minimax optimal rates also investigating their optimal dependence on radius parameter R of the source condition and on the variance σ^2 of the noise (our lower bounds come in slightly different strong and weak versions leading to the natural notion of weak and strong minimax optimality). To the best of our knowledge, this has never been done before.

We emphasize that all derivations of fast rates rely on tools introduced in the seminal work of [94], and depend in particular on the notion of *effective dimension* of the data with respect to the regularization parameter λ, defined as

$$\mathcal{N}(\lambda) := \mathrm{Tr}\left[(B_\nu + \lambda)^{-1} B_\nu\right] = \sum_{i \geq 1} \frac{\mu_{\nu,i}}{\mu_{\nu,i} + \lambda}\,. \tag{2.1.2}$$

Using this tool is essential for obtaining the finer results ("fast rates" taking into account the spectral structure of L) of this chapter since it determines the optimal choice of the regularization parameter. This idea of [24] and [20] is fundamental for our approach, which extends and refines these previous results.

Furthermore, we recall from [24] that the effective dimension $\mathcal{N}(\lambda)$ seems to be just the right parameter to establish an important connection between the operator theoretic and spectral methods and the results obtained via entropy methods (see [28], [86]) since $\mathcal{N}(\lambda)$ encodes via L crucial properties of the marginal distribution ν. However, this connection is not yet fully worked out and further progress in this direction is a challenge for future research aimed at establishing a unified picture.

The importance of the notion of effective dimension will continue to unfold throughout this thesis. In this chapter we will use it in an essential way in the proof of our upper rates: Using precisely polynomial eigenvalue decay (or, perhaps more accurately, restricting to the associated classes of sampling distributions introduced in (2.2.6) and (2.2.7)), we shall accurately compute the effective dimension and

thereby establish our (fast) upper rates (for the regular case). In subsequent chapters we shall control the effective dimension in some more general cases. This is a basic philosophy to improve on the results of this chapter. For the sake of completeness, we also mention at this point that this approach could be somewhat formalized: In one step, one derives minimax optimal rates under implicit assumptions on the effective dimension only (the problem of obtaining lower bounds in this setting is not yet completely solved), and in a second step one verifies these estimates on the effective dimension for certain model classes, defined by specific source conditions and classes of marginals (corresponding e.g. to certain types of eigenvalue decay of the covariance operator). In spirit, this is very much what is done in this thesis, but the approach is not yet completely formalized (at least partially due to the above mentioned problem for lower bounds in this setting).

The outline of the rest of the Chapter is as follows: In Section 2.2, we fix notation and describe our setting in more detail. In particular, we adopt the theory of Carleman operators from the direct problem to our more general setting, including the inverse learning problem. We describe the source conditions, the assumptions on the noise and prior classes, and finally the general class of spectral regularization methods. Granted these preliminaries, we then present in Section 2.3 our main results (Theorem 2.3.4, Theorem 2.3.5 and Corollary 2.3.6). In Section 2.4, we present a concluding discussion on some further aspects of the results. Section 2.5 contains the proofs of the upper bounds, Section 2.6 is devoted to the proof of lower bounds. In the Appendix we establish the concentration inequalities and a perturbation result needed in Section 2.5 and give some supplementary technical lemmata needed in Section 2.6.

2.2 Notation and Preliminaries

In this section, we specify the mathematical setting and assumptions for the model (2.1.1) and reduce it to an equivalent model.

2.2.1 Inverse Problems induced by Carleman Operators

We assume that the input space $\mathcal{X}$ is a standard Borel space endowed with a probability measure ν, and the output space $\mathcal{Y}$ is equal to $\mathbb{R}$. Let $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ be a linear operator, where $\mathcal{H}_1$ is an infinite-dimensional real separable Hilbert space and $\mathcal{H}_2$ some vector space of functions $g : \mathcal{X} \longrightarrow \mathbb{R}$. We do not assume any specific structure on $\mathcal{H}_2$ for now. However, as will become clear shortly, the image $\mathrm{Im}(A) \subset \mathcal{H}_2$ will be endowed with a natural Hilbert space structure as a consequence of the following key assumption:

Assumption 2.2.1. *The evaluation functionals at a given point $x \in \mathcal{X}$:*

$$\begin{aligned} S_x : \mathcal{H}_1 &\longrightarrow \mathbb{R} \\ f &\longmapsto (S_x)(f) := (Af)(x) \end{aligned}$$

are uniformly (w.r.t. $x \in \mathcal{X}$) bounded, i.e., there exists a constant $\kappa < \infty$ such that for any $x \in \mathcal{X}$

$$|S_x(f)| \;\leq\; \kappa \;\|f\|_{\mathcal{H}_1} \;.$$

For all x, the fact that S_x is continuous implies, by Riesz's representation theorem, the existence of an element $F_x \in \mathcal{H}_1$ such that

$$(Af)(x) = \langle f, F_x \rangle_{\mathcal{H}_1}$$

with

$$\|F_x\|_{\mathcal{H}_1} = \|S_x\| \leq \kappa \,,$$

for any $x \in \mathcal{X}$. Define the map

$$\begin{aligned} K : \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R} \\ (x_1, x_2) &\longmapsto K(x_1, x_2) := \langle F_{x_1}, F_{x_2} \rangle_{\mathcal{H}_1} \,, \end{aligned}$$

which is by construction a positive semidefinite (p.s.d.) kernel over $\mathcal{X}$ associated with the so-called feature space $\mathcal{H}_1$, and the feature map $F_\cdot : x \in \mathcal{X} \mapsto F_x \in \mathcal{H}_1$. Observe that for any $x \in \mathcal{X}$, we have the bound $K(x,x) = \|F_x\|^2_{\mathcal{H}_1} \leq \kappa^2$. A fundamental result (see [81], Theorem 4.21) is that to every p.s.d. kernel can be associated a unique reproducing kernel Hilbert space (RKHS). We reproduce this result here, adapted to the considered context:

Proposition 2.2.2. *(Unique RKHS associated with a p.s.d kernel) The real-valued function space*

$$\begin{aligned} \mathcal{H}_K &:= \{ g : \mathcal{X} \longrightarrow \mathbb{R} \mid \exists\, f \in \mathcal{H}_1 \text{ with } \; g(x) = \langle f, F_x \rangle_{\mathcal{H}_1} = (Af)(x) \; \forall x \in \mathcal{X} \} \\ &= \mathrm{Im}(A) \subset \mathcal{H}_2, \end{aligned}$$

equipped with the norm

$$\begin{aligned} \|g\|_{\mathcal{H}_K} &:= \inf \left\{ \|f\|_{\mathcal{H}_1} \;:\; f \in \mathcal{H}_1 \text{ s.t. } \forall x \in \mathcal{X} : g(x) = \langle f, F_x \rangle_{\mathcal{H}_1} = (Af)(x) \right\} \\ &= \inf_{f \in A^{-1}(\{g\})} \|f\|_{\mathcal{H}_1} \end{aligned}$$

is the unique RKHS for which K is a reproducing kernel. Moreover, the operator A is a partial isometry from $\mathcal{H}_1$ to $\mathcal{H}_K$ (i.e., an isometry on the orthogonal of its kernel), and

$$\mathcal{H}_K = \overline{\mathrm{Span}}\{ K(x, .), x \in \mathcal{X} \} \,.$$

From now on, we can therefore forget about the space $\mathcal{H}_2$ and consider A as an operator from $\mathcal{H}_1$ onto $\mathcal{H}_K = \mathrm{Im}(A)$. As a consequence of A being a partial isometry onto $\mathcal{H}_K$, note that this RKHS is separable, since we have assumed that $\mathcal{H}_1$ is. Additionally, we assume

Assumption 2.2.3. *For any $f \in \mathcal{H}_1$, the map $x \mapsto (Af)(x) = \langle f, F_x \rangle_{\mathcal{H}_1}$ is measurable.*

Equivalently, it is assumed that all functions $g \in \mathcal{H}_K$ are measurable. Furthermore, Assumption 2.2.1 implies that $\|Af\|_\infty \leq \kappa \|f\|_{\mathcal{H}_1}$ for all $f \in \mathcal{H}_1$, so that all functions in $\mathcal{H}_K$ are bounded in supremum norm. Therefore, $\mathcal{H}_K$ is a subset of $L^2(\mathcal{X}, \nu)$; let ι denote the associated canonical injection map $\mathcal{H}_K \hookrightarrow L^2(\mathcal{X}, \nu)$.

Together, Assumptions 2.2.3 and 2.2.1 thus imply that the map $F_\cdot : \mathcal{X} \longrightarrow \mathcal{H}_1$ is a bounded *Carleman*

map [44]. We define the associated *Carleman operator*, as

$$\begin{aligned} S_\nu : \mathcal{H}_1 &\longrightarrow L^2(\mathcal{X}, \nu) \\ f &\longmapsto S_\nu f := \iota(Af)\,. \end{aligned}$$

The operator S_ν is bounded and satisfies $\|S_\nu\| \leq \kappa$, since

$$\|S_\nu f\|^2_{L^2(\nu)} = \int_{\mathcal{X}} |(Af)(x)|^2 \, \nu(dx) = \int_{\mathcal{X}} |\langle f, F_x\rangle_{\mathcal{H}_1}|^2 \, \nu(dx) \leq \kappa^2 \|f\|^2_{\mathcal{H}_1} \ .$$

To illustrate the general setting more concretely, we give as an example a classical family of integral operators satisfying the above assumption, where the kernel K is completely explicit.

Example 2.2.4 (Integral operators)**.** *Let $(\mathcal{Z}, \mathfrak{Z}, \mu)$ be a measured space. Assume $\mathcal{H}_1 = L^2(\mathcal{Z}, \mu)$ and the operator we consider is given by*

$$[Af](x) = \int_{\mathcal{Z}} \varphi(x, z) f(z) \, d\mu(z) \ , \qquad x \in \mathcal{X} \ ,$$

where φ is a known measurable function $\mathcal{X} \times \mathcal{Z} \to \mathbb{R}$. Then we have

$$[Af](x) \leq \|f\|_{2,\mu} \, \|\varphi(x, .)\|_{2,\mu} \, ,$$

so that Assumption 2.2.1 is satisfied iff $\sup_{x\in\mathcal{X}} \|\varphi(x, .)\|_{2,\mu} < \infty$. In this case, since $[Af](x) = \langle f, F_x\rangle_{\mathcal{H}_1}$ holds for any $f \in \mathcal{H}_1$, it follows $F_x = \varphi(x, .)$ and

$$K(x, x') = \langle F_x, F_{x'}\rangle_{\mathcal{H}_1} = \int_{\mathcal{Z}} \varphi(x, z)\varphi(x', z) \, d\mu(z).$$

The two next examples are classical particular cases of the above.

Example 2.2.5 (Deconvolution)**.** *One of the most standard inverse problems is that of deconvolution. We let $\mathcal{H}_1 = L^2([0, 1], dt)$ (with dt the Lebesgue measure) and the operator we consider be given by*

$$[Af](x) = \int_0^1 f(t)\varphi(x - t) \, dt \ , \qquad x \in \mathbb{R} \ ,$$

where φ is a known filter belonging to $C^k_0(\mathbb{R})$, the space of k times continuously differentiable functions on the real line with compact support. Then it is clear that A maps into $\mathcal{H}_2 = C^k(\mathbb{R})$ (see e.g. [47], Theorem 1.3.1). It can be easily checked that Assumption 2.2.1 is fulfilled and $[Af](x) = \langle f, F_x\rangle_{\mathcal{H}_1}$ for any $f \in \mathcal{H}_1$ with $F_x = \varphi(x - \cdot)$. The kernel of $\mathrm{Im}(A)$ can explicitly calculated as

$$K(x, y) = \langle F_x, F_y\rangle_{\mathcal{H}_1} = \int_0^1 \varphi(x - t)\varphi(y - t) \, dt \ .$$

Example 2.2.6. *(Differentiating a real function) We consider estimation of a derivative of a real function. To this end, we let $\mathcal{H}_1 := \{f \in L^2[0, 1] : \ \mathbb{E}[f] = 0\}$, the subspace of $L^2([0, 1], dt)$ consisting of functions with mean zero and $\mathcal{H}_2 := C[0, 1]$, the space of continuous functions on $[0, 1]$. Define $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ by*

$$[Af](x) = \int_0^x f(t) \, dt \ .$$

Then $Af = g$ if and only if $f = g'$. It is easily checked that Assumption 2.2.1 is satisfied. To identify the kernel of $\mathrm{Im}(A)$*, the reader can easily convince himself that*

$$[Af](x) = \langle f, F_x \rangle_{L^2} \ ,$$

where $F_x(t) = \mathbb{1}_{[0,x]}(t) - x$. Thus, by definition $K(x,t) = \langle F_x, F_t \rangle_{L^2} = x \wedge t - xt$ and $\mathrm{Im}(A)$ *coincides with the real Sobolev space $H_0^1[0,1]$, consisting of absolutely continuous functions g on $[0,1]$ with weak derivatives of order 1 in $L^2[0,1]$, with boundary condition $g(0) = g(1) = 0$. The associated Carleman operator is given by $S = \iota \circ A$ with $\iota : H_0^1[0,1] \hookrightarrow L^2[0,1]$ and with marginal distribution $\nu = dt$, the Lebesgue measure on $[0,1]$.*

We complete this section by introducing the shortcut notation $\bar{F}_x := \kappa^{-1} F_x$, $\bar{S}_\nu := \kappa^{-1} S_\nu$, $\bar{S}_\nu^\star := \kappa^{-1} S_\nu^\star$. We define $B_\nu := S_\nu^\star S_\nu : \mathcal{H}_1 \longrightarrow \mathcal{H}_1$ and $\bar{B}_\nu := \kappa^{-2} B_\nu$. Then $\bar{B}_\nu$ is positive, self-adjoint and satisfies $\left\|\bar{B}_\nu\right\| \leq 1$. The following Proposition summarizes the main properties of the operators $S_\nu, S_\nu^\star$ and B_ν. Its proof can be found in the Appendix of [26] (Proposition 19).

Proposition 2.2.7. *Under Assumptions 2.2.1 and 2.2.3, the Carleman operator $S_\nu : \mathcal{H}_1 \longrightarrow L^2(\mathcal{X}, \nu)$ is a Hilbert-Schmidt operator with nullspace*

$$\ker(S_\nu) = \mathrm{Span}\{F_x \ : \ x \in \mathrm{support}(\nu)\}^\perp \ .$$

The adjoint operator $S_\nu^\star : L^2(\mathcal{X}, \nu) \longrightarrow \mathcal{H}_1$ is given by

$$S_\nu^\star g = \int_{\mathcal{X}} g(x) F_x \ \nu(dx) \ ,$$

for any $g \in L^2(\mathcal{X}, \nu)$ and where the integral converges in $\mathcal{H}_1$- norm.
Furthermore, if $F_x \otimes F_x^\star$ denotes the operator $f \in \mathcal{H}_1 \mapsto \langle f, F_x \rangle_{\mathcal{H}_1} F_x \in \mathcal{H}_1$, then

$$B_\nu = \int_{\mathcal{X}} F_x \otimes F_x^\star \ \nu(dx) \ , \tag{2.2.1}$$

where the integral converges in trace norm.

It is natural to consider the inverse problem $S_\nu f = g$ (rather than $Af = g$) as the idealized population version (i.e., noise and discretization-free) of (2.1.1), since the former views the output of the operator in the geometry of $L^2(\mathcal{X}, \nu)$, which is the natural population geometry when the sampling measure is ν. Multiplying on both sides by $S_\nu^\star$, we obtain the inverse problem $B_\nu f = S_\nu^\star g$ (called "normal equation" in the inverse problem literature).

Since B_ν is self-adjoint and compact, the spectral theorem ensures the existence of an orthonormal set $\{e_j\}_{j \geq 1}$ such that

$$B_\nu = \sum_{j=1}^{\infty} \mu_j \langle \cdot, e_j \rangle_{\mathcal{H}_1} e_j \tag{2.2.2}$$

and

$$\mathcal{H}_1 = \ker(B_\nu) \oplus \overline{\mathrm{Span}}\{e_j \ : \ j \geq 1\} \ .$$

The numbers μ_j are the positive eigenvalues of B_ν in decreasing order, satisfying $0 < \mu_{j+1} \leq \mu_j$ for all $j > 0$ and $\mu_j \searrow 0$. In the special case where B_ν has finite rank, the above set of positive eigenvalues

and eigenvectors is finite, but to simplify the notation we will always assume that they are countably infinite; formally, we can accommodate for this special situation by allowing that the decreasing sequence of eigenvalues is equal to zero from a certain index on.

Remark 2.2.8. *The considered operators depend on the sampling measure ν and thus also the eigenvalues $(\mu_j)_{j\geq 1}$. For the sake of reading ease, we omit this dependence in the notation; we will also denote henceforth $S = S_\nu$, $\bar{S} = \bar{S}_\nu$, $B = B_\nu$ and $\bar{B} = \bar{B}_\nu$.*

2.2.2 Discretization by random sampling

For discretization, we consider a sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), ..., (x_n, y_n)) \in (\mathcal{X} \times \mathbb{R})^n$ and introduce the associated *sampling operator*

$$\begin{aligned} S_{\mathbf{x}} : \mathcal{H}_1 &\longrightarrow \mathbb{R}^n \\ f &\longmapsto S_{\mathbf{x}} f \,, \end{aligned}$$

with $(S_{\mathbf{x}} f)_j = \langle f, F_{x_j} \rangle_{\mathcal{H}_1}$, $j = 1, ..., n$ and where $\mathbb{R}^n$ is equipped with the inner product of the empirical L^2 structure,

$$\langle \mathbf{y}, \mathbf{y}' \rangle_{R^n} = \frac{1}{n} \sum_{j=1}^{n} y_j y'_j \,.$$

Formally, $S_{\mathbf{x}}$ is the counterpart of S_ν when replacing the sampling distribution ν by the empirical distribution $\widehat{\nu} := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$, and identifying $L^2(\mathcal{X}, \widehat{\nu})$ with $\mathbb{R}^n$ endowed with the above inner product. Additionally, the sampled vector $S_{\mathbf{x}} f$ is corrupted by noise $\boldsymbol{\varepsilon} = (\epsilon_1, \ldots, \epsilon_n)$ to yield the vector of observed values $\mathbf{y} = (y_1, ..., y_n) \in \mathbb{R}^n$:

$$y_j = g(x_j) + \varepsilon_j = (S_{\mathbf{x}} f)_j + \varepsilon_j, \;\; j = 1, ..., n \,, \tag{2.2.3}$$

which can be interpreted as the discretized and noisy counterpart of the population problem $S_\nu f = g$. Replacing the measure ν with the empirical measure $\widehat{\nu}$ in Proposition 2.2.7 gives the following Corollary:

Corollary 2.2.9. *The sampling operator $S_{\mathbf{x}} : \mathcal{H}_1 \longrightarrow \mathbb{R}^n$ is a Hilbert-Schmidt operator with nullspace*

$$\ker(S_{\mathbf{x}}) = \mathrm{Span}\{F_{x_j} \;:\; j = 1, ..., n\}^\perp \,.$$

Furthermore, the adjoint operator $S^\star_{\mathbf{x}} : \mathbb{R}^n \longrightarrow \mathcal{H}_1$ is given by

$$S^\star_{\mathbf{x}} \mathbf{y} = \frac{1}{n} \sum_{j=1}^{n} y_j F_{x_j} \,,$$

and the operator $B_{\mathbf{x}} := S^\star_{\mathbf{x}} S_{\mathbf{x}} : \mathcal{H}_1 \longrightarrow \mathcal{H}_1$ is given by

$$B_{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^{n} F_{x_j} \otimes F^\star_{x_j} \,.$$

With this notation, the normal equation associated with (2.2.3), obtained by multiplying both sides by $S^\star_{\mathbf{x}}$, reads $S^\star_{\mathbf{x}} \mathbf{y} = B_{\mathbf{x}} f + S^\star_{\mathbf{x}} \boldsymbol{\epsilon}$; it is the discretized and noisy counterpart of the population normal equation

introduced in the previous section. The advantage of looking at the normal equations is that both the population and the empirical version act on the same space $\mathcal{H}_1$, so that the latter can be considered as a perturbation of the former (both for the operator and the noise term), an observation which is central to the theory [27].

Similarly, as for the population operators we introduce $\bar{S}_{\mathbf{x}} := \kappa^{-1} S_{\mathbf{x}}$, $\bar{S}^{\star}_{\mathbf{x}} := \kappa^{-1} S^{\star}_{\mathbf{x}}$ and $\bar{B}_{\mathbf{x}} := \kappa^{-2} B_{\mathbf{x}}$.

2.2.3 Statistical model, noise assumption, and prior classes

We recall the considered setting of inverse learning. The sampling is assumed to be random i.i.d., where each observation point (X_i, Y_i) follows the model $Y = Af(X) + \varepsilon$. More precisely, (X_i, Y_i) are i.i.d. with Borel probability distribution ρ on $\mathcal{X} \times \mathbb{R}$. For (X, Y) having distribution ρ, denoting ν the marginal distribution of X, we assume:

Assumption 2.2.10. *The conditional expectation w.r.t. ρ of Y given X exists and it holds for ν-almost all $x \in X$:*

$$\mathbb{E}_{\rho}[Y|X = x] = \bar{S}_x f_{\rho}\,, \text{ for some } f_{\rho} \in \mathcal{H}_1\,. \tag{2.2.4}$$

Furthermore, we will make the following Bernstein-type assumption on the observation noise distribution:

Assumption 2.2.11. *There exists $\sigma > 0$ and $M > 0$ such that for any integer $m \geq 2$:*

$$\mathbb{E}[\,\left|Y - \bar{S}_X f_{\rho}(X)\right|^m \mid X\,] \leq \frac{1}{2} m!\, \sigma^2 M^{m-2} \quad \nu - \text{a.s.} \tag{2.2.5}$$

It is a generally established fact that given any estimator $\widehat{f}$ of f_{ρ}, one can construct a probability measure ρ on $\mathcal{X} \times \mathbb{R}$ such that the rate of convergence of $\widehat{f}$ to f_{ρ} can arbitrarily be slow (see, e.g., [43]). Thus, to derive non-trivial rates of convergence, we concentrate our attention on specific subsets (also called *models*) of the class of probability measures. We will work with the same type of assumptions as considered by [24] and introduce two sets of conditions concerning, on the one hand, the marginal distribution ν of X, and on the other hand, the conditional distribution $\rho(.|.)$ of Y given X.

Let $\mathcal{P}$ denote the set of all probability distributions on $\mathcal{X}$. We define classes of sampling distributions by introducing decay conditions on the eigenvalues μ_i of the operator $\bar{B}_{\nu}$ defined in Section 2.2.1.

For $b > 1$ and $\alpha, \beta > 0$, we define

$$\mathcal{P}^{<}(b, \beta) := \{\nu \in \mathcal{P} :\ \mu_j \leq \beta / j^b \ \ \forall j \geq 1\}\,, \tag{2.2.6}$$

$$\mathcal{P}^{>}(b, \alpha) := \{\nu \in \mathcal{P} :\ \mu_j \geq \alpha / j^b \ \ \forall j \geq 1\} \tag{2.2.7}$$

and

$$\mathcal{P}^{>}_{strong}(b, \alpha) := \{\nu \in \mathcal{P}^{>}(b, \alpha) :\ \exists \gamma > 0\,, j_0 \geq 1 \text{ s.th. } \frac{\mu_{2j}}{\mu_j} \geq 2^{-\gamma} \ \ \forall j \geq j_0\}\,.$$

In the inverse problem literature, such eigenvalue decay assumptions are related to the so-called degree of ill-posedness of the inverse problem $B_{\nu} f = S^{\star} g$. In the present setting, the ill-posedness of the problem is reflected by the eigenvalues of B_{ν} and depends on both the fixed operator A and the sampling distribution

ν.

Example 2.2.12. *Coming back to our Example 2.2.6, the degree of ill-posedness is determined by the decay of the eigenvalues $(\mu_j)_j$ of the positive self-adjoint integral operator $L_K = SS^\star : L^2[0,1] \longrightarrow L^2[0,1]$*

$$[L_K f](x) = \int_0^1 K(x,t) f(t)\, dt\ .$$

Elementary calculations show that the SVD basis is given by $e_j(x) = \sqrt{2}\sin(\pi j x)$ with corresponding eigenvalues $\mu_j = \frac{1}{\pi^2 j^2}$. Thus, $b = 2$ and $\mathcal{P}^<(2, \frac{1}{\pi^2}) \cap \mathcal{P}^>(2, \frac{1}{\pi^2})$ as well as $\mathcal{P}^<(2, \frac{1}{\pi^2}) \cap \mathcal{P}^>_{strong}(2, \frac{1}{\pi^2})$ are not empty.

For a subset $\Omega \subseteq \mathcal{H}_1$, we let $\mathcal{K}(\Omega)$ be the set of regular conditional probability distributions $\rho(\cdot|\cdot)$ on $\mathcal{B}(\mathbb{R}) \times \mathcal{X}$ such that (2.2.4) and (2.2.5) hold for some $f_\rho \in \Omega$. (It is clear that these conditions only depend on the conditional $\rho(.|.)$ of Y given X.) We will focus on a *Hölder-type source condition*, which is a classical smoothness assumption in the theory of inverse problems. Given $r > 0, R > 0$ and $\nu \in \mathcal{P}$, we define

$$\Omega_\nu(r,R) := \{f \in \mathcal{H}_1 : f = \bar{B}_\nu^r h,\ \|h\|_{\mathcal{H}_1} \leq R\}. \tag{2.2.8}$$

Note that for any $r \leq r_0$ we have $\Omega_\nu(r_0,R) \subseteq \Omega_\nu(r,R)$, for any $\nu \in \mathcal{P}$. Since B_ν is compact, the source sets $\Omega_\nu(r,R)$ are precompact sets in $\mathcal{H}_1$.

Then the class of models which we will consider will be defined as

$$\mathcal{M}(r,R,\mathcal{P}') \ :=\ \{\ \rho(dx,dy) = \rho(dy|x)\nu(dx)\ :\ \rho(\cdot|\cdot) \in \mathcal{K}(\Omega_\nu(r,R)),\ \nu \in \mathcal{P}'\ \}\ , \tag{2.2.9}$$

with $\mathcal{P}' = \mathcal{P}^<(b,\beta)$, $\mathcal{P}' = \mathcal{P}^>(b,\alpha)$ or $\mathcal{P}' = \mathcal{P}^>_{strong}(b,\alpha)$.

As a consequence, the class of models depends not only on the smoothness properties of the solution (reflected in the parameters $R > 0,\ r > 0$), but also essentially on the decay of the eigenvalues of B_ν.

2.2.4 Effective Dimension

We introduce the *effective dimension* $\mathcal{N}(\lambda)$, appearing in [24] in a similar context. For $\lambda \in (0,1]$ we set

$$\mathcal{N}(\lambda) = \mathrm{Tr}\left[\ (\bar{B} + \lambda)^{-1}\bar{B}\ \right]\ . \tag{2.2.10}$$

Since by Proposition 2.2.7 the operator $\bar{B}$ is trace-class, $\mathcal{N}(\lambda) < \infty$. Moreover, the following Lemma (see [24], Proposition 3) establishes a connection between the spectral asymptotics of the covariance operator $\bar{B}$ and an upper bound for $\mathcal{N}(\lambda)$.

Lemma 2.2.13. *Assume that the marginal distribution ν of X belongs to $\mathcal{P}^<(b,\beta)$ (with $b > 1$ and $\beta > 0$). Then the effective dimension $\mathcal{N}(\lambda)$ satisfies*

$$\mathcal{N}(\lambda) \leq \frac{\beta b}{b-1}(\kappa^2\lambda)^{-\frac{1}{b}}\ .$$

Furthermore, for $\lambda \leq ||\bar{B}||$, since $\bar{B}$ is positive

$$\mathcal{N}(\lambda) = \sum_{\mu_j \geq \kappa^2\lambda} \frac{\mu_j}{\mu_j + \kappa^2\lambda} + \sum_{\mu_j < \kappa^2\lambda} \frac{\mu_j}{\mu_j + \kappa^2\lambda} \geq \min_{\mu_j \geq \kappa^2\lambda} \left\{ \frac{\mu_j}{\mu_j + \kappa^2\lambda} \right\} \geq \frac{1}{2} ,$$

since the first sum has at least one term.

2.2.5 Equivalence with classical kernel learning setting

With the notation and setting introduced in the previous sections, we point out that the "inverse learning" problem (2.1.1) can, provided Assumptions (2.2.1) and (2.2.3) are met, be reduced to a classical learning problem (hereafter called "direct" learning) under the setting and assumptions of reproducing kernel-based estimation methods. In the direct learning setting, the model is given by (2.1.1) (i.e., $Y_i = g(X_i) + \varepsilon_i$) and the goal is to estimate the function g. Kernel methods *posit* that g belongs to some reproducing kernel Hilbert space[1] $\mathcal{H}_K$ with kernel K and construct an estimate $\widehat{g} \in \mathcal{H}_K$ of g based on the observed data. The reconstruction error $(\widehat{g} - g)$ can be analyzed in $L^2(\nu)-$ norm or in $\mathcal{H}_K-$ norm.

Coming back to the inverse learning setting ($Y_i = (Af)(X_i) + \varepsilon_i$), let $\mathcal{H}_K$ be defined as in the previous sections and assume $f \in \ker(A)^\perp$ (we cannot hope to recover the part of f belonging to $\ker A$ anyway and might as well make this assumption. It is also implied by any form of source condition as introduced in Section 2.2.3).

Consider applying a direct learning method using the reproducing kernel K; this returns some estimate $\widehat{g} \in \mathcal{H}_K$ of g. Now let $A^\dagger$ be the inverse of $A|_{\mathcal{K}er(A)^\perp}$, which is well defined since A is a partial isometry as an operator $\mathcal{H}_1 \mapsto \mathcal{H}_K$ (Proposition 2.2.2). Defining $\widehat{f} := A^\dagger \widehat{g}$, we have

$$\left\| \widehat{f} - f \right\|_{\mathcal{H}_1}^2 = \left\| A^\dagger \widehat{g} - f \right\|_{\mathcal{H}_1}^2 = \left\| A(A^\dagger \widehat{g} - f) \right\|_{\mathcal{H}_K}^2 = \|\widehat{g} - Af\|_{\mathcal{H}_K}^2 = \|\widehat{g} - g\|_{\mathcal{H}_K}^2 .$$

Note that $\widehat{f}$ is, at least in principle, accessible to the statistician, since A (and therefore $A^\dagger$) is assumed to be known. Hence, a bound established for the direct learning setting in the sense of the $\mathcal{H}_K-$ norm reconstruction $\|\widehat{g} - g\|_{\mathcal{H}_K}^2$ also applies to the inverse problem reconstruction error $\left\| \widehat{f} - f \right\|_{\mathcal{H}_1}^2$. Furthermore, it is easy to see that the eigenvalue decay conditions and the source conditions involving the operator B_ν introduced in Section 2.2.3 are, via the same isometry, equivalent to similar conditions involving the kernel integral operator in the direct learning setting, as considered, for instance, in [4, 20, 21, 24, 79]. It follows that estimates in $\mathcal{H}_K-$ norm available from those references are directly applicable to the inverse learning setting. However, as summarized in Table 2.1, for the direct learning problem the results concerning $\mathcal{H}_K-$ norm rates of convergence are far less complete than in $L^2(\nu)-$ norm. In particular, such rates have not been established under consideration of simultaneous source and eigenvalue decay conditions, and neither have the corresponding lower bounds. In this sense, the contribution of the present paper is to complete the picture in Table 2.1, with the inverse learning setting as the underlying motivation.

[1] This can be extended to the case where g is only approximated in $L^2(\nu)$ by a sequence of functions in $\mathcal{H}_K$. For the sake of the present discussion, only the case where it is assumed $g \in \mathcal{H}_K$ is of interest.

2.2.6 Regularization

In this section, we introduce the class of linear regularization methods based on spectral theory for self-adjoint linear operators. These are standard methods for finding stable solutions for ill-posed inverse problems, see, e.g., [34] or [38].

Definition 2.2.14 (**Regularization function**). *Let $g : (0,1] \times [0,1] \longrightarrow \mathbb{R}$ be a function and write $g_\lambda = g(\lambda, \cdot)$. The family $\{g_\lambda\}_\lambda$ is called regularization function, if the following conditions hold:*

(i) There exists a constant $D < \infty$ such that

$$\sup_{0<t\leq 1} |t g_\lambda(t)| \leq D,$$

for any $0 < \lambda \leq 1$.

(ii) There exists a constant $E < \infty$ such that

$$\sup_{0<t\leq 1} |g_\lambda(t)| \leq \frac{E}{\lambda}, \tag{2.2.11}$$

for any $0 < \lambda \leq 1$.

(iii) Defining the residual

$$r_\lambda(t) = 1 - g_\lambda(t)t\,, \tag{2.2.12}$$

there exists a constant $\gamma_0 < \infty$ such that

$$\sup_{0<t\leq 1} |r_\lambda(t)| \leq \gamma_0, \tag{2.2.13}$$

for any $0 < \lambda \leq 1$.

Definition 2.2.15 (**Qualification**). *The qualification of the regularization $\{g_\lambda\}_\lambda$ is the maximal q such that for any $0 < \lambda \leq 1$*

$$\sup_{0<t\leq 1} |r_\lambda(t)| t^q \leq \gamma_q \lambda^q.$$

for some constant $\gamma_q > 0$.

The next lemma provides a simple inequality (see, e.g., [64], Proposition 3) that shall be used later.

Lemma 2.2.16. Let $\{g_\lambda\}_\lambda$ be a regularization function with qualification q. Then, for any $r \leq q$ and $0 < \lambda \leq 1$:

$$\sup_{0<t\leq 1} |r_\lambda(t)| t^r \leq \gamma_r \lambda^r, \tag{2.2.14}$$

where $\gamma_r := \gamma_0^{1-\frac{r}{q}} \gamma_q^{\frac{r}{q}}$.

Given the sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (\mathcal{X} \times \mathbb{R})^n$, we define the regularized approximate solution $f_{\mathbf{z}}^\lambda$ of problem (2.2.3), for a suitable *a priori* parameter choice $\lambda = \lambda_n$, by

$$f_{\mathbf{z}}^{\lambda_n} := g_{\lambda_n}(\bar{B}_{\mathbf{x}}) \bar{S}_{\mathbf{x}}^\star \mathbf{y}\,. \tag{2.2.15}$$

Note that $g_\lambda(\bar{B}_{\mathbf{x}})$ is well defined since $\left\|\bar{B}_{\mathbf{x}}\right\| \leq 1$.

Remark 2.2.17. *(On the explicit calculation of the estimator) Denoting by*

$$\bar{\mathbb{K}}_{\mathbf{x}} = \kappa^{-2}(\frac{1}{n}K(x_i, x_j))_{i,j}$$

the (normalized) kernel matrix and letting $\alpha = (\alpha_1, ..., \alpha_n) = g_{\lambda_n}(\bar{\mathbb{K}}_{\mathbf{x}})\mathbf{y} \in \mathbb{R}^n$*, we have by recalling the formula* $g(A^*A)A^* = A^*g(AA^*)$ *(holding for bounded operators* A *and measurable functions* g *on the spectrum of* A*) and by Corollary 2.2.9*

$$f_{\mathbf{z}}^{\lambda_n} = S_{\mathbf{x}}^{\star} g_{\lambda_n}(\bar{\mathbb{K}}_{\mathbf{x}})\mathbf{y} = \frac{1}{n}\sum_{j=1}^{n} \alpha_j F_{x_j} \ . \tag{2.2.16}$$

Thus the estimator $f_{\mathbf{z}}^{\lambda_n}$ *for the target function can be calculated using the elements* $F_{x_1}, ..., F_{x_n}$ *(which, in turn, determine* $\bar{\mathbb{K}}_{\mathbf{x}}$*, and then the coefficients* $(\alpha_1, ..., \alpha_n)$*) . For the integral operators of a general form given in Example 2.2.4, this has been made completely explicit. In general, explicit formulae of course depend on the operator* A *and the inner product on* $\mathcal{H}_1$*.*

We now consider a different example setting. Assume now that $\mathcal{H}_1$ *also is an RKHS consisting of functions on some measurable space* $\mathcal{Z}$ *with a known, measurable kernel* G*. Then one finds*

$$F_x(z) = \langle F_x, G(z, \cdot)\rangle_{\mathcal{H}_1} = [AG(z, \cdot)](x) \ , \qquad x \in \mathcal{X} \ , z \in \mathcal{Z}$$

and

$$K(x, x') = \langle F_x, F_{x'}\rangle_{\mathcal{H}_1} = [AF_x](x') \ , \qquad x, x' \in \mathcal{X} \ .$$

Given the operator A*, this is completely explicit and requires only forward applications of* A*. In practice, if no closed-form formula can be derived, and since evaluation of* F_x *at a point* z *requires application of* A *to the test function* $G(z, \cdot)$*, a numerical approximation of* AF_x *might include appropriate discretization of* $\mathcal{Z}$*,* F_x *and* A*.*

We close this section by giving some examples which are common both in classical inverse problems [34] and in learning theory [4].

Example 2.2.18. (Spectral Cut-off) *A very classical regularization method is spectral cut-off (or truncated singular value decomposition), defined by*

$$g_\lambda(t) = \begin{cases} \frac{1}{t} & \text{if } t \geq \lambda \\ 0 & \text{if } t < \lambda \ . \end{cases}$$

In this case, $D = E = \gamma_0 = \gamma_q = 1$*. The qualification* q *of this method can be arbitrary.*

Example 2.2.19. (Tikhonov Regularization) *The choice* $g_\lambda(t) = \frac{1}{\lambda + t}$ *corresponds to Tikhonov regularization. In this case we have* $D = E = \gamma_0 = 1$*. The qualification of this method is* $q = 1$ *with* $\gamma_q = 1$*.*

Example 2.2.20. (Landweber Iteration) *The Landweber iteration (gradient descent algorithm with constant step size) is defined by*

$$g_k(t) = \sum_{j=0}^{k-1}(1-t)^j \quad \textit{with } k = 1/\lambda \in \mathbb{N}\,.$$

We have $D = E = \gamma_0 = 1$. The qualification q of this algorithm can be arbitrary with $\gamma_q = 1$ if $0 < q \leq 1$ and $\gamma_q = q^q$ if $q > 1$. The coefficients $(\alpha_j)_j$ in (2.2.16) can be calculated by using the following algorithm:

$\alpha_0 = 0 \in \mathbb{R}^n$
for i= $1, ..., k-1$
$\alpha_i = \alpha_{i-1} + \frac{2}{n}(\mathbf{y} - \mathbb{K}_{\mathbf{x}}\alpha_{i-1})$.

Example 2.2.21. (ν- method) *The $\nu-$ method belongs to the class of so called semi-iterative regularization methods. This method has finite qualification $q = \nu$ with γ_q a positive constant. Moreover, $D = 1$ and $E = 2$. The filter is given by $g_k(t) = p_k(t)$, a polynomial of degree $k-1$, with regularization parameter $\lambda \sim k^{-2}$, which makes this method much faster as e.g. gradient descent. The coefficients $(\alpha_j)_j$ in (2.2.16) can be calculated by using the following algorithm:*

$\alpha_0 = 0 \in \mathbb{R}^n$
$\omega_1 = \frac{4\nu+2}{4\nu+1}$
$\alpha_1 = \alpha_0 + \frac{\omega_1}{n}(\mathbf{y} - \mathbb{K}_{\mathbf{x}}\alpha_0)$
for i= $2, ..., k-1$
$\alpha_i = \alpha_{i-1} + u_i(\alpha_{i-1} - \alpha_{i-2}) + \frac{\omega_i}{n}(\mathbf{y} - \mathbb{K}_{\mathbf{x}}\alpha_{i-1})$
$u_i = \frac{(i-1)(2i-3)(2i+2\nu-1)}{(i+2\nu-1)(2i+4\nu-1)(2i+2\nu-3)}$
$\omega_i = 4\frac{(2i+2\nu-1)(i+\nu-1)}{(i+2\nu-1)(2i+4\nu-1)}$.

For more details concerning the derivation we refer in particular to [34] .

2.3 Main results: upper and lower bounds on convergence rates

Before stating our main results, we recall some basic definitions in order to clarify what we mean by asymptotic upper rate, lower rate and minimax rate optimality. We want to track the precise behavior of these rates not only for what concerns the exponent in the number of examples n, but also in terms of their scaling (multiplicative constant) as a function of some important parameters (namely the noise variance σ^2 and the complexity radius R in the source condition). For this reason, we introduce a notion of a family of rates over a family of models. More precisely, in all the forthcoming definitions, we consider an indexed family $(\mathcal{M}_\theta)_{\theta\in\Theta}$, where for all $\theta \in \Theta$, $\mathcal{M}_\theta$ is a class of Borel probability distributions on $\mathcal{X}\times\mathbb{R}$ satisfying the basic general assumption 2.2.10. We consider rates of convergence in the sense of the p-th moments of the estimation error, where $p > 0$ is a fixed real number.

Definition 2.3.1. *(Upper Rate of Convergence)*
A family of sequences $(a_{n,\theta})_{(n,\theta)\in\mathbb{N}\times\Theta}$ of positive numbers is called **upper rate of convergence in L^p**

for the interpolation norm of parameter $s \in [0, \frac{1}{2}]$ *, over the family of models* $(\mathcal{M}_\theta)_{\theta\in\Theta}$ *, for the sequence of estimated solutions* $(f_{\mathbf{z}}^{\lambda_{n,\theta}})_{(n,\theta)\in\mathbb{N}\times\Theta}$ *, using regularization parameters* $(\lambda_{n,\theta})_{(n,\theta)\in\mathbb{N}\times\Theta}$ *, if*

$$\sup_{\theta\in\Theta} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_\theta} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\|\bar{B}_\nu^s(f_\rho - f_{\mathbf{z}}^{\lambda_{n,\theta}})\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}}}{a_{n,\theta}} < \infty\,.$$

Definition 2.3.2. *(Weak and Strong Minimax Lower Rate of Convergence)*
A family of sequences $(a_{n,\theta})_{(n,\theta)\in\mathbb{N}\times\Theta}$ *of positive numbers is called* **weak minimax lower rate of convergence in** L^p *for the interpolation norm of parameter* $s \in [0, \frac{1}{2}]$ *, over the family of models* $(\mathcal{M}_\theta)_{\theta\in\Theta}$ *, if*

$$\inf_{\theta\in\Theta} \limsup_{n\to\infty} \inf_{f_\bullet} \sup_{\rho\in\mathcal{M}_\theta} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}_\nu^s(f_\rho - f_{\mathbf{z}})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}}}{a_{n,\theta}} > 0\,,$$

where the infimum is taken over all estimators, i.e., measurable mappings $f_\bullet : (\mathcal{X}\times\mathbb{R})^n \longrightarrow \mathcal{H}_1$ *. It is called a* **strong minimax lower rate of convergence in** L^p *if*

$$\inf_{\theta\in\Theta} \liminf_{n\to\infty} \inf_{f_\bullet} \sup_{\rho\in\mathcal{M}_\theta} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}_\nu^s(f_\rho - f_{\mathbf{z}})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}}}{a_{n,\theta}} > 0\,.$$

The difference between weak and strong lower rate can be summarily reformulated in the following way: If r_n denotes the minimax risk, using n observations, then $a_n = \mathcal{O}(r_n)$ must hold if a_n is a strong lower rate, while a_n being a weak lower rate means that $r_n = o(a_n)$ is excluded.

Definition 2.3.3. *(Minimax Optimal Rate of Convergence)*
The sequence of estimated solutions $(f_{\mathbf{z}}^{\lambda_{n,\theta}})_n$ *using the regularization parameters* $(\lambda_{n,\theta})_{(n,\theta)\in\mathbb{N}\times\Theta}$ *is called* **weak/strong minimax optimal in** L^p *for the interpolation norm of parameter* $s \in [0, \frac{1}{2}]$*, over the model family* $(\mathcal{M}_\theta)_{\theta\in\Theta}$*, with* **rate of convergence** *given by the sequence* $(a_{n,\theta})_{(n,\theta)\in\mathbb{N}\times\Theta}$*, if the latter is a weak/strong minimax lower rate as well as an upper rate for* $(f_{\mathbf{z}}^{\lambda_{n,\theta}})_{n,\theta}$*.*

We now formulate our main theorems.

Theorem 2.3.4 (Upper rate)**.** *Consider the model* $\mathcal{M}_{\sigma,M,R} := \mathcal{M}(r, R, \mathcal{P}^<(b,\beta))$ *(as defined in Section 2.2.3), where* $r > 0$*,* $b > 1$ *and* $\beta > 0$ *are fixed, and* $(R, M, \sigma) \in \mathbb{R}_+^3$ *(remember that* (σ, M) *are the parameters in the Bernstein moment condition* (2.2.5)*, in particular* σ^2 *is a bound on the noise variance). Given a sample* $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (\mathcal{X}\times\mathbb{R})^n$*, define* $f_{\mathbf{z}}^\lambda$ *as in* (2.2.15)*, using a regularization function of qualification* $q \geq r + s$*, with the parameter sequence*

$$\lambda_{n,(\sigma,R)} = \min\left(\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{b}{2br+b+1}}, 1\right)\,. \tag{2.3.1}$$

Then for any $s \in [0, \frac{1}{2}]$*, the sequence*

$$a_{n,(\sigma,R)} = R\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{b(r+s)}{2br+b+1}} \tag{2.3.2}$$

is an upper rate of convergence in L^p *for all* $p \geq 1$*, for the interpolation norm of parameter* s*, for the sequence of estimated solutions* $(f_{\mathbf{z}}^{\lambda_{n,(\sigma,R)}})$ *over the family of models* $(\mathcal{M}_{\sigma,M,R})_{(\sigma,M,R)\in\mathbb{R}_+^3}$ *.*

Theorem 2.3.5 (Minimax lower rate)**.** *Let $r > 0, R > 0, b > 1$ and $\alpha > 0$ be fixed. Let ν be a distribution on $\mathcal{X}$ belonging to $\mathcal{P}^{>}(b,\alpha)$. Then the sequence $(a_{n,(\sigma,R)})$ defined in* (2.3.2) *is a weak minimax lower rate of convergence in L^p for all $p \geq 1$, for the model family $\mathcal{M}_{R,M,\sigma} := \mathcal{M}(r,R,\{\nu\})$, $(R,M,\sigma) \in \mathbb{R}_+^3$. If ν belongs to $\mathcal{P}^{>}_{strong}(b,\alpha)$, then the sequence $a_{n,(\sigma,R)}$ is a strong minimax lower rate of convergence in L^p for all $p > 0$, for the model family $\mathcal{M}_{R,M,\sigma}$.*

Finally, we have as a direct consequence:

Corollary 2.3.6 (Minimax optimal rate)**.** *Let $r > 0, b > 1$, $\beta \geq \alpha > 0$ be fixed and assume $\mathcal{P}' = \mathcal{P}^{<}(b,\beta) \cap \mathcal{P}^{>}(b,\alpha) \neq \emptyset$. Then the sequence of estimators $f_{\mathbf{z}}^{\lambda_{n,(\sigma,R)}}$ as defined in* (2.2.15) *is strong minimax optimal in L^p for all $p \geq 1$, under the assumptions and parameter sequence* (2.3.1) *of Theorem 2.3.4 , over the class $\mathcal{M}_{R,M,\sigma} := \mathcal{M}(r,R,\mathcal{P}')$, $(R,M,\sigma) \in \mathbb{R}_+^3$.*

2.4 Discussion

We conclude by briefly discussing some specific points related to our results.

Non-asymptotic, high-probability bounds. The results presented in Section 2.3 are asymptotic in nature and concern moments of the reconstruction error. However, the main underlying technical result is an exponential deviation inequality which holds non-asymptotically. For simplicity of the exposition we have chosen to relegate this result to the Appendix (Proposition 3.4.6 there). Clearly, thanks to such a deviation inequality, we are able to handle moments of all orders of the error. Furthermore, while the asymptotics considered in the previous section always assume that all parameters are fixed as $n \to \infty$, going back to the deviation inequality one could in principle analyze asymptotics of other nonstandard regimes where some parameters are allowed to depend on n .

Adaptivity. For our results we have assumed that the crucial parameters b, r, R concerning the eigenvalue decay of the operator B_ν as well as the regularity of the target function are known, and so is the noise variance σ ; these parameters are used in the choice of regularizing constant λ_n . This is, of course, very unrealistic. Ideally, we would like to have a procedure doing almost as good without knowledge of these parameters in advance. This is the question of adaptivity.

For what concerns the convergence in excess prediction error $L^2(\nu)$, and under the boundedness assumption $|Y| \leq M$, it is well known that a simple hold-out strategy (i.e., choosing, among a finite family of candidate estimators $\widehat{f}_k$, the one achieving minimal error on an held-out validation sample), performed after trimming all candidate estimators $\widehat{f}_k$ to the interval $[-M, M]$, generally speaking selects an estimator close to the best between those considered. One could, for example, consider estimators corresponding to an appropriately chosen discrete (typically geometric) grid of values for λ and adapt the corresponding arguments from [21, 81] ; see also [14] for a general point of view on this question.

However, it remains an open question whether a similar strategy also applies to the error measured in stronger norms: The prediction norm is the only one directly empirically accessible and it does not follow that a value of λ which is good in the sense of prediction (one of which hold-out would select) would automatically also yield good performance for the stronger norms.

In Chapter 5 we study such an adaptive procedure based on Lepski's principle for the oracle selection of

a suitable regularizing constant λ. We even show that an adaptive parameter choice based on Lepski's approach in $L^2(\nu)-$ norm also applies to all other stronger interpolation norms and provides minimax optimal learning rates (up to log-term).

Weak and strong lower bounds. The notion of strong and weak lower bounds introduced in this work (corresponding respectively to a $\liminf$ and $\limsup$ in n) appear to be new. They were motivated by the goal to consider somewhat minimal assumptions on the eigenvalue behavior, i.e., only a one-sided power decay bound, to obtain lower minimax bounds under source condition regularity. It turns out a one-sided power decay bound is the main driver for minimax rates, but excluding arbitrarily abrupt relative variations μ_{2j}/μ_j appears to play a role in distinguishing the weak and strong versions. Such a condition is also called one-sided regular variation, see [7] for extensive considerations on such issues. We believe that this type of assumption can be relevant for the analysis of certain inverse problems when the eigenvalues do not exhibit a two-sided power decay.

Smoothness and source conditions. In considering source conditions (2.2.8) in terms of the operator B_ν as measure of regularity of the target f, we have followed the general approach adopted in previous works on statistical learning using kernels, itself inspired by the setting considered in the (deterministic) inverse problem literature. It is well established in the latter literature that representing the target function in terms of powers of the operator to be inverted is a very natural way to measure its regularity; it can be seen as a way to relate noise and signal in a geometry that is appropriate for the considered ill-posed problem. In our setting, one can, however, wonder why a measure of regularity of the target function should depend on the sampling distribution ν. A high-level answer is that the sampling can itself be seen as a source of noise (or uncertainty), and that it is natural that it enters in the ill-posedness of the problem. For instance, regions in space with sparser sampling will result in more uncertainty. On the other hand, if, say, the support of ν is contained in a low-dimensional manifold, the problem becomes intrinsically lower-dimensional, being understood that we must abandon any hope of estimating outside of the support, and this should also be reflected in the measure of regularity. A more detailed analysis of such issues, and relations to more common notions of regularity, is out of the scope of the present work but certainly an interesting future perspective.

Relaxing Assumption 2.2.1. Assumption 2.2.1 is crucial for our results: It allows us to use the RKHS structure, which is then entirely determined by the operator A. It would be certainly of interest to consider the more general setting where this assumption does not hold, for instance if Af is only assumed to lie in $L^2(\nu)$. In order to follow a similar approach, one would have to introduce separately an RKHS structure having adequate approximation properties. While this setting has been considered in the direct problem case $A = I$ for nonparametric regression (see e.g. [13, 20, 21, 82]), for the general inverse problem this appears to be an open problem, which would in particular require the careful analysis of the interplay between the RKHS structure, the operator A, and the suitable definition of source conditions.

2.5 Proof of Upper Rate

All along the proof, we will use the notation C_a to denote a positive factor only depending on the quantity a. The exact expression of this factor depends on the context and can potentially change from line to line.

Bias-Variance Decomposition

We now consider the following *bias-variance decomposition*

$$\begin{aligned}\bar{B}^s(f_\rho - f_{\mathbf{z}}^\lambda) &= \bar{B}^s(f_\rho - g_\lambda(\bar{B}_{\mathbf{x}})\bar{S}_{\mathbf{x}}^\star \mathbf{y}) \\ &= \bar{B}^s(f_\rho - g_\lambda(\bar{B}_{\mathbf{x}})\bar{B}_{\mathbf{x}} f_\rho) + \bar{B}^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y}) \\ &= \bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}}) f_\rho + \bar{B}^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y})\,, \end{aligned} \tag{2.5.1}$$

where r_λ is given in (2.2.12).

Definition 2.5.1. *We refer to the norm*

$$\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}}) f_\rho\right\|_{\mathcal{H}_1}$$

as the Approximation Error while the norm

$$\left\|\bar{B}^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y})\right\|_{\mathcal{H}_1}$$

is called Sample Error.

We continue with a preliminary inequality. Given $\eta \in (0,1]$, $n \in \mathbb{N}$ and $\lambda \in (0,1]$ assume

$$n \geq 64\lambda^{-1} \max(\mathcal{N}(\lambda), 1) \log^2(8/\eta)\,. \tag{2.5.2}$$

We may apply Proposition A.1.4 to obtain the inequality

$$\left\|(\bar{B}_{\mathbf{x}} + \lambda)^{-1}(\bar{B} + \lambda)\right\| \leq 2\,,$$

with probability at least $1 - \eta$. Combining this with (A.4.3), we get for any $u \in [0,1]$:

$$\begin{aligned}\left\|\bar{B}^u(\bar{B}_{\mathbf{x}} + \lambda)^{-u}\right\| &= \left\|\bar{B}^u(\bar{B} + \lambda)^{-u}(\bar{B} + \lambda)^u(\bar{B}_{\mathbf{x}} + \lambda)^{-u}\right\| \\ &\leq \left\|(\bar{B} + \lambda)(\bar{B}_{\mathbf{x}} + \lambda)^{-1}\right\|^u \leq 2\,. \end{aligned} \tag{2.5.3}$$

From this we deduce readily that, with probability at least $1 - \eta$, we have for any $f \in \mathcal{H}_1$

$$\left\|\bar{B}^s f\right\|_{\mathcal{H}_1} \leq 2\left\|(\bar{B}_{\mathbf{x}} + \lambda)^s f\right\|_{\mathcal{H}_1}. \tag{2.5.4}$$

We upper bound $\left\|\bar{B}^s(f_\rho - f_{\mathbf{z}}^\lambda)\right\|_{\mathcal{H}_1}$ by treating separately the two terms corresponding to the above decomposition, i.e.

$$\left\|\bar{B}^s(f_\rho - f_{\mathbf{z}}^\lambda)\right\|_{\mathcal{H}_1} \leq 2\left(\left\|(\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}}) f_\rho\right\|_{\mathcal{H}_1} + \left\|(\bar{B}_{\mathbf{x}} + \lambda)^s (g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y}))\right\|_{\mathcal{H}_1}\right)\,.$$

Proposition 2.5.2 (Approximation Error)**.** *Let $s \in [0, \frac{1}{2}]$, $r > 0$, $R > 0$, $M > 0$. Suppose $f_\rho \in \Omega_\nu(r, R)$. Let $f_{\mathbf{z}}^\lambda$ be defined as in (2.2.15) using a regularization function of qualification $q \geq r + s$ and put $\bar{\gamma} := \max(\gamma_0, \gamma_q)$. Moreover, let $\eta \in (0,1]$, $\lambda \in (0,1]$ and $n \in \mathbb{N}$.*

1. *(Rough Bound) With probability equal to one*

$$\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}}) f_\rho\right\|_{\mathcal{H}_1} \leq \gamma_0 R\,. \tag{2.5.5}$$

2. *(Refined Bound) If in addition assumption* (2.5.2) *is satisfied, then we have with probability at least* $1-\eta$*:*

$$\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}}) f_\rho\right\|_{\mathcal{H}_1} \leq 16\bar{\gamma} C_r R \log\left(4\eta^{-1}\right) \lambda^s \left(\lambda^r + \frac{1}{\sqrt{n}} \mathbb{1}_{(1,\infty)}(r)\right), \tag{2.5.6}$$

with $C_r = \max(rC, 1)$*, where* C *is explicitly given in Proposition A.4.1, equation* (A.4.2)*.*

Proof of Proposition 2.5.2. 1. A rough bound immediately follows using (2.2.13) and from $f_\rho \in \Omega_\nu(r, R)$:

$$\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}}) f_\rho\right\|_{\mathcal{H}_1} \leq \gamma_0 R\,. \tag{2.5.7}$$

2. Since $f_\rho \in \Omega_\nu(r, R)$, we have

$$\left\|(\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}}) f_\rho\right\|_{\mathcal{H}_1} \leq R \,\left\|(\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}}) \bar{B}^r\right\|. \tag{2.5.8}$$

We now concentrate on the operator norm appearing in the RHS of the above bound, and distinguish between two cases. The first case is $r \geq 1$, for which we write

$$(\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}}) \bar{B}^r = (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}}) \bar{B}^r_{\mathbf{x}} + (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}^r - \bar{B}^r_{\mathbf{x}}). \tag{2.5.9}$$

The operator norm of the first term is estimated via

$$\begin{aligned}
\left\|(\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}}) \bar{B}^r_{\mathbf{x}}\right\| &\leq \sup_{t\in[0,1]} (t+\lambda)^s t^r r_\lambda(t) \\
&\leq \sup_{t\in[0,1]} t^{s+r} r_\lambda(t) + \lambda^s \sup_{t\in[0,1]} t^r r_\lambda(t) \\
&\leq 2\bar{\gamma}\lambda^{s+r},
\end{aligned} \tag{2.5.10}$$

by applying (twice) Lemma 2.2.16 and the assumption that the qualification q of the regularization is greater than $r+s$; we also introduced $\bar{\gamma} := \max(\gamma_0, \gamma_q)$. The second term in equation (2.5.9) is estimated via

$$\begin{aligned}
\left\|(\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}^r - \bar{B}^r_{\mathbf{x}})\right\| &\leq \left\|(\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})\right\| \,\left\|\bar{B}^r - \bar{B}^r_{\mathbf{x}}\right\| \\
&\leq 2\bar{\gamma} r C\, \lambda^s \,\left\|\bar{B} - \bar{B}_{\mathbf{x}}\right\|,
\end{aligned}$$

where C is given in Proposition A.4.1, equation (A.4.2). For the first factor we have used the same device as previously for the term in (2.5.10) based on Lemma 2.2.16, and for the second factor we used Proposition A.4.1. Finally using Proposition *A*.1.5 to upper bound $\left\|\bar{B} - \bar{B}_{\mathbf{x}}\right\|$, collecting the previous estimates we obtain with probability at least $1 - \eta/2$:

$$\left\|(\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}}) f_\rho\right\|_{\mathcal{H}_1} \leq \bar{\gamma} r C R \log\left(4\eta^{-1}\right) \left(\lambda^r + \frac{1}{\sqrt{n}}\right) \lambda^s. \tag{2.5.11}$$

We turn to the case $r < 1$, for which we want to establish a similar inequality. Instead of (2.5.9)

we use:

$$\begin{aligned}\left\|(\bar{B}_{\mathbf{x}}+\lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}^r\right\| &= \left\|(\bar{B}_{\mathbf{x}}+\lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}}+\lambda)^r(\bar{B}_{\mathbf{x}}+\lambda)^{-r}\bar{B}^r\right\| \\ &\leq 2\left\|(\bar{B}_{\mathbf{x}}+\lambda)^{r+s} r_\lambda(\bar{B}_{\mathbf{x}})\right\| \\ &\leq 8\bar{\gamma}\lambda^{r+s}, \end{aligned} \tag{2.5.12}$$

where we have used the (transposed version of) inequality (2.5.3) (valid with probability at least $1-\eta/2$); and, for the last inequality, an argument similar the one leading to (2.5.10) (using this time that $(t+\lambda)^{r+s} \leq 2(t^{r+s}+\lambda^{r+s})$ for all $t \geq 0$ since $r+s \leq 2$ in the case we are considering). Combining this with (2.5.11) , (2.5.8) and (2.5.4) implies that inequality (2.5.6) holds with probability at least $1-\eta$.

□

Proposition 2.5.3 (Sample Error)**.** *Let $s \in [0, \frac{1}{2}]$, $r > 0$, $R > 0$, $M > 0$. Suppose $f_\rho \in \Omega_\nu(r, R)$. Let $f_{\mathbf{z}}^\lambda$ be defined as in* (2.2.15) *using a regularization function of qualification $q \geq r+s$ and put $\bar{\gamma} := \max(\gamma_0, \gamma_q)$. Moreover, let $\eta \in (0, 1]$, $\lambda \in (0, 1]$ and $n \in \mathbb{N}$.*

1. *(Rough Bound) With probability at least $1-\eta$:*

$$\left\|\bar{B}^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f - \bar{S}_{\mathbf{x}}^\star \mathbf{y})\right\|_{\mathcal{H}_1} \leq C_{E,M,\sigma} \log(2\eta^{-1}) \frac{1}{\lambda\sqrt{n}} . \tag{2.5.13}$$

2. *(Refined Bound) If in addition assumption* (2.5.2) *holds, then we have with probability at least $1-\eta$:*

$$\left\|\bar{B}^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y})\right\|_{\mathcal{H}_1} \leq C_{s,D,E} \log(8\eta^{-1}) \lambda^s \left(\frac{M}{n\lambda} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n\lambda}}\right) . \tag{2.5.14}$$

Proof of Proposition 2.5.3. 1. Using (2.2.11) and the second part of Proposition A.1.2 , we obtain that with probability at least $1-\eta$

$$\begin{aligned}\left\|\bar{B}^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y})\right\|_{\mathcal{H}_1} &\leq 2\log(2\eta^{-1}) \frac{E}{\lambda}\left(\frac{M}{n} + \sqrt{\frac{\sigma^2}{n}}\right) \\ &\leq C_{E,M,\sigma} \frac{1}{\lambda\sqrt{n}} \log(2\eta^{-1}) . \end{aligned}$$

2. We further split by writing

$$(\bar{B}_{\mathbf{x}}+\lambda)^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y}) \quad = \quad H_{\mathbf{x}}^{(1)} \cdot H_{\mathbf{x}}^{(2)} \cdot h_{\mathbf{z}}^\lambda \tag{2.5.15}$$

with

$$\begin{aligned} H_{\mathbf{x}}^{(1)} &:= (\bar{B}_{\mathbf{x}}+\lambda)^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}}+\lambda)^{\frac{1}{2}}, \\ H_{\mathbf{x}}^{(2)} &:= (\bar{B}_{\mathbf{x}}+\lambda)^{-\frac{1}{2}}(\bar{B}+\lambda)^{\frac{1}{2}}, \\ h_{\mathbf{z}}^\lambda &:= (\bar{B}+\lambda)^{-\frac{1}{2}}(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y}) \end{aligned}$$

and proceed by bounding each factor separately.

For the first term, we have (for any $\lambda \in (0,1]$ and $\mathbf{x} \in \mathcal{X}^n$), and remembering that $s \leq 1/2$:

$$\begin{aligned}
\left\| H_{\mathbf{x}}^{(1)} \right\| &\leq \sup_{t\in[0,1]} (t+\lambda)^{s+\frac{1}{2}} g_\lambda(t) \\
&\leq \lambda^{s+\frac{1}{2}} \sup_{t\in[0,1]} g_\lambda(t) + \sup_{t\in[0,1]} \left| t^{s+\frac{1}{2}} g_\lambda(t) \right| \\
&\leq E\lambda^{s-\frac{1}{2}} + \left(\sup_{t\in[0,1]} |t g_\lambda(t)| \right)^{s+\frac{1}{2}} \left(\sup_{t\in[0,1]} |g_\lambda(t)| \right)^{\frac{1}{2}-s} \\
&\leq E\lambda^{s-\frac{1}{2}} + D^{s+\frac{1}{2}} E^{\frac{1}{2}-s} \lambda^{s-\frac{1}{2}} = C_{s,D,E} \lambda^{s-\frac{1}{2}}, \qquad (2.5.16)
\end{aligned}$$

where we have used Definition 2.2.14 (i), (ii).

The probabilistic bound on $H_{\mathbf{x}}^{(2)}$ follows from Proposition A.1.4, which we can apply using assumption (2.5.2), combined with Proposition A.4.2. This ensures with probability at least $1 - \eta/4$

$$\left\| H_{\mathbf{x}}^{(2)} \right\| \leq \sqrt{2}\,. \qquad (2.5.17)$$

Finally, the probabilistic bound on $h_{\mathbf{z}}^{\lambda}$ follows from Proposition A.1.2: With probability at least $1 - \eta/4$, we have

$$\left\| h_{\mathbf{z}}^{\lambda} \right\|_{\mathcal{H}_1} \leq 2\log(8\eta^{-1}) \left(\frac{M}{n\sqrt{\lambda}} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n}} \right). \qquad (2.5.18)$$

As a result, combining (2.5.16), (2.5.17) and (2.5.18) with (2.5.15) gives with probability at least $1 - \eta$

$$\left\| (\bar{B}_{\mathbf{x}} + \lambda)^s g_\lambda(\bar{B}_{\mathbf{x}}) (\bar{B}_{\mathbf{x}} f - \bar{S}_{\mathbf{x}}^{\star} \mathbf{y}) \right\|_{\mathcal{H}_1} \leq C_{s,D,E} \log(8\eta^{-1}) \lambda^s \left(\frac{M}{n\lambda} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n\lambda}} \right). \qquad (2.5.19)$$

Combining the last bound with (2.5.4) completes the proof.

□

Corollary 2.5.4. *Let $s \in [0, \frac{1}{2}]$, $\sigma > 0, M > 0, r > 0, R > 0, \beta > 0, b > 1$ and assume the generating distribution of (X, Y) belongs to $\mathcal{M}(r, R, \mathcal{P}^<(b, \beta))$ (defined in Section 2.2.3). Let $f_{\mathbf{z}}^{\lambda}$ be the estimator defined as in* (2.2.15) *using a regularization function of qualification $q \geq r + s$ and put $\bar{\gamma} := \max(\gamma_0, \gamma_q)$. Then, there exists $n_0 > 0$ (depending on the above parameters), so that for all $n \geq n_0$, if we set*

$$\lambda_n = \min\left(\left(\frac{\sigma}{R\sqrt{n}} \right)^{\frac{2b}{2br+b+1}}, 1 \right), \qquad (2.5.20)$$

then with probability at least $1 - \eta$:

$$\left\| \bar{B}^s (f_\rho - f_{\mathbf{z}}^{\lambda_n}) \right\|_{\mathcal{H}_1} \leq C_{s,\beta,\bar{\gamma},D,E} C_b \, \log(8\eta^{-1}) R \left(\frac{\sigma}{R\sqrt{n}} \right)^{\frac{2b(r+s)}{2br+b+1}},$$

provided $\log(\eta^{-1}) \leq C_{b,\beta,\sigma,R}\, n^{\frac{br}{2br+b+1}}$ *and with* $C_b = \sqrt{\frac{b}{b-1}}$.

Remark: In the above corollary, n_0 can possibly depend on all parameters, but the constant in front of the upper bound does not depend on R, σ, nor M. In this sense, this result tracks precisely the effect

of these important parameters on the scaling of the rate, but remains asymptotic in nature: It cannot be applied if, say, R, σ of M also depend on n (because the requirement $n \geq n_0$ might then lead to an impossiblity.) If some parameters are allowed to change with n , one should go back to the non-asymptotic statement of Proposition 2.5.2 and Proposition 2.5.3 for an analysis of the rates.

Proof of Corollary 2.5.4. We check that the assumptions of both Proposition 2.5.2 and Proposition 2.5.3 are satisfied provided n is big enough. Concerning assumption (2.5.2), let us recall that by Lemma 2.2.10:

$$\mathcal{N}(\lambda) \leq \frac{\beta b}{b-1} \lambda^{-1/b} \,. \tag{2.5.21}$$

Consequently, (2.5.2) is ensured by the sufficient condition

$$n \geq C_{b,\beta} \log^2(8\eta^{-1}) \lambda^{-\frac{1}{b}-1} \;\Leftarrow\; \log(\eta^{-1}) \leq C_{b,\beta,\sigma,R}\, n^{\frac{br}{2br+b+1}} \,. \tag{2.5.22}$$

Applying Proposition 2.5.2, Proposition 2.5.3, Lemma 2.2.10 again and folding the effect of the parameters we do not intend to track precisely into a generic multiplicative constant, we obtain using decomposition (2.5.1) that with probability $1-\eta$

$$\left\| \bar{B}^s (f_\rho - f_{\mathbf{z}}^\lambda) \right\|_{\mathcal{H}_1} \leq C_\bullet \, \log(8\eta^{-1}) \lambda^s \left(R \left(\lambda^r + \frac{1}{\sqrt{n}} \mathbb{1}_{(1,\infty)}(r) \right) + \left(\frac{M}{n\lambda} + \frac{\sigma C_b}{\sqrt{n}} \lambda^{-\frac{b+1}{2b}} \right) \right) . \tag{2.5.23}$$

with $\bullet = (s, \bar{\gamma}, D, E, \beta)$ and $C_b = \sqrt{\frac{b}{b-1}}$.

Observe that the choice (2.5.20) implies that $rn^{-\frac{1}{2}} = o(\lambda_n^r)$. Therefore, up to requiring n large enough and multiplying the front factor by 2 , we can disregard the term $r/\sqrt{n}$ in the second factor of the above bound. Similarly, by comparing the exponents in n, one can readily check that

$$\frac{M}{n\lambda_n} = o\left(C_b \sqrt{\frac{1}{n} \lambda_n^{-\frac{b+1}{b}}} \right) ,$$

so that we can also disregard the term $M(n\lambda_n)^{-1}$ for n large enough (again, up multiplying the front factor by 2) and concentrate on the two remaining main terms of the upper bound in (2.5.23), which are $R\lambda^r$ and $\sigma \lambda^{-\frac{b+1}{2b}} n^{-\frac{1}{2}}$. The proposed choice of λ_n balances precisely these two terms and easy computations lead to the announced conclusion. □

We now come to the proof of our main Theorem for the upper bound. To simplify notation and argument we will adopt the following conventions:

- The dependence of multiplicative constants C on various parameters will (generally) be omitted, except for σ, M, R, η and n which we want to track precisely.

- The expression "for n big enough" means that the statement holds for $n \geq n_0$, with n_0 potentially depending on all model parameters (including σ, M and R), but not on η.

Proof of Theorem 2.3.4

Proof of Theorem 2.3.4. We would like to "integrate" the bound of Corollary 2.5.4 over η to obtain a bound in L^p norm (see Lemma A.3.1), unfortunately, the condition on η prevents this since very large deviations are excluded. To alleviate this, we use a much coarser "fallback" upper bound which is valid for all $\eta \in (0,1]$. From (2.5.5) and (2.5.13) , we conclude that

$$\mathbb{P}\Big[\big\|\bar{B}^s(f_\rho - f_{\mathbf{z}}^{\lambda})\big\|_{\mathcal{H}_1} \geq a' + b' \log \eta^{-1} \Big] \leq \eta \,,$$

for all $\eta \in (0,1]$, with $a' := C_{\sigma,M,R} \max\left(\frac{1}{\lambda\sqrt{n}}, 1\right)$ and $b' := \frac{C_{\sigma,M}}{\lambda\sqrt{n}}$. On the other hand, Corollary 2.5.4, ensured that

$$\mathbb{P}\Big[\big\|\bar{B}^s(f_\rho - f_{\mathbf{z}}^{\lambda_{n,(\sigma,R)}})\big\|_{\mathcal{H}_1} \geq a + b \log \eta^{-1} \Big] \leq \eta \,, \text{ for } \log \eta^{-1} \leq \log \eta_0^{-1} := C_{\sigma,R} n^{\frac{br}{2br+b+1}} \,,$$

with $a = b := CR\left(\frac{\sigma}{R\sqrt{n}}\right)^{\frac{2b(r+s)}{2br+b+1}} = C a_{n,(\sigma,R)}$, provided that n is big enough.

We can now apply Corollary A.3.2, which encapsulates some tedious computations, to conclude that for any $p \leq \frac{1}{2}\log \eta_0^{-1}$ and n big enough:

$$\mathbb{E}\Big[\big\|\bar{B}^s(f_\rho - f_{\mathbf{z}}^{\lambda_{n,(\sigma,R)}})\big\|_{\mathcal{H}_1}^p \Big] \leq C_p \left(a_{n,(\sigma,R)}^p + \eta_0 \left((a')^p + 2(b' \log \eta_0^{-1})^p \right) \right) .$$

Now for fixed σ, M, R , and p , the quantities a', b' are powers of n , while $\eta_0 = \exp(-C_{\sigma,r} n^{\nu_{b,r}})$ for $\nu_{b,r} > 0$. The condition $p \leq \frac{1}{2}\log \eta_0^{-1}$ is thus satisfied for n large enough and we have

$$\limsup_{n\to\infty} \sup_{\rho \in \mathcal{M}_{\sigma,M,R}} \frac{\mathbb{E}_{\rho^{\otimes n}}\Big[\big\|\bar{B}^s(f_\rho - f_{\mathbf{z}}^{\lambda_{n,(\sigma,R)}})\big\|_{\mathcal{H}_1}^p \Big]^{\frac{1}{p}}}{a_{n,(\sigma,R)}} \leq C \,,$$

(where we reiterate that the constant C above may depend on all parameters including p , but not on σ, M nor R). Therefore, taking the supremum over (σ, M, R) yields the desired conclusion. □

2.6 Proof of Lower Rate

We will apply the general reduction scheme in Appendix A.5 , Proposition A.5.1 to our target distance $d_s : \Omega_\nu(r,R) \times \Omega_\nu(r,R) \longrightarrow \mathbb{R}_+$, given by

$$d_s(f_1, f_2) \;=\; \big\|\bar{B}^s\,(f_1 - f_2)\big\|_{\mathcal{H}_1} \;,$$

with $s \in [0, \frac{1}{2}]$ and $\nu \in \mathcal{P}^{>}(b,\alpha)$. We will establish the lower bounds in the particular case where the distribution of Y given X is Gaussian with variance σ^2 (which satisfies the Bernstein moment condition (2.2.5) with $M = \sigma$) . The main effort is to construct a finite subfamily belonging to the model of interest and suitably satisfying the assumptions of Proposition A.5.1; this is the goal of the forthcoming propositions and lemmata.

Proposition 2.6.1. *Let $\nu \in \mathcal{P}^{>}(b,\alpha)$, for $b > 1$, $\alpha > 0$. Assume that $r > 0, R > 0$. To each $f \in \Omega_\nu(r,R)$*

and $x \in \mathcal{X}$ we associate the following measure:

$$\rho_f(dx, dy) := \rho_f(dy|x)\nu(dx)\,, \text{ where } \rho_f(dy|x) := \mathcal{N}(\bar{S}_x f, \sigma^2)\,. \tag{2.6.1}$$

Then:

(i) The measure ρ_f belongs to the class $\mathcal{M}(r, R, \mathcal{P}^{>}(b, \alpha))$, defined in (2.2.9).

(ii) Given $f_1, f_2 \in \Omega_\nu(r, R)$, the Kullback-Leibler divergence between ρ_1 and ρ_2 satisfies

$$\mathcal{K}(\rho_1, \rho_2) = \frac{1}{2\sigma^2} \big\| \sqrt{\bar{B}}(f_1 - f_2) \big\|^2_{\mathcal{H}_1}\,.$$

Proof. Point (i) follows directly from the definition of the class $\mathcal{M}(r, R, \mathcal{P}^{>}(b, \alpha))$. For point (ii), note that the Kullback-Leibler divergence between two Gaussian distributions with identical variance σ^2 and mean difference Δ is $\Delta/2\sigma^2$. Since ρ_1, ρ_2 have the same X-marginal ν, it holds

$$\begin{aligned}\mathcal{K}(\rho_1, \rho_2) = \mathbb{E}[\mathcal{K}(\rho_1(.|X), \rho_2(.|X))] &= \frac{1}{2\sigma^2} \int \big(\bar{S}_x(f_1 - f_2)\big)^2 \, d\nu(x) \\ &= \frac{1}{2\sigma^2} \big\| \bar{S}(f_1 - f_2) \big\|^2_{L^2(\nu)} = \frac{1}{2\sigma^2} \big\| \sqrt{\bar{B}}(f_1 - f_2) \big\|^2_{\mathcal{H}_1}\,.\end{aligned}$$

□

The following lemma is a variant from [24], Proposition 6, which will be useful in the subsequent proposition.

Lemma 2.6.2. *For any $m \geq 28$ there exist an integer $N_m > 3$ and $\pi_1, ..., \pi_{N_m} \in \{-1, +1\}^m$ such that for any $i, j \in \{1, ..., N_m\}$ with $i \neq j$ it holds*

$$\log(N_m - 1) > \frac{m}{36} > \frac{2}{3}\,, \tag{2.6.2}$$

and

$$\sum_{l=1}^{m} (\pi_i^l - \pi_j^l)^2 \geq m\,, \tag{2.6.3}$$

where $\pi_i = (\pi_i^1, ..., \pi_i^m)$.

Proposition 2.6.3. *Assume $\nu \in \mathcal{P}^{>}(b, \alpha)$. Let $0 \leq s \leq 1/2$, $R > 0$ and $r > 0$. For any $\varepsilon_0 > 0$ there exist $\varepsilon \leq \varepsilon_0$, $N_\varepsilon \in \mathbb{N}$ and functions $f_1, ..., f_{N_\varepsilon} \in \mathcal{H}_1$ satisfying*

(i) $f_i \in \Omega_\nu(r, R)$ for any $i = 1, ..., N_\varepsilon$ and

$$\big\| \bar{B}^s(f_i - f_j) \big\|^2_{\mathcal{H}_1} > \varepsilon^2\,,$$

for any $i, j = 1, ..., N_\varepsilon$ with $i \neq j$.

(ii) Let $\rho_i := \rho_{f_i}$ be given by (2.6.1)*). Then it holds*

$$\mathcal{K}(\rho_i, \rho_j) \leq C_{b,r,s}\, R^2 \sigma^{-2} \left(\frac{\varepsilon}{R}\right)^{\frac{2r+1}{r+s}}\,,$$

for any $i, j = 1, ..., N_\varepsilon$ *with* $i \neq j$.

(iii) $\log(N_\varepsilon - 1) \geq C_{\alpha,b,r,s} \left(\frac{R}{\varepsilon}\right)^{\frac{1}{b(r+s)}}$.

If ν *belongs to the subclass* $\mathcal{P}^{>}_{strong}(b, \alpha)$*, then the assertions from (i), (ii) and (iii) are valid for all* $\varepsilon > 0$ *small enough (depending on the parameters* r, R, s, α, b *as well as* j_0, γ *coming from the choice of* ν *in* $\mathcal{P}^{>}_{strong}(b, \alpha)$*; the multiplicative constants in (ii), (iii) then also depend on* γ *.)*

Proof. We first prove the proposition under the stronger assumption that ν belongs to $\mathcal{P}^{>}_{strong}(b, \alpha)$. We recall from (2.2.2) that we denote $(e_l)_{l\geq 1}$ an orthonormal family of $\mathcal{H}_1$ of eigenvectors of $\bar{B}$ corresponding to the eigenvalues $(\mu_l)_{l\geq 1}$, which satisfy by definition of $\mathcal{P}^{>}_{strong}(b, \alpha)$:

$$\forall l \geq 0 \ : \mu_l \geq \alpha l^{-b} \tag{2.6.4}$$

and

$$\forall l \geq l_0 \ : \mu_{2l} \geq 2^{-\gamma} \mu_l \ , \tag{2.6.5}$$

for some $l_0 \in \mathbb{N}$ and for some $\gamma > 0$. For any given $\varepsilon < R 2^{-\gamma(r+s)} \left(\frac{\alpha^{1/b}}{\max(28, l_0)}\right)^{b(r+s)}$ we pick $m = m(\varepsilon) := \max\{\ l \geq 1 : \ \mu_l \geq 2^\gamma (\varepsilon R^{-1})^{\frac{1}{r+s}}\ \}$. Note that $m \geq \max(28, l_0)$, following from the choice of ε and from (2.6.4).

Let $N_m > 3$ and $\pi_1, ..., \pi_{N_m} \in \{-1, +1\}^m$ be given by Lemma 2.6.2 and define

$$g_i := \frac{\varepsilon}{\sqrt{m}} \sum_{l=m+1}^{2m} \pi_i^{(l-m)} \left(\frac{1}{\mu_l}\right)^{r+s} e_l \ , \quad i = 1, ..., m \ . \tag{2.6.6}$$

We have by (2.6.5) and from the definition of m

$$\|g_i\|^2_{\mathcal{H}_1} = \frac{\varepsilon^2}{m} \sum_{l=m+1}^{2m} \left(\frac{1}{\mu_k}\right)^{2(r+s)} \leq \varepsilon^2 \mu_{2m}^{-2(r+s)} \leq \varepsilon^2 2^{2\gamma(r+s)} \mu_m^{-2(r+s)} \leq R^2 \, .$$

For $i = 1, ..., N_m$ let $f_i := \bar{B}^r g_i \in \Omega_\nu(r, R)$, with g_i as in (2.6.6). Then

$$\begin{aligned} \| \bar{B}^s (f_i - f_j) \|^2_{\mathcal{H}_1} &= \| \bar{B}^{r+s} (g_i - g_j) \|^2_{\mathcal{H}_1} \\ &= \frac{\varepsilon^2}{m} \sum_{l=m+1}^{2m} (\pi_i^{l-m} - \pi_j^{l-m})^2 \left(\frac{1}{\mu_l}\right)^{2(r+s)} \mu_l^{2(r+s)} \geq \varepsilon^2 \, , \end{aligned}$$

by (2.6.3) , and the proof of (i) is finished. For $i = 1, ..., N_\varepsilon$, let $\rho_i = \rho_{f_i}$ be defined by (2.6.1). Then,

using the definition of m, the Kullback-Leibler divergence satisfies

$$\begin{aligned}\mathcal{K}(\rho_i,\rho_j) &= \frac{1}{2\sigma^2}\,\left\|\sqrt{\bar{B}}(f_i-f_j)\right\|_{\mathcal{H}_1}^2 \\ &= \frac{1}{2\sigma^2}\,\left\|\bar{B}^{r+1/2}(g_i-g_j)\right\|_{\mathcal{H}_1}^2 \\ &= \frac{\varepsilon^2}{2\sigma^2 m}\sum_{l=m+1}^{2m}(\pi_i^{l-m}-\pi_j^{l-m})^2\left(\frac{1}{\mu_l}\right)^{2(r+s)}\mu_l^{2r+1} \\ &\leq 2\sigma^{-2}\mu_{m+1}^{1-2s}\varepsilon^2 \\ &\leq 2^{1+\gamma(1-2s)}\,\sigma^{-2}R^2\left(\frac{\varepsilon}{R}\right)^{\frac{1+2r}{r+s}}\,,\end{aligned}$$

which shows (ii). Finally, (2.6.2), (2.6.4), (2.6.5) and the definition of m imply

$$\log(N_m-1)\geq\frac{m}{36}\geq\frac{\alpha^{1/b}}{36}\mu_m^{-1/b}\geq\frac{\alpha^{1/b}}{36}2^{-\gamma/b}\mu_{2m}^{-1/b}\geq\frac{\alpha^{1/b}}{36}2^{-2\gamma/b}\left(\frac{R}{\varepsilon}\right)^{\frac{1}{b(r+s)}}\,,$$

thus (iii) is established.

We now assume that ν belongs to $\mathcal{P}^{>}(b,\alpha)$ and only satisfies condition (2.6.4). Let any $\varepsilon_0>0$ be given. We pick $m\in\mathbb{N}$ satisfying $m\geq 28$ and the two following conditions:

$$\mu_m\leq 2^{b+1}(R^{-1}\varepsilon_0)^{\frac{1}{r+s}}\,, \tag{2.6.7}$$

$$\frac{\mu_{2m}}{\mu_m}\geq 2^{-b-1}\,. \tag{2.6.8}$$

Since the sequence of eigenvalues (μ_m) converges to 0, condition (2.6.7) must be satisfied for any m big enough, say $m\geq m_0(\varepsilon_0)$. Subject to that condition, we argue by contradiction that there must exist m satisfying (2.6.8). If that were not the case, we would have by immediate recursion for any $l>0$, introducing $m':=2^l m_0(\varepsilon_0)$:

$$\mu_{m'}<2^{-l(b+1)}\mu_{m_0(\varepsilon_0)}=\left(\frac{m'}{m_0(\varepsilon_0)}\right)^{-b-1}\mu_{m_0(\varepsilon_0)}=C_{\varepsilon_0}(m')^{-(b+1)}\,,$$

which would (eventually, for l big enough) contradict (2.6.4). Therefore, there must exist an $m>m_0$ satisfying the required conditions. Now put

$$\varepsilon:=2^{-(b+1)(r+s)}R\mu_m^{r+s}\leq\varepsilon_0\,, \tag{2.6.9}$$

where the inequality is from requirement (2.6.7). For $i=1,...,N_m$, we define g_i as in (2.6.6). Then $||g_i||_{\mathcal{H}_1}\leq R$. Again, let $f_i:=\bar{B}^r g_i\in\Omega_\nu(r,R)$ and the same calculations as above (with γ replaced by $b+1$) lead to $(i),(ii)$ and (iii). □

Proof of Theorem 2.3.5. Let the parameters r,R,s,b,α,σ be fixed for the rest of the proof, and the marginal distribution $\nu\in\mathcal{P}^{>}(b,\alpha)$ also be fixed.

Our aim is to apply Proposition A.5.1 to the distance $d_s(f_1,f_2):=\left\|\bar{B}^s\,(f_1-f_2)\right\|_{\mathcal{H}_1}$ ($s\in[0,\frac{1}{2}]$), on the class $\Theta:=\Omega_\nu(r,R)$, where for any $f\in\Omega_\nu(r,R)$, the associated distribution is $P_f:=\rho_f^{\otimes n}$ with ρ_f defined as from Proposition 2.6.1 (i); more precisely, we will apply this proposition along a well-chosen sequence $(n_k,\varepsilon_k)_{k\geq 0}$. From Proposition 2.6.3, we deduce that there exist a decreasing null sequence $(\varepsilon_k)>0$ such that for any ε belonging to the sequence, there exists N_ε and functions $f_1,...,f_{N_\varepsilon}$ satisfying (i)-(ii)-(iii).

In the rest of this proof, we assume $\varepsilon = \varepsilon_k$ is a value belonging to the null sequence. Point (i) gives requirement (i) of Proposition A.5.1. We turn to requirement (A.5.2). Let $\rho_j = \rho_{f_j}$ be given by (2.6.1). Then by Proposition 2.6.3 (ii)-(iii) :

$$\begin{aligned}\frac{1}{N_\varepsilon - 1}\sum_{j=1}^{N_\varepsilon - 1} \mathcal{K}(\rho_j^{\otimes n}, \rho_{N_\varepsilon}^{\otimes n}) &= \frac{n}{N_\varepsilon - 1}\sum_{j=1}^{N_\varepsilon - 1} \mathcal{K}(\rho_j, \rho_{N_\varepsilon}) \\ &\leq nC_{b,r,s}\, R^2\sigma^{-2}\left(\frac{\varepsilon}{R}\right)^{\frac{2r+1}{r+s}} \\ &\leq nC_{\alpha,b,r,s}\, R^2\sigma^{-2}\left(\frac{\varepsilon}{R}\right)^{\frac{2br+b+1}{b(r+s)}} \log(N_\varepsilon - 1) \\ &=: \omega\, \log(N_\varepsilon - 1)\,.\end{aligned}$$

Choosing $n := \left\lfloor \left(8C_{\alpha,b,r,s}\, R^2\sigma^{-2}\left(\varepsilon R^{-1}\right)^{\frac{2br+2r+1}{b(r+s)}}\right)^{-1} \right\rfloor$ ensures $\omega \leq \frac{1}{8}$ and therefore requirement (A.5.2) is satisfied. Then Proposition A.5.1 entails:

$$\begin{aligned}\inf_{\hat{f}_\bullet}\max_{1\leq j\leq N_\varepsilon} \rho_j^{\otimes n}\left(\left\|\bar{B}^s(\hat{f}_\bullet - f_j)\right\|_{\mathcal{H}_1} \geq \frac{\varepsilon}{2}\right) &\geq \frac{\sqrt{N_\varepsilon - 1}}{1+\sqrt{N_\varepsilon - 1}}\left(1 - 2\omega - \sqrt{\frac{2\omega}{\log(N_\varepsilon - 1)}}\right) \\ &\geq \frac{1}{2}\left(\frac{3}{4} - \sqrt{\frac{3}{8}}\right) \\ &> 0\,.\end{aligned}$$

This inequality holds for any (n_k, ε_k) for ε_k in the decreasing null sequence and n_k given by the above formula; we deduce that $n_k \to \infty$ with

$$\varepsilon_k \geq C_{\alpha,b,r,s} R\left(\frac{\sigma}{R\sqrt{n_k}}\right)^{\frac{2b(r+s)}{2br+b+1}}.$$

Thus, applying $(A.5.1)$ and taking the limsup gives the result.

Now suppose that that ν belongs to $\mathcal{P}^{>}_{strong}(b,\alpha)$. Define

$$\varepsilon := R(8C_{\alpha,b,r,s})^{-\frac{b(r+s)}{2br+b+1}}\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{b(r+s)}{2br+b+1}},$$

then for any n sufficiently large, points (i)-(ii)-(iii) of Proposition 2.6.3 will hold. The same calculations as above now hold for any n large enough. Finally, taking the lim inf finishes the proof. □

Proof of Corollary 2.3.6. The main point is only to ensure that the *strong* minimax lower bound applies, for this we simply check that $\mathcal{P}^{>}_{strong}(b,\alpha) \supset \mathcal{P}' = \mathcal{P}^{<}(b,\beta) \cap \mathcal{P}^{>}(b,\alpha)$. For any $\nu \in \mathcal{P}^{<}(b,\beta) \cap \mathcal{P}^{>}(b,\alpha)$, the eigenvalues of the operator $\bar{B}$ satisfy $\alpha j^{-b} \leq \mu_j \leq \beta j^{-b}$ for all $j \geq 1$. It follows that for any $j \geq 1$:

$$\frac{\mu_{2j}}{\mu_j} \geq \frac{\alpha}{\beta}2^{-b},$$

so that the conditions for $\nu \in \mathcal{P}^{>}_{strong}(b,\alpha)$ are met (with parameters $\gamma := b + \log_2\frac{\beta}{\alpha}$, $l_0 = 1$). Since $\mathcal{P}'$ is assumed to be non empty, for any $\nu \in \mathcal{P}'$ the strong lower minimax bound of Theorem 2.3.5 applies to the family $\mathcal{M}_{R,M,\sigma} := \mathcal{M}(r,R,\{\nu\})$ and a fortiori to the family $\mathcal{M}_{R,M,\sigma} := \mathcal{M}(r,R,\mathcal{P}')$ whose models

are larger. On the other hand since, $\mathcal{M}(r, R, \mathcal{P}') \subset \mathcal{M}(r, R, \mathcal{P}^<(b, \beta))$, the upper bound of Theorem 2.3.4 applies and we are done. $\square$

Chapter 3

Minimax Rates beyond the regular Case

Up to now we have considered classes of marginals such that the induced covariance operator $\bar{B}$ of the kernel feature map - via equation (2.2.1) - has eigenvalues μ_j falling into a window $[\alpha j^{-b}, \beta j^{-b}]$ (see Section 2.2.3), where upper and lower bound as a function of $j \in \mathbb{N}$ only differ by a multiplicative constant. We call this the regular case. The aim of this chapter is to relax this condition on the allowed window of eigenvalues. This will enlarge the class of allowed marginals (or sampling distributions).

We recall that in a distribution free approach it is imperative to avoid specific assumptions on the data generating distribution and the induced marginal, the sampling distribution. We further recall that these were restricted by our assumptions on the decay of the eigenvalues of the covariance operator $\bar{B} = \bar{B}_\nu$. It is therefore very much motivated by the general philosophy of a distribution free approach to the regression problem to also consider classes of much more general eigenvalue decay as $\mu_j \asymp j^{-b}$ and to prove minimax optimality for rates of convergence for the associated sets of allowed marginals. We remind the reader that minimax optimality of rates of convergence is a global property of the set of allowed marginals (or, more exactly, of the whole model class) depending in some sense on the worst possible case of convergence within that set. One therefore expects that broadening that set will lead to less refined notions of minimax optimal rates as compared to our previous more specific class. This is precisely what we will see (compare Remark 3.1.1).

But, for complex data, there is no strong reason to expect that the decay of the eigenvalues of the associated covariance operator should be strictly polynomial, and we would like to cover behavior as general as possible for the eigenvalue decay, for instance:

- decay rates including other slow varying functions, such as $\mu_{v,i} \asymp i^{-b}(\log i)^c(\log\log i)^d$;
- eigenvalue sequence featuring plateaus separated by relative gaps;
- shifting or switching along the sequence between different polynomial-type regimes,

which all might correspond to different types of structure of the data at different scales. With the

distribution-free point of view in mind, we therefore try to characterize minimax rates for target function classes of the form (2.5.8), striving for assumptions as weak as possible on the eigenvalue sequence. In this section we present a first result in this direction (i.e. beyond the regular case) and show that kernel methods also achieve minimax optimal rates in this general case.

3.1 Setting

For the sake of the reader we briefly recall the setting of Chapter 2 . We let $\mathbf{Z} = \mathcal{X} \times \mathbb{R}$ denote the sample space, where the input space $\mathcal{X}$ is a standard Borel space endowed with a fixed unknown probability measure ν. Let $\mathcal{P}$ denote the set of all probability distributions on $\mathcal{X}$. The kernel space $\mathcal{H}_1$ is assumed to be separable, equipped with a measurable positive semi-definite kernel K. Moreover, we consider the covariance operator $\bar{B} = \bar{B}_\nu = \mathbb{E}[\bar{B}_x] = \mathbb{E}[\bar{F}_X \otimes \bar{F}_X^\star]$, which can be shown to be positive, self-adjoint and trace class (and hence in particular compact). Given a sample $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, we define the sampling operator $\bar{S}_\mathbf{x} : \mathcal{H}_1 \longrightarrow \mathbb{R}^n$ by $(S_\mathbf{x} f)_i = \langle f, \bar{F}_{x_i} \rangle_{\mathcal{H}_1}$. The empirical covariance operator is given by $\bar{B}_\mathbf{x} = \bar{S}_\mathbf{x}^\star \bar{S}_\mathbf{x}$. Throughout we denote by μ_j the positive eigenvalues of $\bar{B}$ in decreasing order, satisfying $0 < \mu_{j+1} \leq \mu_j$ for all $j > 0$ and $\mu_j \searrow 0$. For any $t > 0$ we denote by

$$\mathcal{F}(t) := \#\{j \in \mathbb{N} : \mu_j \geq t\} \tag{3.1.1}$$

the cardinality of eigenvalues above the threshold t. Note that $\mathcal{F}$ is left-continuous and decreasing as t grows with $\mathcal{F}(t) = 0$ for any $t > 1$, and $\mathcal{F}(t)$ has limit $+\infty$ as $t \downarrow 0$. Given $r > 0$, we set $\mathcal{G}(t) := \frac{t^{2r+1}}{\mathcal{F}(t)}$ (possibly taking the value ∞ if $\mathcal{F}(t) = 0$), which is left-continous and increasing on $(0, 1]$ with $\mathcal{G}(0^+) = 0$. Define the generalized inverse for any $u > 0$ by

$$\mathcal{G}^{-1}(u) := \max\{t : \mathcal{G}(t) \leq u\} \ . \tag{3.1.2}$$

Some properties of $\mathcal{F}, \mathcal{G}$ and $\mathcal{G}^{-1}$ are collected in Lemma 3.4.1. The generalized inverse $\mathcal{G}^{-1}$ will be convenient to formulate our main result in Section 3.2. We emphasize that the functions $\mathcal{F}, \mathcal{G}$ and $\mathcal{G}^{-1}$ by definition all depend on the sampling distribution ν (via the covariance operator $\bar{B} = \bar{B}_\nu$). For brevity, this dependence will be suppressed in our notation.

We shall keep the assumptions from our earlier Sections 2.2.3 and 2.2.6 with the exception of broadening the allowed window of eigenvalues.

More precisely, we shall assume (2.2.10) for the conditional expectation wrt. ρ of Y given X, the Bernstein-type assumption on the observation noise distribution (2.2.5) and the Source Condition (2.2.8). The class of spectral regularization procedures is given in Section 2.2.6. We enlarge the allowed class of marginals by introducing:

$$\mathcal{P}^{>}(\nu_*) := \{\nu \in \mathcal{P} : \ \exists j_0 \geq 1 \text{ s.th. } \frac{\mu_{2j}}{\mu_j} \geq 2^{-\nu_*} \ \ \forall j \geq j_0\} \ , \tag{3.1.3}$$

$$\mathcal{P}^{<}(\nu^*) := \{\nu \in \mathcal{P} : \ \exists j_0 \geq 1 \text{ s.th. } \frac{\mu_{2j}}{\mu_j} \leq 2^{-\nu^*} \ \ \forall j \geq j_0\} \ . \tag{3.1.4}$$

Let $\theta = (M, \sigma, R) \in \mathbb{R}_+^3$ (remember that (σ, M) are the parameters in the Bernstein moment condition

(2.2.5), in particular σ^2 is a bound on the noise variance). Then the class of models which we will consider will be defined as

$$\mathcal{M}_\theta := \mathcal{M}(\theta, r, \mathcal{P}') \;:=\; \{\, \rho(dx,dy) = \rho(dy|x)\nu(dx) \;:\; \rho(\cdot|\cdot) \in \mathcal{K}(\Omega_\nu(r,R)),\; \nu \in \mathcal{P}' \,\}\,, \tag{3.1.5}$$

with $\mathcal{P}' = \mathcal{P}^>(\nu_*)$ or $\mathcal{P}' = \mathcal{P}^<(\nu^*)$ and $1 < \nu^* \le \nu_*$.

Remark 3.1.1. *We emphasize that the classes of marginals considered in* (3.1.3) *and* (3.1.4) *are larger than the classes introduced in Section 2.2.3, but the results which we shall obtain are less refined. To clarify this point, we remark that it follows immediately from the definitions that*

$$\mathcal{P}^>(b,\alpha) \cap \mathcal{P}^<(b,\beta) \subset \mathcal{P}^> \left(b - \log_2 \left(\frac{\alpha}{\beta} \right) \right)$$

and

$$\mathcal{P}^>(b,\alpha) \cap \mathcal{P}^<(b,\beta) \subset \mathcal{P}^< \left(b - \log_2 \left(\frac{\beta}{\alpha} \right) \right).$$

We shall prove that in the setting of this chapter rates of convergence depend on the values of ν^* and ν_*. Specialized to our previous classes this means that the prefactors α, β now enter the rates of convergence (which previously only depended on the exponent b). In other words, the results of this section on the enlarged classes of marginals do not reproduce the more refined results of Section 2.3 for the previous classes of marginals.

Remark 3.1.2. *Finally, we remark that the condition in* (3.1.3) *implies a ploynomial lower bound on the eigenvalues in the following way: From $\mu_{2^k j_0} \ge 2^{-k\nu_*}\mu_{j_0}$ one has*

$$\mu_j \ge C_{j_0,\nu_*} j^{-\nu_*}\,, \qquad \text{for any } j \ge j_0\,, \tag{3.1.6}$$

with $C_{j_0,\nu_*} = j_0^{\nu_*}\mu_{j_0}$. In particular, if $\frac{\mu_{2j}}{\mu_j} \ge 2^{-\nu_*}$ holds for any $j \ge 1$, we have $\mathcal{P}^>(\nu_*) \subset \mathcal{P}^>(\nu_*, \mu_1)$. The bound (3.1.6) *will be important for deriving an adaptive estimator in Chapter 5, Example 3 , since it imposes a lower bound for the eigenvalue counting function $\mathcal{F}(\lambda)$, provided λ is sufficiently small (see Lemma 3.4.4). A similar upper bound follows from* (3.1.4)*:*

$$\mu_j \le C_{j_0,\nu^*} j^{-\nu^*}\,, \qquad \text{for any } j \ge j_0\,, \tag{3.1.7}$$

with $C_{j_0,\nu^*} = j_0^{\nu^*}\mu_{j_0}$.

3.2 Main results

Here we present our main results: Upper rates and minimax lower rates (which in the setting of this chapter turn out to be automatically *strong*), yielding minimax-optimality.

Concerning notation, we recall that, given a sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (\mathcal{X} \times \mathbb{R})^n$, we define the estimator $f_{\mathbf{z}}^\lambda$ for a suitable a-priori parameter choice $\lambda = \lambda_n$ by

$$f_{\mathbf{z}}^{\lambda_n} := g_{\lambda_n}(\bar{B}_{\mathbf{x}})\bar{S}_{\mathbf{x}}^\star \mathbf{y}\,. \tag{3.2.1}$$

Theorem 3.2.1 (Upper rate)**.** *Consider the model $\mathcal{M}_\theta := \mathcal{M}(r, R, \mathcal{P}^>(\nu^*))$ as defined in 3.1.5, where*

$r > 0$ and $\nu^* > 0$ are fixed and $\theta := (R, M, \sigma) \in \mathbb{R}^3_+$. Given a sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (\mathcal{X} \times \mathbb{R})^n$, define $f^\lambda_{\mathbf{z}}$ as in (3.2.1)*, using a regularization function of qualification $q \geq r + s$, with the parameter sequence*

$$\lambda_{n,\theta} = \min\left(\mathcal{G}^{-1}\left(\frac{\sigma^2}{R^2 n}\right), 1\right) . \tag{3.2.2}$$

Then for any $s \in [0, \frac{1}{2}]$, the sequence

$$a_{n,\theta} = R\mathcal{G}^{-1}\left(\frac{\sigma^2}{R^2 n}\right)^{r+s} \tag{3.2.3}$$

is an upper rate of convergence in L^p for all $p \geq 1$, for the interpolation norm of parameter s, for the sequence of estimated solutions $(f^{\lambda_{n,\theta}}_{\mathbf{z}})$ over the family of models $(\mathcal{M}_\theta)_{\theta \in \mathbb{R}^3_+}$, i.e.

$$\sup_{\theta \in \mathbb{R}^3_+} \limsup_{n \to \infty} \sup_{\rho \in \mathcal{M}_\theta} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\|\bar{B}^s_\nu (f_\rho - f^{\lambda_{n,\theta}}_{\mathbf{z}})\|^p_{\mathcal{H}_1}\right]^{\frac{1}{p}}}{a_{n,\theta}} < \infty .$$

Theorem 3.2.2 (Minimax lower rate)**.** *Let $r > 0$, $R > 0$, $\theta := (R, M, \sigma) \in \mathbb{R}^3_+$ and $\nu_* > 0$ be fixed. Let ν be a distribution on $\mathcal{X}$ belonging to $\mathcal{P}^>(\nu_*)$. Then the sequence $(a_{n,\theta})$ defined in* (3.2.3) *is a strong minimax lower rate of convergence in L^p for all $p \geq 1$, for the model family $\mathcal{M}_\theta := \mathcal{M}(r, R, \mathcal{P}^>(\nu_*))$, i.e.*

$$\inf_{\theta \in \mathbb{R}^3_+} \liminf_{n \to \infty} \inf_{f_\bullet} \sup_{\rho \in \mathcal{M}_\theta} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s_\nu (f_\rho - f_{\mathbf{z}})\right\|^p_{\mathcal{H}_1}\right]^{\frac{1}{p}}}{a_{n,\theta}} > 0 .$$

Corollary 3.2.3 (Minimax optimal rate)**.** *Let $r > 0$, $\nu_* \geq \nu^* > 0$ be fixed and assume $\mathcal{P}' = \mathcal{P}^<(\nu^*) \cap \mathcal{P}^>(\nu_*) \neq \emptyset$. Then the sequence of estimators $f^{\lambda_{n,\theta}}_{\mathbf{z}}$ as defined in* (3.2.1) *is strong minimax optimal in L^p for all $p \geq 1$, under the assumptions and parameter sequence* (3.2.2) *of Theorem 3.2.1 , over the class $\mathcal{M}_\theta := \mathcal{M}(r, R, \mathcal{P}')$, $\theta := (R, M, \sigma) \in \mathbb{R}^3_+$.*

3.3 Discussion

Range of applications. The assumptions we made on the spectrum decay, namely (3.1.3) and (3.1.4), are much weaker than the usual assumptions of polynomial decay on the eigenvalues. Therefore, our results establish that in this much broader situation, usual kernel regularization methods can achieve minimax rates over the regularity classes $\Omega_\nu(r, R)$. In particular, these conditions accommodate for changing behavior of the spectrum at different scales, as well as other situations delineated in the introduction. Still, our conditions do not encompass totally arbitary sequences: (3.1.3) in particular implies that the eigenvalues cannot decrease with a polynomial rate with exponent larger than ν_* . While the latter constant can be chosen arbitrary large and only results in a change of constant factor in the rates, it excludes for example exponentially decreasing eigenvalues.

While we do not prove results about exponential eigenvalue decay on the level of sharpness of our results in this thesis (in particular, keeping track of the noise variance and the radius in the source condition), we are convinced that - with some additional work - such results are entirely accessible using only techniques of the present thesis (at least staying within the framework of our Hölder type source conditions). For

somewhat less elaborate results in this direction we refer to [62] . However, we remark that in such cases (typically observed when using the Gaussian kernel), in practice the kernel parameters (e.g. the bandwidth) are also tuned in addition to λ , which might reflect the fact that for badly tuned bandwidth of the kernel, tuning of the regularization parameter λ alone might not give satisfactory (minimax) results. On the other hand, the results we obtained might provide an additional motivation for using "rougher" kernels than Gaussian, leading to a softer decay of eigenvalues, in which case minimax adaptivity is at hand over a large regularity class. The latter is true however only if the qualification of the method is large, which is not the case for the usual kernel ridge regression: hence, rougher kernels should be used with methods having a large qualification (for instance L^2 boosting).We emphasize, however, that in the inverse learning setting - in contradistinction to the direct learning setting - one ist not free to choose the kernel; a given kernel is then part of the overall problem.

As a general remark we want to add that while we expect deep relations between a given structure of data and types of eigenvalue decay, a precise analysis of such relations is very much open both on the level of examination of special examples and general mathematical theorems.

Adaptivity. Our results establish the existence of a suitable regularization parameter λ such that the associated estimator attains the minimax rate if the regularity class parameter are known in advance. The latter is of course not realistic, but in the case of L^2 (prediction) error, the principle of using a grid for λ and then using a hold-out sample to select a value of λ from the data is known to be able to select a value close to the optimal choice in a broad domain of situations (see, for instance, [21]), so that in principle we can generalize our results to data-dependent minimax adaptivity even in the absence of a priori knowledge of the regularity parameters. For a much more detailed discussion of adaptivity we refer to the last chapter of this thesis. In particular, using a (new) version of the balancing principle or Lepskii's method adapted to the framework of this thesis - which is different from hold-out - we shall obtain estimates in $\mathcal{H}_1$-norm and all interpolating norms.

3.4 Proofs

The proof of Theorem 3.2.1 and Theorem 3.2.2 will be given not only in both $\mathcal{H}_1$- norm and $L^2(\nu)$-norm, but also for all intermediate norms $||B^s f||_{\mathcal{H}_1}$, where $s \in [0, 1/2]$. Note that $s = 0$ corresponds to $\mathcal{H}_1$-norm, while $s = 1/2$ corresponds to $L^2(\nu)$-norm.

To simplify notation we will adopt the following conventions: The dependence of multiplicative constants C on various parameters will (generally) be omitted, except for σ, M, R, η and n which we want to track precisely . The expression "for n big enough" means that the statement holds for $n \geq n_0$, with n_0 potentially depending on all model parameters (including σ, M and R), but not on η .

3.4.1 Preliminaries

We start by collecting some useful properties for the functions $\mathcal{F}$ and $\mathcal{G}$ in the following lemma.

Lemma 3.4.1. *1. Let $c \leq 1$ be fixed. Then for any t,*

$$\mathcal{G}(ct) \leq c\mathcal{G}(t)\,.$$

2. Assume $\nu \in \mathcal{P}^<(\nu^*)$ holds. Let $C \geq 1$ be fixed. Then for any t small enough,

$$\mathcal{F}(t) \leq 4C^{\frac{1}{\nu^*}}\mathcal{F}(Ct) \qquad \text{and} \qquad \mathcal{G}(Ct) \leq 4C^{2r+1+\frac{1}{\nu^*}}\mathcal{G}(t)\,.$$

3. Assume $\nu \in \mathcal{P}^<(\nu^*)$. For any $u > 0$ it holds $\mathcal{G}\big(\mathcal{G}^{-1}(u)\big) \leq u$ and for u small enough,

$$\mathcal{G}\big(\mathcal{G}^{-1}(u)\big) \geq \frac{u}{4}\,.$$

Proof of Lemma 3.4.1. For point 1 of the Lemma, let $c \leq 1$ be fixed; just write by definition of $\mathcal{G}$ and the fact that $\mathcal{F}$ is nonincreasing:

$$\mathcal{G}(ct) = \frac{c^{2r+1}t^{2r+1}}{\mathcal{F}(ct)} \leq c^{2r+1}\mathcal{G}(t) \leq c\mathcal{G}(t)\,.$$

For point 2, let $j_0 \geq 1$ such that $\frac{\mu_{2j}}{\mu_j} \leq 2^{-\nu^*}$ for any $j \geq j_0$ and let t_0 be small enough such that $\mathcal{F}(t_0) \geq j_0$.

Let $C \geq 1$ be fixed and $t \leq C^{-1}t_0$, so that $k_t := \mathcal{F}(Ct) \geq j_0$. By definition $\mu_{k_t+1} < Ct$. Furthermore for any $i \geq 1$ we have $\mu_{2^i(k_t+1)} \leq 2^{-i\nu^*}\mu_{(k_t+1)}$ by repetition. Choosing $i := 1 + \lfloor \frac{\log_2 C}{\nu^*} \rfloor$, we have $2^{-i\nu^*}C \leq 1$ and $2^i \leq 2C^{\frac{1}{\nu^*}}$. Combining the first inequality with what precedes we deduce $\mu_{2^i(k_t+1)} < t$ and thus $\mathcal{F}(t) \leq 2^i(k_t+1) - 1 \leq 4C^{\frac{1}{\nu^*}}\mathcal{F}(Ct)$. We deduce

$$\mathcal{G}(Ct) = \frac{C^{2r+1}t^{2r+1}}{\mathcal{F}(Ct)} \leq \frac{4C^{2r+1+\frac{1}{\nu^*}}t^{2r+1}}{\mathcal{F}(t)} = 4C^{2r+1+\frac{1}{\nu^*}}\mathcal{G}(t)\,.$$

We turn to point 3. Since $\mathcal{G}$ is left-continuous, the supremum in the definition (2.2) of its inverse $\mathcal{G}^{-1}$ is indeed a maximum (also the set over which the max is taken is nonempty since $\mathcal{G}(0^+) = 0$ and $u > 0$), and therefore must satisfy $\mathcal{G}(\mathcal{G}^{-1}(u)) \leq u$.

Consider now $\mathcal{F}(t^+) := \#\{j \in \mathcal{N} : \mu_i > t\}$, let t_0' be small enough such that $\mathcal{F}(t_0') \geq 2j_0$, and assume $t < \min(t_0', \mu_1)$. The second component of the latter minimum ensures $\mathcal{F}(t^+) \geq 1$. If $t \notin \{\mu_i, i \geq 1\}$, then $\mathcal{F}$ is continous in t and $\mathcal{F}(t) = \mathcal{F}(t^+)$. Otherwise, $t = \mu_k$ with $k = \mathcal{F}(t) \geq 2j_0$, so that $\frac{\mu_k}{\mu_{\lfloor \frac{k}{2} \rfloor}} \leq 2^{-\nu^*} < 1$, that is to say $t < \mu_{\lfloor \frac{k}{2} \rfloor}$, implying $F(t^+) \geq \lfloor \frac{k}{2} \rfloor \geq \frac{1}{2}\mathcal{F}(t) - 1$ and finally $\mathcal{F}(t) \leq 4\mathcal{F}(t^+)$.

Consider now u small enough such that $t = \mathcal{G}^{-1}(u) < \min(t_0', \mu_1)$ as above. Then $\mathcal{G}(t^+) \geq u$ and

$$\mathcal{G}(\mathcal{G}^{-1}(u)) = \mathcal{G}(t) = \frac{t^{2r+1}}{\mathcal{F}(t)} \geq \frac{1}{4}\frac{t^{2r+1}}{\mathcal{F}(t^+)} = \frac{1}{4}\mathcal{G}(t^+) \geq \frac{u}{4}\,.$$

□

Lemma 3.4.2 (Effective dimensionality, upper bound)**.** *Assume $\nu \in \mathcal{P}^<(\nu^*)$. The effective dimension*

(defined in (2.1.2)*) satisfies for any* λ *sufficiently small and some* $C_{\nu^*} < \infty$

$$\mathcal{N}(\lambda) \leq C_{\nu^*} \mathcal{F}(\lambda) \,. \tag{3.4.1}$$

Proof of Lemma 3.4.2. Let $j_0 \geq 1$ such that $\frac{\mu_{2j}}{\mu_j} \leq 2^{-\nu^*}$ for any $j \geq j_0$ and let λ_0 be small enough such that $\mathcal{F}(\lambda_0) \geq j_0$. For $\lambda \leq \lambda_0$, denote $j_\lambda := \mathcal{F}(\lambda)$. Then, using $\mu_j < \mu_j + \lambda$ and $\lambda < \mu_j + \lambda$, we obtain

$$\mathcal{N}(\lambda) = \sum_{j=1}^{j_\lambda} \frac{\mu_j}{\mu_j + \lambda} + \sum_{j > j_\lambda} \frac{\mu_j}{\mu_j + \lambda} \leq j_\lambda + \frac{1}{\lambda} \sum_{j > j_\lambda} \mu_j \,.$$

Focussing on the tail sum we see that

$$\begin{aligned} \sum_{j \geq j_\lambda} \mu_j = \sum_{l=0}^{\infty} \sum_{j=(j_\lambda+1)2^l}^{(j_\lambda+1)2^{l+1}-1} \mu_j \leq (j_\lambda + 1) \mu_{j_\lambda+1} \sum_{l=0}^{\infty} 2^l 2^{-l\nu*} = (j_\lambda + 1) \mu_{j_\lambda+1} (1 - 2^{1-\nu*})^{-1} \\ \leq \lambda (\mathcal{F}(\lambda) + 1)(1 - 2^{1-\nu*})^{-1} \,, \end{aligned}$$

where the first inequality comes from the fact that the sequence $(\mu_j)_j$ is decreasing and by repetition; and the second inequality comes from the definition of j_λ. Collecting all ingredients we obtain for any $\lambda \leq \lambda_0$:

$$\mathcal{N}(\lambda) \leq \mathcal{F}(\lambda)(1 + 2(1 - 2^{1-\nu*})^{-1}) \,.$$

□

Lemma 3.4.3 (Effective dimensionality, lower bound)**.** *The effective dimension (defined in* (2.1.2)*) satisfies for any* $\lambda \in [0, 1]$ *and* $\nu \in \mathcal{P}$

$$\mathcal{N}(\lambda) \geq \frac{1}{2} \mathcal{F}(\lambda) \,. \tag{3.4.2}$$

Proof. For $\lambda \in [0, 1]$ denote $j_\lambda := \mathcal{F}(\lambda)$. Since $\mu_j > \lambda$ for any $j \leq j_\lambda$, one has

$$\mathcal{N}(\lambda) = \sum_{j=1}^{j_\lambda} \frac{\mu_j}{\mu_j + \lambda} + \sum_{j > j_\lambda} \frac{\mu_j}{\mu_j + \lambda} \geq \sum_{j=1}^{j_\lambda} \frac{\mu_j}{\mu_j + \lambda} \geq \frac{1}{2} j_\lambda = \frac{1}{2} \mathcal{F}(\lambda) \,.$$

□

Lemma 3.4.4 (Counting Function, lower bound)**.** *Assume* $\nu \in \mathcal{P}^{>}(\nu_*)$*. Then for any* λ *small enough*

$$\mathcal{F}(\lambda) \geq C'_{\nu_*} \lambda^{-\frac{1}{\nu_*}} \,,$$

for some $C'_{\nu_*} > 0$*.*

Proof. Assume $\frac{\mu_{2j}}{\mu_j} \geq 2^{-\nu_*}$ holds for any $j \geq j_0$. Denote $j_\lambda := \mathcal{F}(\lambda)$. Let $\lambda_0 \in (0, 1]$ such that $j_\lambda \geq j_0$ for any $\lambda \leq \lambda_0$. Then, by definition of j_λ and from (3.1.6), for any $\lambda \leq \lambda_0$

$$\mu_{j_\lambda} \geq \max\left(\lambda, C_{\nu_*} j_\lambda^{-\nu_*}\right) \,.$$

In case $C_{\nu_*} j_\lambda^{-\nu_*} \geq \lambda$ we have the chain of inequalities

$$\mu_{j_\lambda} \geq C_{\nu_*} j_\lambda^{-\nu_*} \geq \lambda > \mu_{j_\lambda+1} \geq C_{\nu_*} (j_\lambda + 1)^{-\nu_*} \ .$$

Since $1 \leq j_\lambda$, we obtain $(j_\lambda+1)^{-1} \geq \frac{1}{2} j_\lambda^{-1}$ and thus $\lambda > C_{\nu_*} 2^{-\nu_*} j_\lambda^{-\nu_*}$, implying the result. If $\lambda \geq C_{\nu_*} j_\lambda^{-\nu_*}$ we immediately have the bound. □

Corollary 3.4.5 (Counting Function, upper bound)**.** *Assume $\nu \in \mathcal{P}^<(\nu^*)$. Then for any λ small enough*

$$\mathcal{F}(\lambda) \leq C_{\nu^*} \lambda^{-\frac{1}{\nu^*}} \ , \quad C_{\nu^*} = 4 \mu_1^{\frac{1}{\nu^*}} \ .$$

Proof. The proof follows from Lemma 3.4.1 by setting $C = \mu_1 \frac{1}{\lambda}$, which is larger than 1 provided λ is sufficiently small. Then

$$\mathcal{F}(\lambda) \leq 4 \left(\frac{\mu_1}{\lambda} \right)^{\frac{1}{\nu^*}} \mathcal{F}(\mu_1) \ .$$

Note that $\mathcal{F}(\mu_1) = 1$. □

3.4.2 Proof of upper rate

The proof of Theorem 3.2.1 relies on the non-asymptotic result, established in Section 2.5, Proposition 2.5.2 and Proposition 2.5.3. We repeat these results in streamlined form for better understanding.

Proposition 3.4.6. *Let $s \in [0, \frac{1}{2}]$ and assume $f_\rho \in \Omega_\nu(r, R)$, with $\nu \in \mathcal{P}$, $r > 0$ and $R > 0$. Let $f_{\mathbf{z}}^\lambda$ be defined as in* (3.2.1) *using a regularization function of qualification $q \geq r + s$. Then, for any $\eta \in (0, 1)$, $\lambda \in (0, 1]$ and $n \in \mathbb{N}$ satisfying*

$$n \geq 64 \lambda^{-1} \max(\mathcal{N}(\lambda), 1) \log^2 (8/\eta) \ , \tag{3.4.3}$$

we have with probability at least $1 - \eta$:

$$\left\| \bar{B}_\nu^s (f_\rho - f_{\mathbf{z}}^\lambda) \right\|_{\mathcal{H}_1} \leq C_\blacktriangle \ \log(8\eta^{-1}) \lambda^s \left(R \left(\lambda^r + \frac{1}{\sqrt{n}} \mathbb{1}_{(1,\infty)}(r) \right) + \left(\frac{M}{n\lambda} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n\lambda}} \right) \right) \ . \tag{3.4.4}$$

Proof of Theorem 3.2.1: Let all assumptions of Theorem 3.2.1 be satisfied. Provided n is big enough, we have $\mathcal{F}(\lambda_n) \geq 1$ and by Lemma 3.4.2 it holds $\mathcal{N}(\lambda_n) \leq C. \lambda_n^{2r+1} / \mathcal{G}(\lambda_n)$, following from the definition of $\mathcal{G}$. By the definition of λ_n and by Lemma 3.4.1, (iii), for n sufficiently large, $\mathcal{G}(\lambda_n) \geq C_{\sigma,R} \frac{1}{n}$, so assumption (3.4.3) is satisfied if $\log(8/\eta) \leq C_{\sigma,R} \lambda_n^{-r}$. Hence, with probability at least $1 - \eta$

$$\left\| \bar{B}^s (f_\rho - f_{\mathbf{z}}^{\lambda_n}) \right\|_{\mathcal{H}_1} \leq C_\blacktriangle \ \log(8\eta^{-1}) \lambda_n^s \left(R \left(\lambda_n^r + \frac{1}{\sqrt{n}} \mathbb{1}_{(1,\infty)}(r) \right) + \left(\frac{M}{n\lambda_n} + \sigma \sqrt{\frac{\lambda_n^{2r}}{n\mathcal{G}(\lambda_n)}} \right) \right) \ . \tag{3.4.5}$$

Observe that the choice (3.2.2) implies that $n^{-\frac{1}{2}} = o(\lambda_n^r)$, since $\sigma^2 / R^2 n = \mathcal{G}(\lambda_n) \leq \lambda_n^{2r+1}$. Therefore, up to requiring n large enough and multiplying the front factor by 2 , we can disregard the term $1/\sqrt{n}$ in the second factor of the above bound. Similarly, one can readily check that

$$\frac{M}{n\lambda_n} = o \left(\sqrt{\frac{\lambda_n^{2r}}{n\mathcal{G}(\lambda_n)}} \right) \ ,$$

so that we can also disregard the term $(n\lambda_n)^{-1}$ for n large enough and concentrate on the two remaining main terms of the upper bound in (3.4.5), which are $R\lambda_n^r$ and $\sigma\lambda_n^r\mathcal{G}(\lambda_n)^{-\frac{1}{2}}n^{-\frac{1}{2}}$. The proposed choice of λ_n balances precisely these two terms and easy computations lead to

$$\left\|\bar{B}^s(f_\rho - f_{\mathbf{z}}^{\lambda_n})\right\|_{\mathcal{H}_1} \leq C_{\blacktriangle}\ \log(8\eta^{-1})R\lambda_n^{r+s}\,, \tag{3.4.6}$$

with probability at least $1-\eta$, provided $\eta \geq \eta_n := 8\exp\left(-C_{\sigma,R}\lambda_n^{-r}\right)$ and provided n is sufficiently large. Note that λ_n^{-r} is increasing and so $\eta_n \to 0$ as $n \to \infty$.

For establishing a less accurate bound which covers the whole interval $(0,1]$ we may use (2.5.5) and 2.5.13 from the previous chapter. We conclude that

$$\mathbb{P}\Big[\ \left\|\bar{B}^s(f_\rho - f_{\mathbf{z}}^{\lambda})\right\|_{\mathcal{H}_1} \geq a' + b'\log\eta^{-1}\Big] \leq \eta\,,$$

for all $\eta \in (0,1]$, with $a' := C_{\sigma,M,R}\max\left(\frac{1}{\lambda\sqrt{n}},1\right)$ and $b' := \frac{C_{\sigma,M}}{\lambda\sqrt{n}}$. The result follows exactly as in the proof of Theorem 2.3.4 in the regular case by applying Corollary A.3.2. □

3.4.3 Proof of minimax lower rate

Let $r>0$, $R>0$ and $s \in [0,1/2]$ be fixed. Assume the generating distribution ρ belongs to the class $\mathcal{M}_\theta := \mathcal{M}(\theta, r, \mathcal{P}')$, with $\theta = (M,\sigma,R) \in \mathbb{R}_+^3$ and $\mathcal{P}' = \mathcal{P}^{>}(\nu_*)$, as defined in (3.1.5). In order to obtain minimax lower bounds, we proceed as in the previous chapter by applying the general reduction scheme from Section A.5. Again, The main idea is to find N_ε functions $f_1,\ldots,f_{N_\varepsilon}$ belonging to the source sets $\Omega_\nu(r,R)$, depending on ε sufficiently small, with $N_\varepsilon \to \infty$ as $\varepsilon \to 0$, such that any two of these functions are ε-separated with respect to $||\bar{B}^s(\cdot)||_{\mathcal{H}_1}$- norm, but such that the associated distributions $\rho_{f_j} =: \rho_j \in \mathcal{M}$ (see definition (3.4.8) below) have small Kullback-Leibler divergence $\mathcal{K}$ to each other and are therefore statistically close. We shall use

$$\inf_{\hat{f}_\bullet}\sup_{\rho\in\mathcal{M}_\theta}\mathbb{E}_{\mathbf{z}\sim\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho - f_{\mathbf{z}})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} \geq \varepsilon\inf_{\hat{f}_\bullet}\max_{1\leq j\leq N_\varepsilon}\rho_j\left[\left\|\bar{B}^s(f_\rho - \hat{f}_{\mathbf{z}})\right\|_{\mathcal{H}_1} \geq \varepsilon\right]\,, \tag{3.4.7}$$

(see (A.5.1)) , where the infimum is taken over all estimators $\hat{f}_\bullet$ of f_ρ. The above RHS is then lower bounded through Proposition A.5.1 given in Section A.5 which is a consequence of Fano's lemma.

We will establish the lower bounds in the particular case where the distribution of Y given X is Gaussian with variance σ^2 (which satisfies the Bernstein noise assumption (2.2.5) with $M=\sigma$) . The main effort is to construct a finite subfamily belonging to the model of interest and suitably satisfying the assumptions of Proposition A.5.1 in Section A.5. More precisely, to each $f \in \Omega_\nu(r,R)$ and $x \in \mathcal{X}$ we associate the following measure:

$$\rho_f(dx,dy) := \rho_f(dy|x)\nu(dx)\,,\ \text{where } \rho_f(dy|x) := \mathcal{N}(\bar{S}_x f, \sigma^2)\,. \tag{3.4.8}$$

Then the measure ρ_f belongs to the class $\mathcal{M}(\theta,r,\mathcal{P}^{>}(\nu_*))$ and given $f_1,f_2 \in \Omega_\nu(r,R)$, the Kullback divergence between ρ_1 and ρ_2 satisfies

$$\mathcal{K}(\rho_1,\rho_2) = \frac{1}{2\sigma^2}\left\|\sqrt{B}(f_1-f_2)\right\|_{\mathcal{H}_1}^2\,. \tag{3.4.9}$$

We will need the following Proposition:

Proposition 3.4.7. *For any $\varepsilon > 0$ sufficiently small (depending on the parameters ν_*, r, R, s), there exist $N_\varepsilon \in \mathbb{N}$ and functions $f_1, \dots, f_{N_\varepsilon} \in \Omega_\nu(r, R)$ satisfying*

1. *For any $i, j = 1, \dots, N_\varepsilon$ with $i \neq j$ one has $\| \bar{B}^s(f_i - f_j) \|^2_{\mathcal{H}_1} > \varepsilon^2$ and*

$$\mathcal{K}(\rho_i, \rho_j) \leq C_{\nu_*, s}\, R^2 \sigma^{-2} \left(\frac{\varepsilon}{R}\right)^{\frac{2r+1}{r+s}} , \tag{3.4.10}$$

2. $\log(N_\varepsilon - 1) \geq \frac{1}{36} \mathcal{F}(2^{\nu_*} \left(\frac{\varepsilon}{R}\right)^{\frac{1}{r+s}})$.

Proof of Proposition 3.4.7. We recall that we denote $(e_l)_{l \in \mathbb{N}}$ an orthonormal family of $\mathcal{H}_1$ of eigenvectors of $\bar{B}$, corresponding to the eigenvalues $(\mu_l)_{l \in \mathbb{N}}$. Assume the sampling distribution belongs to $\mathcal{P}^{>}(\nu_*)$, i.e. $\frac{\mu_{2j}}{\mu_j} \geq 2^{-\nu_*}$ for any $j \geq j_0$, for some $j_0 \in \mathbb{N}$. Let $max := \max(28, j_0)$. Choose $\varepsilon < 2^{-\nu_*(r+s)} R\, \mu_{max}$ and pick $m = m(\varepsilon) := \mathcal{F}(2^{\nu_*}(\varepsilon R^{-1})^{\frac{1}{r+s}})$. Note that $m \geq 28$, following from the choice of ε, so Lemma 2.6.2 applies.

Let $N_m > 3$ and $\pi_1, \dots, \pi_{N_m} \in \{-1, +1\}^m$ be given by Lemma 2.6.2 and define

$$g_i := \frac{\varepsilon}{\sqrt{m}} \sum_{l=m+1}^{2m} \pi_i^{(l-m)} \left(\frac{1}{\mu_l}\right)^{r+s} e_l \, . \tag{3.4.11}$$

We have by the definition of m

$$\|g_i\|^2_{\mathcal{H}_1} = \frac{\varepsilon^2}{m} \sum_{l=m+1}^{2m} \left(\frac{1}{\mu_l}\right)^{2(r+s)} \leq \varepsilon^2 \mu_{2m}^{-2(r+s)} \leq \varepsilon^2 2^{2\nu_*(r+s)} \mu_m^{-2(r+s)} \leq R^2 \, .$$

For $i = 1, \dots, N_m$ let $f_i := \bar{B}^r g_i \in \Omega_\nu(r, R)$, with g_i as in (3.4.11). Then

$$\| \bar{B}^s(f_i - f_j) \|^2_{\mathcal{H}_1} \geq \varepsilon^2 \, ,$$

as a consequence of Lemma 2.6.2. For $i = 1, \dots, N_\varepsilon$, let $\rho_i = \rho_{f_i}$ be defined by (3.4.8). Then, using the definition of m, the Kullback divergence satisfies

$$\begin{aligned} \mathcal{K}(\rho_i, \rho_j) = \frac{1}{2\sigma^2} \left\| \sqrt{\bar{B}}(f_i - f_j) \right\|^2_{\mathcal{H}_1} &\leq (2\sigma)^{-2} \mu_{m+1}^{1-2s} \varepsilon^2 \\ &\leq 2^{\nu_*(1-2s)}\, (2\sigma^2)^{-1} R^2 \left(\frac{\varepsilon}{R}\right)^{\frac{1+2r}{r+s}} , \end{aligned}$$

which completes the proof of the first part. Finally, again by Lemma 2.6.2 and the definition of m

$$\log(N_m - 1) \geq \frac{m}{36} = \frac{1}{36} \mathcal{F}\left(2^{\nu_*} \, \left(\varepsilon R^{-1}\right)^{\frac{1}{r+s}}\right)$$

and the proof is complete. □

Proof of Theorem 3.2.2: Our aim is to apply Proposition A.5.1 and we will check that all required conditions are satisfied. From Proposition 3.4.7 we deduce that for any ε sufficiently small, there exists N_ε and functions $f_1, \dots, f_{N_\varepsilon} \in \Omega_\nu(r, R)$ fulfilling points 1 and 2 . The first part of point 1 gives requirement

(i) of Proposition A.5.1. Requirement (ii) follows directly from (3.4.10) and from point 2 in Proposition 3.4.7:

$$\frac{1}{N_\varepsilon - 1} \sum_{j=1}^{N_\varepsilon - 1} \mathcal{K}(\rho_j^{\otimes n}, \rho_{N_\varepsilon}^{\otimes n}) \leq n36 C'_{\nu_*,s}\, R^2\sigma^{-2}\, \mathcal{G}\left(2^{\nu_*}\left(\varepsilon R^{-1}\right)^{\frac{1}{r+s}}\right) \log(N_\varepsilon - 1)$$
$$=: \omega\, \log(N_\varepsilon - 1)\,,$$

with $C'_{\nu_*,s} = 2^{-2\nu_*(r+s)-1} < 1$. Define $\varepsilon := 2^{-\nu_*}\frac{R}{288}\mathcal{G}^{-1}(\frac{\sigma^2}{R^2 n})^{r+s}$, then by Lemma 3.4.1, the requirements of Proposition A.5.1 will hold (in particular, $\omega < 1/8$) for any n sufficiently large and

$$\inf_{\hat{f}_\bullet} \max_{1\leq j\leq N_\varepsilon} \rho_j^{\otimes n}\left(\left\|\bar{B}_\nu^s(\hat{f}_\bullet - f_j)\right\|_{\mathcal{H}_1} \geq \frac{\varepsilon}{2}\right) \geq \frac{\sqrt{N_\varepsilon - 1}}{1+\sqrt{N_\varepsilon - 1}}\left(1 - 2\omega - \sqrt{\frac{2\omega}{\log\left(N_\varepsilon - 1\right)}}\right) > 0\,.$$

Taking the liminf finishes the proof. □

Chapter 4

Distributed Learning

In this chapter we treat distributed learning algorithms in the regular case (i.e. polynomial decay of eigenvalues of the covariance operator, Hölder-type source condition) which represent one strategy to deal with large data sets. A central problem is in analyzing how far the original sample may be divided into subsamples without destroying the minimax optimality for the rates of convergence and the appropriate choice of regularization parameter. The maximal number of subsamples as a fraction of the total sample size consistent with this requirement is not yet known. Therefore we complement our analytical result (giving a sufficient condition) by numerical experiments with the aim of finding evidence for necessity.

After explaining the basic algorithm in Section 4.1, we present our main results in Section 4.2 and our numerical analysis in Section 4.3. Following a brief discussion in Section 4.4 we provide proofs of our main results in Section 4.5.

4.1 Distributed Learning Algorithm

We let $D = \{(x_j, y_j)\}_{j=1}^n \subset (\mathcal{X}\times\mathcal{Y})^n$ be the dataset, which we partition into m disjoint subsets $D_1, ..., D_m$, each having size $\frac{n}{m}$. In the following, we assume that the total sample size n is divisible by m. Denote the jth data vector, $1 \leq j \leq m$, by $(\mathbf{x}_j, \mathbf{y}_j) \in (\mathcal{X} \times \mathbb{R})^{\frac{n}{m}}$. On each subset we compute a local estimator for a suitable *a priori* parameter choice $\lambda = \lambda_n$ according to

$$f_{D_j}^{\lambda_n} := g_{\lambda_n}(\bar{B}_{\mathbf{x}_j})\bar{S}_{\mathbf{x}_j}^{\star}\mathbf{y}_j \ . \tag{4.1.1}$$

By f_D^λ we will denote the estimator using the whole sample $m = 1$. The final estimator is given by simply averaging the local ones:

$$\bar{f}_D^\lambda := \frac{1}{m}\sum_{j=1}^m f_{D_j}^\lambda \ . \tag{4.1.2}$$

4.2 Main Results

This section presents our main results. Theorem 4.2.1 and Theorem 4.2.2 contain separate estimates on the approximation error and the sample error and lead to Corollary 4.2.3 which gives an upper bound for the error $\left\|\bar{B}^s(f_\rho - \bar{f}_D^\lambda)\right\|_{\mathcal{H}_1}$ and presents an upper rate of convergence for the sequence of distributed learning algorithms.

The minimax optimal rate for the single machine problem as presented in Theorem 2.3.4 yields an estimate on the difference between the single machine and the distributed learning algorithm in Corollary 4.2.4.

As in the previous Chapter 2 , we want to track the precise behavior of these rates not only for what concerns the exponent in the number of examples n, but also in terms of their scaling (multiplicative constant) as a function of the noise variance σ^2 and the complexity radius R in the source condition. For this reason, we again introduce a notion of a family of rates over a family of models. More precisely, we consider an indexed family $(\mathcal{M}_\theta)_{\theta\in\Theta}$, where for all $\theta \in \Theta$, $\mathcal{M}_\theta$ is a class of Borel probability distributions on $\mathcal{X} \times \mathbb{R}$ satisfying the basic general assumptions (2.2.4) and (2.2.5). We consider rates of convergence in the sense of the p-th moments of the estimation error, where $1 \leq p < \infty$ is a fixed real number.

Our proofs are based on a classical bias-variance decomposition as follows: Introducing

$$\tilde{f}_D^\lambda = \frac{1}{m}\sum_{j=1}^m g_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}_{\mathbf{x}_j} f_\rho \,, \tag{4.2.1}$$

we write

$$\begin{aligned} \bar{B}^s(f_\rho - \bar{f}_D^\lambda) &= \bar{B}^s(\, f_\rho - \tilde{f}_D^\lambda\,) + \bar{B}^s(\, \tilde{f}_D^\lambda - \bar{f}_D^\lambda\,) \\ &= \frac{1}{m}\sum_{j=1}^m \bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j}) f_\rho + \frac{1}{m}\sum_{j=1}^m \bar{B}^s g_\lambda(\bar{B}_{\mathbf{x}_j})(\bar{B}_{\mathbf{x}_j} f_\rho - S^*_{\mathbf{x}_j}\mathbf{y}_j) \,. \end{aligned} \tag{4.2.2}$$

In all the forthcoming results in this section, we let $s \in [0, \frac{1}{2}]$, $p \geq 1$ and consider the model $\mathcal{M}_{\sigma,M,R} := \mathcal{M}(r, R, \mathcal{P}^<(b,\beta))$ where $r > 0$, $b > 1$ and $\beta > 0$ are fixed, and $\theta = (R, M, \sigma)$ varies in $\Theta = \mathbb{R}^3_+$. Given a sample $D \subset (\mathcal{X} \times \mathbb{R})^n$ of size n, define $\bar{f}_D^{\lambda_n}$, $f_D^{\lambda_n}$ as in Section 4.1 and $\tilde{f}_D^{\lambda_n}$ as in (4.2.1), using a regularization function g_λ of qualification $q \geq r + s$, with parameter sequence

$$\lambda_n := \lambda_{n,(\sigma,R)} := \min\left(\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{b}{2br+b+1}}, 1\right) \,, \tag{4.2.3}$$

independent of M. Furthermore, define the sequence

$$a_n := a_{n,(\sigma,R)} := R\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{b(r+s)}{2br+b+1}} \,. \tag{4.2.4}$$

We recall from the introduction that we shall always assume that n is a multiple of m. With these preparations, our main results are:

Theorem 4.2.1 (Approximation Error)**.** *If the number m of subsample sets satisfies*

$$m \leq n^{\alpha} \,, \quad \alpha < \frac{2b \min\{r,1\}}{2br+b+1} \,, \tag{4.2.5}$$

then

$$\sup_{(\sigma,M,R)\in\mathbb{R}^3_+} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho - \tilde{f}_D^{\lambda_n})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}}}{a_n} < \infty \,.$$

Theorem 4.2.2 (Sample Error)**.** *If the number m of subsample sets satisfies*

$$m \leq n^{\alpha} \,, \quad \alpha < \frac{2br}{2br+b+1} \,, \tag{4.2.6}$$

then

$$\sup_{(\sigma,M,R)\in\mathbb{R}^3_+} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(\tilde{f}_D^{\lambda_n} - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}}}{a_n} < \infty \,.$$

And, as a consequence (by (4.2.2) and by applying the triangle inequality):

Corollary 4.2.3. *If the number m of subsample sets satisfies*

$$m \leq n^{\alpha} \,, \quad \alpha < \frac{2b \min\{r,1\}}{2br+b+1} \,, \tag{4.2.7}$$

then the sequence (4.2.4) is an upper rate of convergence in L^p, for the interpolation norm of parameter s, for the sequence of estimated solutions $(\bar{f}_D^{\lambda_{n,(\sigma,R)}})$ over the family of models $(\mathcal{M}_{\sigma,M,R})_{(\sigma,M,R)\in\mathbb{R}^3_+}$, i.e.

$$\sup_{(\sigma,M,R)\in\mathbb{R}^3_+} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}}}{a_n} < \infty \,.$$

Combining Corollary 4.2.3 with Theorem 2.3.4 immediately yields:

Corollary 4.2.4. *If the number m of subsample sets satisfies*

$$m \leq n^{\alpha} \,, \quad \alpha < \frac{2b \min\{r,1\}}{2br+b+1} \,, \tag{4.2.8}$$

then

$$\sup_{(\sigma,M,R)\in\mathbb{R}^3_+} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_D^{\lambda_n} - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}}}{a_n} < \infty \,.$$

4.3 Numerical Studies

In this section we numerically study the reconstruction error (corresponding to $s = 0$) in Corollary 4.2.3 (in expectation with $p = 2$) both in the single machine and distributed learning setting. Our main interest is in finding numerical evidence for the "optimality" of our theoretical exponent α parametrizing the size of subsamples in terms of the total sample size, $m = n^{\alpha}$. In addition we shall demonstrate in which way parallelization serves as an additional regularization.

More specifically, we shall analyze our Example 2.2.6 (*Differentiating a real function*) which we recall here. We set $\mathcal{H}_1 = \{f \in L^2[0,1] : \mathbb{E}[f] = 0\}$ and let $A : \mathcal{H}_1 \longrightarrow \mathrm{Im}(A) = H_0^1[0,1]$ be given by

$$[Af](x) = \int_0^x f(t)\, dt = \langle f, F_x \rangle_{L^2} \ , \tag{4.3.1}$$

where $F_x(t) = \mathbb{1}_{[0,x]}(t) - x$. The kernel is given by $K(x,t) = \langle F_x, F_t \rangle_{L^2} = x \wedge t - xt$.
For all experiments in this section, we simulate data from the inverse regression model

$$Y_i = Af_\rho(X_i) + \epsilon_i \ , \qquad i = 1, ..., n \ ,$$

where the input variables $X_i \sim Unif[0,1]$ are uniformly distributed and the noise variables $\varepsilon_i \sim N(0, \sigma^2)$ are normally distributed with standard deviation $\sigma = 0.005$. We choose the target function f_ρ according to two different cases, namely $r < 1$ (*low smoothness*) and $r = \infty$ (*high smoothness*). To accurately determine the degree of smoothness $r > 0$, we apply Proposition 4.3.1 below by explicitly calculating the Fourier coefficients $(\langle f_\rho, e_j \rangle_{\mathcal{H}_1})_{j \in \mathbb{N}}$, where $e_j(x) = \sqrt{2}\cos(\pi j x)$, for $j \in \mathbb{N}^*$, forms an ONB of $\mathcal{H}_1$. Recall that the rate of eigenvalue decay is explicitly given by $b = 2$, meaning that we have full control over all parameters in (4.2.8). From [34] we need

Proposition 4.3.1. *Let $\mathcal{H}_1, \mathcal{H}_2$ be separable Hilbert spaces and $S : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ be a compact linear operator with singular system $\{\sigma_j, \varphi_j, \psi_j\}$[1]. Denoting by $S^\dagger$ the generalized inverse [2] of S, one has for any $r > 0$ and $g \in \mathcal{H}_2$:*

g is in the domain of $S^\dagger$ and $S^\dagger g \in \mathrm{Im}((S^*S)^r)$ if and only if

$$\sum_{j=0}^{\infty} \frac{|\langle g, \psi_j \rangle_{\mathcal{H}_2}|^2}{\sigma_j^{2+4r}} \ < \ \infty \ .$$

In our case, $\mathcal{H}_1$ is as above, $\mathcal{H}_2$ is $L^2([0,1])$ with Lebesgue measure and $S = A$ (as a map from $\mathcal{H}_1$ to $\mathcal{H}_2$). Thus $\mathrm{Im}(S) = H_0^1[0,1]$ is dense in $\mathcal{H}_2$ making $(\mathrm{Im}(S))^\perp$ trivial and giving $SS^\dagger = 1$ on $\mathrm{Im}(S)$. Furthermore, $\varphi_j = e_j$ is a normalized eigenbasis of $B = S^*S$ with eigenvalues $\sigma_j^2 = (\pi j)^{-2}$. Since $||Ae_j||_{L^2}^2 = \sigma_j^2$ we obtain for $g = Af \in \mathrm{Im}(A)$

$$\langle g, \psi_j \rangle_{L^2} = \langle Af, \frac{Ae_j}{||Ae_j||} \rangle_{L^2} = \sigma_j \langle f, e_j \rangle_{L^2} .$$

Thus, applying Proposition 4.3.1 to $g = Af$ we obtain

Corollary 4.3.2. *For $S = A$ and $B = S^*S$ as above we have for any $r > 0$:*

$f \in \mathrm{Im}(B^r)$ if and only if

$$\sum_{j=1}^{\infty} j^{4r} |\langle f, e_j \rangle_{L^2}|^2 < \infty \ .$$

Thus, as expected, abstract smoothness measured by the parameter r in the source condition corresponds in this special case to decay of the classical Fourier coefficients which - by the classical theory of Fourier

[1] i.e., the φ_j are the normalized eigenfunctions of S^*S with eigenvalues σ_j^2 and $\psi_j = S\varphi_j / ||S\varphi_j||$; thus $S = \sum \sigma_j \langle \varphi_j, \cdot \rangle \psi_j$

[2] the unique unbounded linear operator with domain $\mathrm{Im}(S) \oplus (\mathrm{Im}(S))^\perp$ in $\mathcal{H}_2$ vanishing on $(\mathrm{Im}(S))^\perp$ and satisfying $SS^\dagger = 1$ on $\mathrm{Im}(S)$, with range orthogonal to the null space $N(S)$

series - measures smoothness of the periodic continuation of $f \in L^2([0,1])$ to the real line.

Low smoothness

We choose $f_\rho(x) = x - \frac{1}{2}$ which clearly belongs to $\mathcal{H}_1$. A straightforward calculation gives the Fourier coefficient $\langle f_\rho, e_j \rangle = -2(\pi j)^{-2}$ for j odd (vanishing for j even). Thus, by the above criterion, f_ρ satisfies the source condition $f_\rho \in Ran(B^r)$ precisely for $0 < r < 0.75$. According to Theorem 2.3.4, the worst case rate in the single machine problem is given by $n^{-\gamma}$, with $\gamma = 0.25$. Regularization is done using the $\nu-$ method (see Example 2.2.21), with qualification $q = \nu = 1$. Recall that the stopping index T serves as the regularization parameter λ, where $T \sim \lambda^{-2}$. We consider sample sizes from the set

$$N := \{500, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 6000, 7000, 8000, 9000\} .$$

In the model selection step, we estimate the performance of different models and choose the *oracle stopping time* $\hat{T}_{oracle}$ by minimizing (an approximation of) the reconstruction error:

$$\hat{T}_{oracle} = \arg\min_T \left(\frac{1}{M} \sum_{j=1}^{M} \left\| f_\rho - \hat{f}_j^T \right\|_{\mathcal{H}_1}^2 \right)^{\frac{1}{2}}$$

over $M = 30$ runs (for any $n \in N$).

In the model assessment step, we partition the dataset into $m \sim n^\alpha$ subsamples, for any $\alpha \in \{0, 0.05, 0.1, ..., 0.85\}$. On each subsample we regularize using the oracle stopping time $\hat{T}_{oracle}$ (determined by using the whole sample). Corresponding to Corollary 4.2.3, the accuracy should be comparable to the one using the whole sample as long as $\alpha < 0.5$. In Figure 4.1 (left panel) we plot the reconstruction error $||\bar{f}^{\hat{T}} - f_\rho||_{\mathcal{H}_1}$ versus the ratio $\alpha = \log(m)/\log(n)$ for different sample sizes $n \in N$. We execute each simulation $M = 30$ times. The plot supports our theoretical finding. The right panel shows the reconstruction error versus the total number of samples $n \in N$ using different partitions of the data. The black curve ($\alpha = 0$) corresponds to the baseline error ($m = 0$, no partition of data). Error curves below a threshold $\alpha < 0.6$ are roughly comparable, whereas curves above this threshold show a gap in performances. However, while this figure correctly shows an error increasing with α, our data seem to be not sufficiently pushed into the asymptotic domain to correctly predict the rate (given by the slope of the various approximate lines): Calculating the slope from our data by the R-tool for linear regression gives Table 4.1.

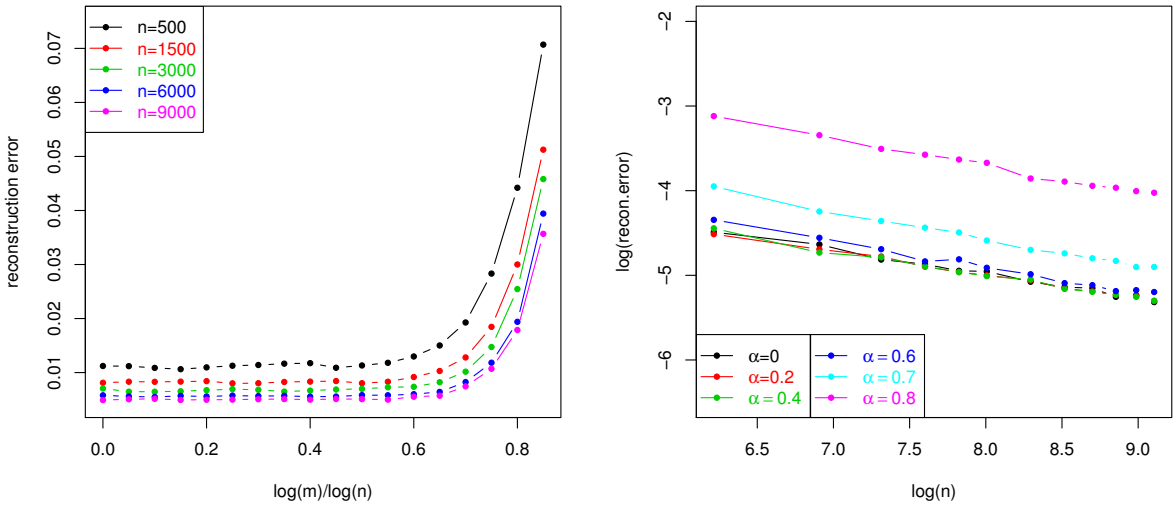

Figure 4.1: The reconstruction error $||\bar{f}_D^{T_{oracle}} - f_\rho||_{\mathcal{H}_1}$ in the low smoothness case. Left plot: Reconstruction error curves for various (but fixed) sample sizes as a function of the number of partitions. Right plot: Reconstruction error curves for various (but fixed) numbers of partitions as a function of the sample size (on log-scale).

Table 4.1: Estimated rates in low smoothness regime for different numbers of partitions. Theoretical value is $\gamma = 0.25$ up to threshold $\alpha = 0.5$.

α	$\hat{\gamma}$ ($\gamma = 0.25$)	97.5% confidence interval
0	0.284	[0.265 , 0.303]
0.05	0.275	[0.257 , 0.293]
0.1	0.272	[0.247 , 0.298]
0.15	0.265	[0.250 , 0.279]
0.2	0.273	[0.261 , 0.285]
0.25	0.273	[0.258 , 0.290]
0.3	0.274	[0.262 , 0.286]
0.35	0.272	[0.248 , 0.296]
0.4	0.283	[0.264 , 0.302]
0.45	0.280	[0.261 , 0.299]
0.5	0.266	[0.252 , 0.280]
0.55	0.281	[0.262 , 0.300]
0.6	0.303	[0.281 , 0.326]
0.65	0.328	[0.314 , 0.342]
0.7	0.326	[0.308 , 0.344]
0.75	0.324	[0.300 , 0.348]
0.8	0.320	[0.299 , 0.341]

Clearly, the increasing values of $\hat{\gamma}$ for $0.5 \leq \alpha \leq 0.8$ seem to be spurious, since our data have not yet sufficiently reached the asymptotic domain. Thus the rate of convergence is not reliably predicted by our data.

In another experiment we study the performances in case of (very) different regularization: only partitioning the data (no regularization), underregularization (higher stopping index) and overregularization (lower stopping index). The outcome of this experiment amplifies the regularization effect of parallelizing. Figure 4.2 shows the main point: Overregularization is always hopeless, underregularization is better. In the extreme case of none or almost none regularization there is a sharp minimum in the reconstruction error which is only slightly larger than the minimax optimal value for the oracle regularization parameter and which is achieved at an attractively large degree of parallelization. Qualitatively, this agrees very well with the intuitive notion that parallelizing serves as additional regularization. It is unclear if and how this effect could be systematically used as a computational tool (possibly trading time saving parallelization and avoiding the estimation of the correct regularization parameter against a loss in convergence).

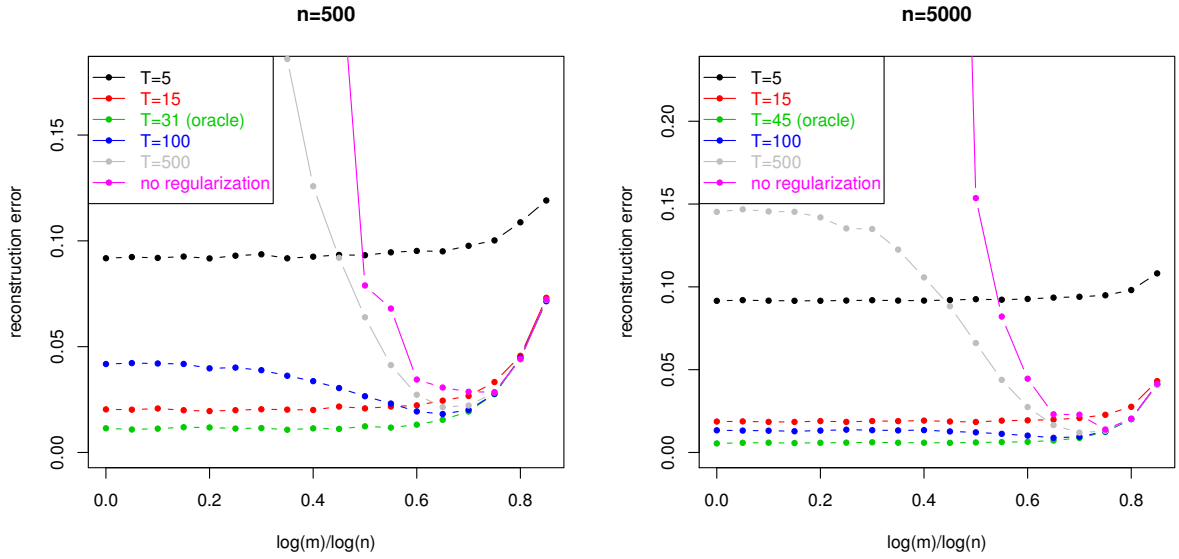

Figure 4.2: The reconstruction error $||\bar{f}_D^{\lambda} - f_\rho||_{\mathcal{H}_1}$ in the low smoothness case. Left plot: Error curves for different stopping times for $n = 500$ samples, as a function of the number of partitions. Right plot: Error curves for different stopping times for $n = 5000$ samples, as a function of the number of partitions.

We emphasize that numerical results seem to indicate that parallelization is possible to a slightly larger degree than indicated by our theoretical estimate. A similar result was reported in the paper [95], which also treats the low smoothness case. This point is not yet completely understood.

High smoothness

We choose $f_\rho(x) = \cos(2\pi x)$, which corresponds to just one non-vanishing Fourier coefficient and by our criterion Corollary 4.3.2 has $r = \infty$. In view of our main Corollary 4.2.3 this requires a regularization method with higher qualification; we take the *Gradient Descent* method (see Example 2.2.20).

The appearance of the term $2b\min\{1, r\}$ in our theoretical result 4.2.3 gives a predicted value $\alpha = 0$ and implies that parallelization is strictly forbidden for infinite smoothness. This is clearly verified by our computations which strongly support that the appearance of the above $2b\min\{1, r\}$ in the formula for the maximal α is not just an artefact of technically suboptimal estimates.

More specifically, the left panel in Figure 4.3 shows the absence of any plateau for the reconstruction error as a function of α. This corresponds to the right panel showing that no group of values of α performs roughly equivalently.

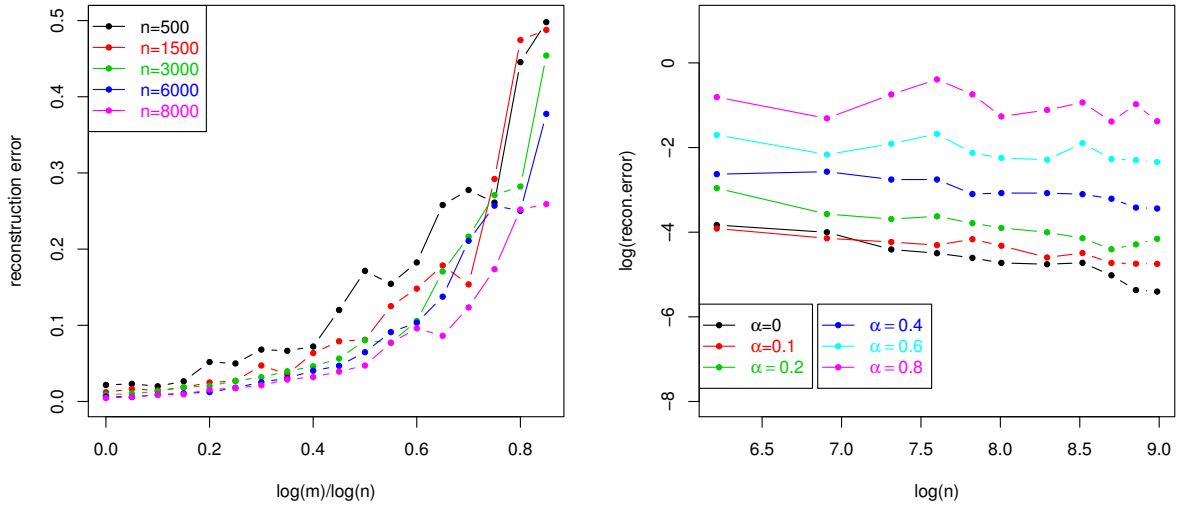

Figure 4.3: The reconstruction error $||\bar{f}_D^{\lambda_{oracle}} - f_\rho||_{\mathcal{H}_1}$ in the high smoothness case. Left plot: Reconstruction error curves for various (but fixed) sample sizes as a function of the number of partitions. Right plot: Reconstruction error curves for various (but fixed) numbers of partitions as a function of the sample size (on log-scale).

Plotting different values of regularization in Figure 4.4 we again identify overregularization as hopeless, while severe underregularization exhibits a sharp minimum in the reconstruction error. But its value at roughly 0.25 is much less attractive compared to the case of low smoothness where the error is an order of magnitude less.

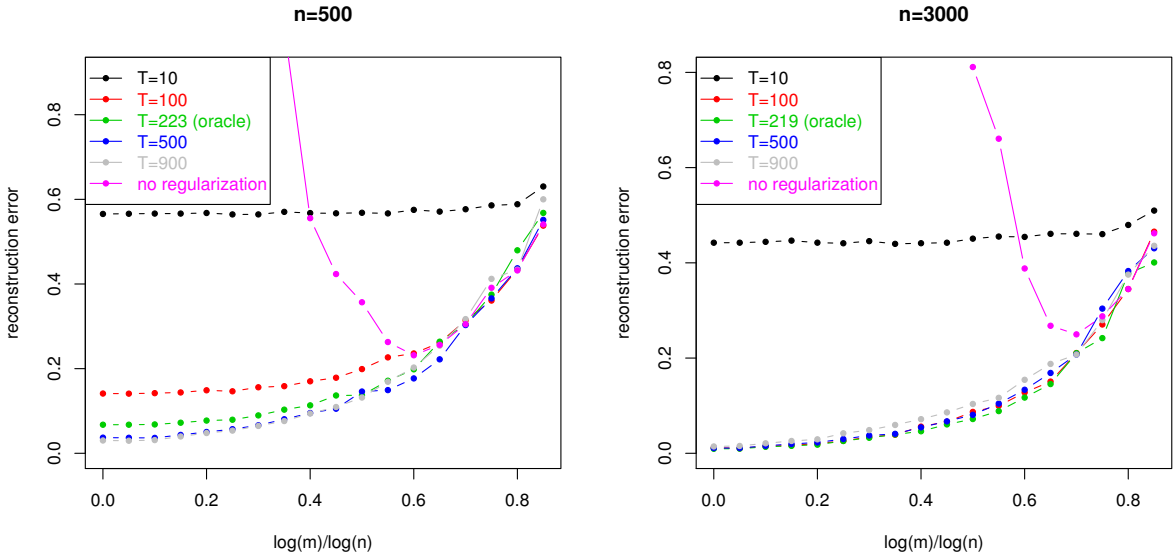

Figure 4.4: The reconstruction error $||\bar{f}_D^{\lambda} - f_\rho||$ in the high smoothness case. Left plot: Error curves for different stopping times for $n = 500$ samples, as a function of the number of partitions. Right plot: Error curves for different stopping times for $n = 5000$ samples, as a function of the number of partitions.

Finally, Table 4.2 shows that, as expected, rates of convergence are really bad for $\alpha \neq 0$ (especially so, considering the confidence intervall).

Table 4.2: Estimated rates in high smoothness regime for different numbers of partitions. The theoretical value is $\gamma = 0.5$ up to threshold $\alpha = 0$.

α	$\hat{\gamma}$ ($\gamma = 0.5$)	97.5% confidence interval
0	0.544	[0.428 , 0.660]
0.05	0.545	[0.463 , 0.628]
0.1	0.307	[0.227 , 0.387]
0.15	0.422	[0.281 , 0.563]
0.2	0.448	[0.346 , 0.550]
0.25	0.364	[0.302 , 0.426]
0.3	0.380	[0.295 , 0.465]
0.35	0.246	[0.136 , 0.356]
0.4	0.319	[0.225 , 0.413]
0.45	0.350	[0.272 , 0.428]
0.5	0.374	[0.250 , 0.497]
0.55	0.179	[0.013 , 0.346]
0.6	0.186	[0.026 , 0.347]
0.65	0.327	[0.121 , 0.533]
0.7	0.179	[0.026 , 0.331]
0.75	0.111	[-0.053 , 0.275]
0.8	0.145	[-0.109 , 0.399]

4.4 Discussion

Minimax Optimality: We have shown that for a large class of spectral regularization methods the error of the distributed algorithm $||\bar{B}^s(\bar{f}_D^{\lambda_n} - f_\rho)||_{\mathcal{H}_1}$ satisfies the same upper bound as the error $\left\|\bar{B}^s(f_D^{\lambda_n} - f_\rho)\right\|_{\mathcal{H}_1}$ for the single machine problem, if the regularization parameter λ_n is chosen according to (4.2.3), provided the number of subsamples grows sufficiently slowly with the sample size n. Since, by [17], the rates for the latter are minimax optimal, our rates in Corollary 4.2.3 are minimax optimal also.

Comparison with previous results [58] and [95]: In [95] the authors derive Minimax-optimal rates in 3 cases: finite rank kernels, sub- Gaussian decay of eigenvalues of the kernel and polynomial decay, provided m satisfies a certain upper bound, depending on the rate of decay of the eigenvalues under two crucial assumptions on the eigenfunctions of the integral operator associated to the kernel: For any $j \in \mathbb{N}$

$$\mathbb{E}[\phi_j(X)^{2k}] \leq \rho_k^{2k} \,, \tag{4.4.1}$$

for some $k \geq 2$ and $\rho_k < \infty$ or even stronger, it is assumed that the eigenfunctions are uniformly bounded, i.e.

$$\sup_{x \in \mathcal{X}} |\phi_j(x)| \leq \rho \,, \tag{4.4.2}$$

or any $j \in \mathbb{N}$ and some $\rho < \infty$. We shall describe in more detail the case of polynomially decaying eigenvalues, which corresponds to our setting. Assuming eigenvalue decay $\mu_j \lesssim j^{-b}$ with $b > 1$, the authors choose a regularization parameter $\lambda_n = n^{-\frac{b}{b+1}}$ and

$$m \lesssim \left(\frac{n^{\frac{b(k-4)-k}{b+1}}}{\rho^{4k} \log^k(n)} \right)^{\frac{1}{k-2}} .$$

leading to an error in L^2- norm

$$\mathbb{E}[||\bar{f}_D^{\lambda_n} - f_\rho||_{L^2}^2] \lesssim n^{-\frac{b}{b+1}} \,,$$

being minimax optimal.

For $k < 4$, this is not a useful bound, since $m \to 1$ as $n \to \infty$ in this case (for any sort of eigenvalue decay). On the other hand, if k and b might be taken arbitrarily large - corresponding to almost bounded eigenfunctions and arbitrarily large polynomial decay of eigenvalues - m might be chosen proportional to $n^{1-\epsilon}$, for any $\epsilon > 0$. As might be expected, replacing the L^{2k} bound on the eigenfunctions by a bound in L^∞, gives an upper bound on m which simply is the limit for $k \to \infty$ in the bound given above, namely

$$m \lesssim \frac{n^{\frac{b-1}{b+1}}}{\rho^4 \log n} \,,$$

which for large b behaves as above. Granted bounds on the eigenfunctions in L^{2k} for (very) large k, this is a strong result. While the decay rate of the eigenvalues can be determined by the smoothness of K (see, e.g., [36] and references therein), it is a widely open question which general properties of the kernel imply estimates as in (4.4.1) and (4.4.2) on the eigenfunctions.

The author in [96] even gives a counterexample and presents a C^∞ Mercer kernel on $[0, 1]$ where the eigenfunctions of the corresponding integral operator are *not* uniformly bounded. Thus, smoothness of

the kernel is not a sufficient condition for (4.4.2) to hold.

Moreover, we point out that the upper bound (4.4.1) on the eigenfunctions (and thus the upper bound for m in [95]) depends on the (unknown) marginal distribution ν (only the strongest assumption, a bound in sup-norm (4.4.2), does not depend on ν). Concerning this point, our approach is "agnostic".

As already mentioned in the Introduction, these bounds on the eigenfunctions have been eliminated in [58], for KRR, imposing polynomial decay of eigenvalues as above. This is very similar to our approach. As a general rule, our bounds on m and the bounds in [58] are worse than the bounds in [95] for eigenfunctions in (or close to) L^∞, but in the complementary case where nothing is known on the eigenfunctions m still can be chosen as an increasing function of n, namely $m = n^\alpha$. More precisely, choosing λ_n as in (4.2.3), the authors in [58] derive as an upper bound

$$m \lesssim n^\alpha \ , \qquad \alpha = \frac{2br}{2br + b + 1} \ ,$$

with r being the smoothness parameter arising in the source condition. We recall here that due to our assumption $q \geq r + s$, the smoothness parameter r is restricted to the interval $(0, \frac{1}{2}]$ for KRR ($q = 1$) and L^2 risk ($s = \frac{1}{2}$).

Our results (which hold for a general class of spectral regularization methods) are in some ways comparable to [58]. Specialized to KRR, our estimates for the exponent α in $m = O(n^\alpha)$ coincide with the result given in [58] . Furthermore we emphasize that [95] and [58] estimate the DL-error only for $s = 1/2$ in our notation (corresponding to $L^2(\nu)-$ norm), while our result holds for all values of $s \in [0, 1/2]$ which smoothly interpolates between $L^2(\nu)-$ norm and RKHS$-$ norm and, in addition, for all values of $p \in [1, \infty)$. Additionally, we precisely analyze the dependence of the noise variance σ^2 and the complexity radius R in the source condition.[3]

Concerning general strategy, while [58] uses a novel second order decomposition in an essential way, our approach is more classical. We clearly distinguish between estimating the approximation error and the sample error. We write the variance as a sum of i.i.d random variables, which allows to use Rosenthal's inequality. Compared to our previous result for the single machine problem in Chapter 2 , this is an essentially new ingredient in our proof.

Number of Subsamples: We follow the line of reasoning in earlier work on distributed learning insofar as we only prove *sufficient conditions* on the cardinality $m = n^\alpha$ of subsamples compatible with minimax optimal rates of convergence. On the complementary problem of proving *necessity*, analytical results are unknown to the best of our knowledge. However, our numerical results seem to indicate that the exponent α might actually be taken larger than we have proved so far in the low smoothness regime.

Adaptivity: It is clear from the theoretical results that both the regularization parameter λ and the allowed cardinality of subsamples m depend on the parameters r and b, which in general are unknown. Thus, an adaptive approach to both parameters b and r for choosing λ and m is of interest. In [95] the authors sketch a heuristic argument to estimate λ from the data of the subsamples via cross-validation, *provided the number of subsamples has been fixed by an a priori choice.* This leaves completely open the important issue how an optimal m could adaptively be inferred from the given data. To the best of our

[3] While this thesis was written, the authors in [42] concurrently worked at distributed learning algorithms for the same class of spectral regularization schemes. They establish the same upper bound for the number of allowed subsamples in $L^2(\nu)-$ norm, but with unpublished proof at the time of submission of this thesis.

knowledge, there are yet no rigorous results on adaptivity in this more general sense. Progress in this field may well be crucial in finally assessing the relative merits of the distributed learning approach as compared with alternative strategies to effectively deal with large data sets.

We sketch an alternative naive approach to adaptivity, based on hold-out in the direct case, where we consider each $f \in \mathcal{H}_K$ also as a function in $L^2(\mathcal{X}, \nu)$. We split the data $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n$ into a training and validation part $\mathbf{z} = (\mathbf{z}^t, \mathbf{z}^v)$ of cardinality m_t, m_v. We further subdivide $\mathbf{z}^t$ into m_k subsamples, roughly of size m_t/m_k, where $m_k \leq m_t$, $k = 1, 2, \ldots$ is some strictly decreasing sequence. For each k and each subsample $\mathbf{z}_j$, $1 \leq j \leq m_k$, we define the estimators $\hat{f}^\lambda_{\mathbf{z}_j}$ as in (4.1.1) and their average

$$\bar{f}^\lambda_{k,\mathbf{z}^t} := \frac{1}{m_k} \sum_{j=1}^{m_k} \hat{f}^\lambda_{\mathbf{z}_j} \ . \tag{4.4.3}$$

Here, λ varies in some sufficiently fine lattice Λ. Then evaluation on $\mathbf{z}^v$ gives the associated empirical L^2- error

$$\mathrm{Err}^\lambda_k(\mathbf{z}^v) := \frac{1}{m_v} \sum_{i=1}^{m_v} |y^v_i - \bar{f}^\lambda_{k,\mathbf{z}^t}(x^v_i)|^2 \ , \quad \mathbf{z}^v = (\mathbf{y}^v, \mathbf{x}^v) \ , \quad \mathbf{y}^v = (y^v_1, \ldots, y^v_{m_v}) \ , \tag{4.4.4}$$

leading us to define

$$\hat{\lambda}_k := \mathrm{Argmin}_{\lambda \in \Lambda} \mathrm{Err}^\lambda_k(\mathbf{z}^v) \ , \qquad \mathrm{Err}(k) := \mathrm{Err}^{\hat{\lambda}_k}_k(\mathbf{z}^v). \tag{4.4.5}$$

Then, an appropriate stopping criterion for k might be to stop at

$$k^* := \min\{k \geq 3 \ : \ \Delta(k) \leq \delta \inf_{2 \leq j < k} \Delta(j)\} \ , \qquad \Delta(j) := |\mathrm{Err}(j) - \mathrm{Err}(j-1)| \ , \tag{4.4.6}$$

for some $\delta < 1$ (which might require tuning). The corresponding regularization parameter is $\hat{\lambda} = \hat{\lambda}_{k^*}$, given by (4.4.5). At least intuitively, it is then reasonable to define a purely data driven estimator as

$$\widehat{f}_n := \bar{f}^{\hat{\lambda}}_{k^*,\mathbf{z}^t} \ . \tag{4.4.7}$$

Note that the training data $\mathbf{z}^t$ enter the definition of $\widehat{f}_n$ via the explicit formula (4.4.3) encoding our kernel based approach, while $\mathbf{z}^v$ serves to determine $(k^*, \hat{\lambda}^*)$ via minimization of the empirical L^2- error and some form of the discrepancy principle, which tells one to stop where $\mathrm{Err}(j)$ does not appreciably improve anymore. It is open if such a procedure achieves optimal rates, and we have to leave this for future research.

4.5 Proofs

For ease of reading we make use of the following conventions:

- we are interested in a precise dependence of multiplicative constants on the parameters σ, M, R, η, m, n and p
- the dependence of multiplicative constants on various other parameters, including the kernel parameter κ, the norm parameter $s \in [0, \frac{1}{2}]$, the parameters arising from the regularization method,

$b > 1$, $\beta > 0$, $r > 0$ etc. will (generally) be omitted and simply be indicated by the symbol ▲

- the value of $C_\blacktriangle$ might change from line to line
- the expression "for n sufficiently large" means that the statement holds for $n \geq n_0$, with n_0 potentially depending on all model parameters (including σ, M and R), but not on η.

We intend to estimate each term in the decomposition (4.2.2) separately. This is the content of the following Propositions. Recall that ν denotes the input sampling distribution and $\mathcal{P}$ the set of all probability distributions on the input space $\mathcal{X}$.

At first, a preliminary Lemma:

Lemma 4.5.1. *Recall the definition of $\mathcal{B}_{\frac{n}{m}}(\lambda)$ in (A.2.1), where $\lambda \in (0,1]$:*

$$\mathcal{B}_{\frac{n}{m}}(\lambda) := \left[1 + \left(\frac{2m}{n\lambda} + \sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right)^2\right] . \tag{4.5.1}$$

If λ_n is defined by (4.2.3) and if

$$m_n \leq n^\alpha , \qquad \alpha < \frac{2br}{2br+b+1} ,$$

one has

$$\mathcal{B}_{\frac{n}{m_n}}(\lambda_n) \leq 2 ,$$

provided n is sufficiently large.

Proof of Lemma 4.5.1. Recall that $\mathcal{N}(\lambda_n) \leq C_b \lambda_n^{-\frac{1}{b}}$ and $\sigma\sqrt{\frac{\lambda_n^{-\frac{1}{b}}}{n\lambda_n}} = R\lambda_n^r$. Using the definition of λ_n in (4.2.3) yields

$$\frac{2m_n}{n\lambda} = o\left(\sqrt{m}_n \lambda_n^r\right) ,$$

provided

$$m_n \leq n^\alpha , \qquad \alpha < \frac{2(br+1)}{2br+b+1} .$$

Finally, $\sqrt{m}_n \lambda_n^r = o(1)$ if

$$m_n \leq n^\alpha , \qquad \alpha < \frac{2br}{2br+b+1} .$$

□

Approximation Error

Lemma 4.5.2. *Let $\nu \in \mathcal{P}$, $v \in \mathbb{R}$ and let $\mathbf{x} \in \mathcal{X}^{\frac{n}{m}}$ be an iid sample, drawn according to ν. Assume the regularization $(g_\lambda)_\lambda$ has qualification $q \geq v + 1 + s$. Then with probability at least $1 - \eta$*

$$\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}}) \bar{B}_{\mathbf{x}}^v (\bar{B} - \bar{B}_{\mathbf{x}})\right\|_{\mathcal{H}_1} \leq C_\blacktriangle \log^4(4\eta^{-1}) \lambda^{s+v+1} \mathcal{B}_{\frac{n}{m}}^{s+1}(\lambda) \left(\frac{2m}{n\lambda} + \sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right)$$

for some $C_\blacktriangle < \infty$.

Proof of Lemma 4.5.2. From (A.2.3) and from Proposition A.1.3, since $q \geq s+v+1$, one has

$$\begin{aligned}\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}_{\mathbf{x}}^v(\bar{B}-\bar{B}_{\mathbf{x}})\right\|_{\mathcal{H}_1} &\leq C_\blacktriangle \log^{2(s+1)}(4\eta^{-1})\mathcal{B}_{\frac{n}{m}}^{s+1}(\lambda) \\ &\qquad \left\|(\bar{B}_{\mathbf{x}}+\lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}_{\mathbf{x}}^v(\bar{B}_{\mathbf{x}}+\lambda)\right\| \left\|(\bar{B}+\lambda)^{-1}(\bar{B}-\bar{B}_{\mathbf{x}})\right\| \\ &\leq C_\blacktriangle \log^4(4\eta^{-1})\lambda^{s+v+1}\mathcal{B}_{\frac{n}{m}}^{s+1}(\lambda)\left(\frac{2m}{n\lambda}+\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right) ,\end{aligned}$$

for any $\lambda \in (0,1]$, $\eta \in (0,1]$, with probability at least $1-\eta$. We also used that $s \leq \frac{1}{2}$. □

Lemma 4.5.3. *Let $\nu \in \mathcal{P}$, $v \in \mathbb{R}$ and let $\mathbf{x} \in \mathcal{X}^{\frac{n}{m}}$ be an iid sample, drawn according to ν. Assume the regularization $(g_\lambda)_\lambda$ has qualification $q \geq v+s$. Then for any $\lambda \in (0,1]$, $\eta \in (0,1]$, with probability at least $1-\eta$*

$$\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}_{\mathbf{x}}^v\right\| \leq C_\blacktriangle \log^{2s}(2\eta^{-1})\mathcal{B}_{\frac{n}{m}}^{s}(\lambda)\lambda^{s+v} ,$$

for some $C_\blacktriangle < \infty$.

Proof of Lemma 4.5.3. Using (A.2.3), since $q \geq v+s$

$$\begin{aligned}\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}_{\mathbf{x}}^v\right\| &\leq C_\blacktriangle \log^{2s}(2\eta^{-1})\mathcal{B}_{\frac{n}{m}}^{s}(\lambda)\left\|(\bar{B}_{\mathbf{x}}+\lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}_{\mathbf{x}}^v\right\| \\ &\leq C_\blacktriangle \log^{2s}(2\eta^{-1})\mathcal{B}_{\frac{n}{m}}^{s}(\lambda)\lambda^{s+v} ,\end{aligned}$$

with probability at least $1-\eta$. □

Proposition 4.5.4 (Expectation of Approximation Error)**.** *Let $f_\rho \in \Omega_\nu(r,R)$, $\lambda \in (0,1]$ and let $\mathcal{B}_{\frac{n}{m}}(\lambda)$ be defined in* (4.5.1)*. Assume the regularization has qualification $q \geq r+s$. For any $p \geq 1$ one has:*

1. *If $r \leq 1$, then*
$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho - \tilde{f}_D^\lambda)\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} \leq C_p R\, \lambda^{s+r}\, \mathcal{B}_{\frac{n}{m}}^{s+r}(\lambda) .$$
2. *If $r > 1$, then*
$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho - \tilde{f}_D^\lambda)\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} \leq C_p R\lambda^s \mathcal{B}_{\frac{n}{m}}^{s+1}(\lambda)\left(\lambda^r + \lambda\left(\frac{2m}{n\lambda}+\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right)\right) .$$

In 1. and 2. the constant C_p does not depend on $(\sigma, M, R) \in \mathbb{R}_+^3$.

Proof of Proposition 4.5.4. Since $f_\rho \in \Omega_\nu(r,R)$

$$\begin{aligned}\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho - \tilde{f}_D^\lambda)\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} &= \mathbb{E}_{\rho^{\otimes n}}\left[\left\|\frac{1}{m}\sum_{j=1}^m \bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})f_\rho\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} \\ &\leq \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})f_\rho\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} \\ &\leq \frac{R}{m}\sum_{j=1}^m \mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^r\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} . \qquad (4.5.2)\end{aligned}$$

The first inequality is just the triangle inequality for the p- norm $||f||_p = \mathbb{E}[||f||^p_{\mathcal{H}_1}]^{\frac{1}{p}}$. We bound the expectation for each separate subsample of size $\frac{n}{m}$ by first deriving a probabilistic estimate and then we integrate.

Consider first the case where $r \leq 1$. Using (A.2.3) and Cordes Inequality Proposition A.4.2 , one has for any $j = 1, ..., m$

$$\begin{aligned}\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^r\right\| &\leq C_\blacktriangle \log^{2(s+r)}(4\eta^{-1})\mathcal{B}^{s+r}_{\frac{n}{m}}(\lambda)\left\|(\bar{B}_{\mathbf{x}_j}+\lambda)^s r_\lambda(\bar{B}_{\mathbf{x}_j})(\bar{B}_{\mathbf{x}_j}+\lambda)^r\right\| \\ &\leq C_\blacktriangle \log^3(4\eta^{-1})\lambda^{s+r}\mathcal{B}^{s+r}_{\frac{n}{m}}(\lambda)\ ,\end{aligned}$$

with probability at least $1-\eta$ and where $\mathcal{B}^{s+r}_{\frac{n}{m}}(\lambda)$ is defined in (4.5.1). Recall that the regularization has qualification $q \geq r+s$. From Lemma A.3.3, by integration one has

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^r\right\|^p\right]^{\frac{1}{p}} \leq C_{\blacktriangle,p}\ \lambda^{s+r}\ \mathcal{B}^{s+r}_{\frac{n}{m}}(\lambda)\ ,$$

for some $C_{\blacktriangle,p} < \infty$, not depending on σ, M, R. Finally, from (4.5.2)

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho - \tilde{f}^\lambda_D)\right\|^p_{\mathcal{H}_1}\right]^{\frac{1}{p}} \leq C_{\blacktriangle,p} R\ \lambda^{s+r}\ \mathcal{B}^{s+r}_{\frac{n}{m}}(\lambda)\ .$$

In the case where $r \geq 1$, we write $r = k+u$, with $k = \lfloor r \rfloor$ and $u = r-k < 1$. We shall use the decomposition

$$\bar{B}^k = \sum_{l=0}^{k-1} \bar{B}^l_{\mathbf{x}}(\bar{B}-\bar{B}_{\mathbf{x}})\bar{B}^{k-(l+1)} + \bar{B}^k_{\mathbf{x}}\ . \tag{4.5.3}$$

We proceed by bounding (4.5.2) according to decomposition (4.5.3) . For any $j = 1, ...m$, one has

$$\begin{aligned}\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^{k+u}\right\|^p\right]^{\frac{1}{p}} &\leq \sum_{l=0}^{k-1}\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^l_{\mathbf{x}_j}(\bar{B}-\bar{B}_{\mathbf{x}_j})\bar{B}^{k-(l+1)+u}\right\|^p\right]^{\frac{1}{p}} \\ &\quad + \mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^k_{\mathbf{x}_j}\bar{B}^u\right\|^p\right]^{\frac{1}{p}} \\ &\leq \sum_{l=0}^{k-1}\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^l_{\mathbf{x}_j}(\bar{B}-\bar{B}_{\mathbf{x}_j})\right\|^p\right]^{\frac{1}{p}} \\ &\quad + \mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^k_{\mathbf{x}_j}\bar{B}^u\right\|^p\right]^{\frac{1}{p}}\ .\end{aligned} \tag{4.5.4}$$

Here we use that $\bar{B}^{k-(l+1)+u}$ is bounded by 1. By Lemma 4.5.3 and by (A.2.3), with probability at least $1-\eta$

$$\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^k_{\mathbf{x}_j}\bar{B}^u\right\| \leq C_\blacktriangle \log^{2(s+u)}(2\eta^{-1})\mathcal{B}^{s+u}_{\frac{n}{m}}(\lambda)\lambda^{s+r}$$

and thus integration using Lemma A.3.3 yields

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^r_{\mathbf{x}_j}\right\|^p\right]^{\frac{1}{p}} \leq C_{\blacktriangle,p}\mathcal{B}^{s+u}_{\frac{n}{m}}(\lambda)\lambda^{s+r}\ . \tag{4.5.5}$$

For estimating the first term in (4.5.4) we may use Lemma 4.5.2. For any $l = 0, ..., k-1$, $j = 1, ..., m$

with probability at least $1-\eta$

$$\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^l_{\mathbf{x}_j}(\bar{B}-\bar{B}_{\mathbf{x}_j})\right\| \leq C_{\blacktriangle}\log^4(8\eta^{-1})\lambda^{s+l+1}\mathcal{B}^{s+1}_{\frac{n}{m}}(\lambda)\left(\frac{2m}{n\lambda}+\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right) .$$

Again by integration, since $\lambda^l \leq 1$ for any $l=0,...,k-1$, one has

$$\sum_{l=0}^{k-1}\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s r_\lambda(\bar{B}_{\mathbf{x}_j})\bar{B}^l_{\mathbf{x}_j}(\bar{B}-\bar{B}_{\mathbf{x}_j})\right\|^p\right]^{\frac{1}{p}} \leq C_{\blacktriangle,p}\lfloor r\rfloor\lambda^{s+1}\mathcal{B}^{s+1}_{\frac{n}{m}}(\lambda)\left(\frac{2m}{n\lambda}+\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right) . \tag{4.5.6}$$

Finally, combining (4.5.5) and (4.5.6) with (4.5.2) gives in the case where $r>1$

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho-\tilde{f}^\lambda_D)\right\|^p_{\mathcal{H}_1}\right]^{\frac{1}{p}} \leq C_{\blacktriangle}\lambda^s\mathcal{B}^{s+1}_{\frac{n}{m}}(\lambda)\left(\lambda^r+\lambda\left(\frac{2m}{n\lambda}+\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right)\right) .$$

The rest of the proof follows from (4.5.4).

□

Proof of Theorem 4.2.1. Let λ_n defined by (4.2.3). According to Lemma 4.5.1, we have $\mathcal{B}_{\frac{n}{m_n}}(\lambda_n)\leq 2$ provided $\alpha < \frac{2br}{2br+b+1}$. We immediately obtain from the first part of Proposition 4.5.4 in the case where $r\leq 1$

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho-\tilde{f}^{\lambda_n}_D)\right\|^p_{\mathcal{H}_1}\right]^{\frac{1}{p}} \leq C_{\blacktriangle,p}R\,\lambda_n^{s+r} = C_{\blacktriangle,p}\,a_n .$$

We turn to the case where $r>1$. We apply the second part of Proposition 4.5.4. By Lemma 4.5.1 we have

$$\begin{aligned}\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho-\tilde{f}^{\lambda_n}_D)\right\|^p_{\mathcal{H}_1}\right]^{\frac{1}{p}} &\leq C_{\blacktriangle,p}\,R\lambda_n^{s+1}\left(\frac{2m_n}{n\lambda_n}+\sqrt{\frac{m_n\mathcal{N}(\lambda_n)}{n\lambda_n}}\right)\\ &\leq C_{\blacktriangle,p}\,R\lambda_n^{s+1}\left(\frac{2m_n}{n\lambda_n}+\frac{R}{\sigma}\sqrt{m_n}\lambda_n^r\right) ,\end{aligned}$$

where we used that $\mathcal{N}(\lambda_n)\leq C_b\lambda_n^{-1/b}$. Observe that

$$\frac{2m_n}{n\lambda_n} = o\left(\sqrt{m}_n\lambda_n^r\right) ,$$

provided

$$m_n \leq n^\alpha , \qquad \alpha < \frac{2(br+1)}{2br+b+1} .$$

Furthermore, for n sufficiently large, $\frac{R}{\sigma}\sqrt{m_n}\lambda_n\leq 1$, provided that

$$\alpha < \frac{2b}{2br+b+1} .$$

As a result, for any $1\leq p$

$$\limsup_{n\to\infty}\sup_{\rho\in\mathcal{M}_{\sigma,M,R}}\frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(f_\rho-\tilde{f}^{\lambda_n}_D)\right\|^p_{\mathcal{H}_1}\right]^{\frac{1}{p}}}{a_n} \leq C_{\blacktriangle,p} ,$$

for some $C_{\blacktriangle,p} < \infty$, not depending on σ, M, R. □

Sample Error

The main idea for deriving an upper bound for the sample error is to identify it as a sum of unbiased Hilbert space- valued i.i.d. variables and then to apply a suitable version of Rosenthal's inequality.

Given $\lambda \in (0,1]$, we define the random variable $\xi_\lambda : (\mathcal{X} \times \mathbb{R})^{\frac{n}{m}} \longrightarrow \mathcal{H}_1$ by

$$\xi_\lambda(\mathbf{x}, \mathbf{y}) := \bar{B}^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}^*_{\mathbf{x}} \mathbf{y}) \, .$$

Recall that according to Assumption 2.2.10, the conditional expectation w.r.t. ρ of Y given X satisfies

$$\mathbb{E}_\rho[Y|X = x] = \bar{S}_x f_\rho \, ,$$

implying that ξ_λ is unbiased (since $\bar{B}_{\mathbf{x}} = \bar{S}^*_{\mathbf{x}} \bar{S}_{\mathbf{x}}$). Thus,

$$\bar{B}^s(\tilde{f}^\lambda_D - \bar{f}^\lambda_D) \; = \; \frac{1}{m} \sum_{j=1}^m \xi_\lambda(\mathbf{x}_j, \mathbf{y}_j) \tag{4.5.7}$$

is a sum of centered i.i.d. random variables.

Furthermore, we need the following result from [70], Theorem 5.2 , which generalizes Rosenthal's inequalities from [76] (originally only formulated for real valued random variables) to random variables with values in a Banach space. For Hilbert spaces this looks particularly nice.

Proposition 4.5.5. *Let $\mathcal{H}$ be a Hilbert space and $\xi_1, ..., \xi_m$ be a finite sequence of independent, mean zero $\mathcal{H}$- valued random variables. If $2 \leq p < \infty$, then there exists a constant $C_p > 0$, only depending on p, such that*

$$\left(\mathbb{E} \left\| \frac{1}{m} \sum_{j=1}^m \xi_j \right\|^p_{\mathcal{H}} \right)^{\frac{1}{p}} \; \leq \; \frac{C_p}{m} \; \max \left\{ \left(\sum_{j=1}^m \mathbb{E}||\xi_j||^p_{\mathcal{H}} \right)^{\frac{1}{p}} , \left(\sum_{j=1}^m \mathbb{E}||\xi_j||^2_{\mathcal{H}} \right)^{\frac{1}{2}} \right\} \, . \tag{4.5.8}$$

We remark in passing that [32] , Corollary 1.22 , contains the interesting result that in addition to the upper bound in (4.5.8) there is also a corresponding lower bound where the constant C_p is replaced by another constant $C'_p > 0$, only depending on p.

Proposition 4.5.6 (Expectation of Sample Error)**.** *Let ρ be a source distribution belonging to $\mathcal{M}_{\sigma,M,R}$, $s \in [0, \frac{1}{2}]$ and let $\lambda \in (0,1]$. Define $\mathcal{B}_{\frac{n}{m}}(\lambda)$ as in* (4.5.1)*. Assume the regularization has qualification $q \geq r + s$. For any $p \geq 1$ one has:*

$$\mathbb{E}_{\rho^{\otimes n}} \left[\left\| \bar{B}^s(\tilde{f}^\lambda_D - \bar{f}^\lambda_D) \right\|^p_{\mathcal{H}_1} \right]^{\frac{1}{p}} \leq C_p \, m^{-\frac{1}{2}} \mathcal{B}_{\frac{n}{m}}(\lambda)^{\frac{1}{2}+s} \lambda^s \, \left(\frac{mM}{n\lambda} + \sigma \sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}} \right) \, ,$$

where C_p does not depend on $(\sigma, M, R) \in \mathbb{R}^3_+$.

Proof of Proposition 4.5.6. Let $\lambda \in (0,1]$ and $p \geq 2$. From Proposition 4.5.5

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s\left(\tilde{f}_D^\lambda - \bar{f}_D^\lambda\right)\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} = \mathbb{E}_{\rho^{\otimes n}}\left[\left\|\frac{1}{m}\sum_{j=1}^m \xi_\lambda(\mathbf{x}_j,\mathbf{y}_j)\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} \tag{4.5.9}$$

$$\leq \frac{C_p}{m}\ \max\left\{\left(\sum_{j=1}^m \mathbb{E}_{\rho^{\otimes n}}\left[||\xi_\lambda(\mathbf{x}_j,\mathbf{y}_j)||_{\mathcal{H}_1}^p\right]\right)^{\frac{1}{p}}, \left(\sum_{j=1}^m \mathbb{E}_{\rho^{\otimes n}}\left[||\xi_\lambda(\mathbf{x}_j,\mathbf{y}_j)||_{\mathcal{H}_1}^2\right]\right)^{\frac{1}{2}}\right\}\ .$$

Again, the estimates in expectation will follow from integration a bound holding with high probability. By (A.2.3), one has for any $j = 1, ..., m$

$$\begin{aligned} ||\xi_\lambda(\mathbf{x}_j,\mathbf{y}_j)||_{\mathcal{H}_1} &= ||\bar{B}^s g_\lambda(\bar{B}_{\mathbf{x}_j})(\bar{B}_{\mathbf{x}_j} f_\rho - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j)||_{\mathcal{H}_1} \\ &\leq 8\log^{2s}(4\eta^{-1})\mathcal{B}_{\frac{n}{m}}(\lambda)^s \\ &\quad ||(\bar{B}_{\mathbf{x}_j}+\lambda)^s g_\lambda(\bar{B}_{\mathbf{x}_j})(\bar{B}_{\mathbf{x}_j} f_\rho - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j)||_{\mathcal{H}_1}\ , \end{aligned} \tag{4.5.10}$$

holding with probability at least $1-\frac{\eta}{2}$, where $B_{\frac{n}{m}}(\lambda)$ is defined in (4.5.1). We proceed by further splitting as in (2.5.15) (with n replaced by $\frac{n}{m}$)

$$(\bar{B}_{\mathbf{x}_j}+\lambda)^s g_\lambda(\bar{B}_{\mathbf{x}_j})(\bar{B}_{\mathbf{x}_j} f_\rho - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j) = H^{(1)}_{\mathbf{x}_j}\cdot H^{(2)}_{\mathbf{x}_j}\cdot h^\lambda_{\mathbf{z}_j}\ ,$$

with

$$\begin{aligned} H^{(1)}_{\mathbf{x}_j} &:= (\bar{B}_{\mathbf{x}_j}+\lambda)^s g_\lambda(\bar{B}_{\mathbf{x}_j})(\bar{B}_{\mathbf{x}_j}+\lambda)^{\frac{1}{2}}, \\ H^{(2)}_{\mathbf{x}_j} &:= (\bar{B}_{\mathbf{x}_j}+\lambda)^{-\frac{1}{2}}(\bar{B}+\lambda)^{\frac{1}{2}}, \\ h^\lambda_{\mathbf{z}_j} &:= (\bar{B}+\lambda)^{-\frac{1}{2}}(\bar{B}_{\mathbf{x}_j} f_\rho - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j)\ . \end{aligned}$$

The first term is estimated as in (2.5.16) and gives

$$H^{(1)}_{\mathbf{x}_j} \leq C_\blacktriangle \lambda^{s-\frac{1}{2}}\ . \tag{4.5.11}$$

The second term is now bounded using (A.2.3) once more. One has with probability at least $1-\frac{\eta}{4}$

$$H^{(2)}_{\mathbf{x}_j} \leq 8\log(8\eta^{-1})\mathcal{B}_{\frac{n}{m}}(\lambda)^{\frac{1}{2}}\ . \tag{4.5.12}$$

Finally, $h^\lambda_{\mathbf{z}_j}$ is estimated using Proposition A.1.2:

$$h^\lambda_{\mathbf{z}_j} \leq 2\log(8\eta^{-1})\left(\frac{mM}{n\sqrt{\lambda}} + \sigma\sqrt{\frac{m\mathcal{N}(\lambda)}{n}}\right)\ , \tag{4.5.13}$$

holding with probability at least $1-\frac{\eta}{4}$. Thus, combining (4.5.11), (4.5.12) and (4.5.13) with (4.5.10) gives for any $j = 1, ..., m$

$$||\xi_\lambda(\mathbf{x}_j,\mathbf{y}_j)||_{\mathcal{H}_1} \leq C_\blacktriangle \log^{2(s+1)}(8\eta^{-1})\mathcal{B}_{\frac{n}{m}}(\lambda)^{\frac{1}{2}+s}\lambda^s\ \left(\frac{mM}{n\lambda} + \sigma\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right)\ ,$$

with probability at least $1-\eta$. Integration using Lemma A.3.3 gives for any $p \geq 2$

$$\sum_{j=1}^{p} \mathbb{E}_{\rho^{\otimes n}}\left[\left\|\xi_\lambda(\mathbf{x}_j, \mathbf{y}_j)\right\|_{\mathcal{H}_1}^p\right] \leq C_{\blacktriangle,p} m\, \mathcal{A}^p \,,$$

with

$$\mathcal{A} := \mathcal{A}_{\frac{n}{m}}(\lambda) := \mathcal{B}_{\frac{n}{m}}(\lambda)^{\frac{1}{2}+s}\lambda^s \left(\frac{mM}{n\lambda} + \sigma\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right) .$$

Combining this with (4.5.9) implies, since $p \geq 2$

$$\begin{aligned}
\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(\tilde{f}_D^\lambda - \bar{f}_D^\lambda)\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} &\leq \frac{C_p}{m}\max\left((m\mathcal{A}^p)^{\frac{1}{p}}, \left(m\mathcal{A}^2\right)^{\frac{1}{2}}\right) \\
&= \frac{C_p}{m}\,\mathcal{A}\max\left(m^{\frac{1}{p}}, m^{\frac{1}{2}}\right) \\
&= \frac{C_p}{\sqrt{m}}\,\mathcal{A}\,,
\end{aligned}$$

where C_p does not depend on $(\sigma, M, R) \in \mathbb{R}_+^3$. The result for the case $1 \leq p \leq 2$ immediately follows from Hölder's inequality. □

Proof of Theorem 4.2.2. Let λ_n defined by (4.2.3). According to Lemma 4.5.1 we have $\mathcal{B}_{\frac{n}{m}}(\lambda_n) \leq 2$ provided $\alpha < \frac{2br}{2br+b+1}$. We immediately obtain from Proposition 4.5.6

$$\begin{aligned}
\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(\tilde{f}_D^{\lambda_n} - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} &\leq \frac{C_p}{\sqrt{m}}\,\lambda_n^s \left(\frac{mM}{n\lambda_n} + \sigma\sqrt{\frac{m\mathcal{N}(\lambda_n)}{n\lambda_n}}\right) \\
&\leq C_p\lambda_n^s \left(\frac{\sqrt{m}M}{n\lambda_n} + \sigma\sqrt{\frac{\mathcal{N}(\lambda_n)}{n\lambda_n}}\right) .
\end{aligned}$$

Again, we use that $\mathcal{N}(\lambda_n) \leq C_b\lambda_n^{-1/b}$ and

$$\frac{\sqrt{m}_n M}{n\lambda_n} = o\left(\sigma\sqrt{\frac{\lambda_n^{-1/b}}{n\lambda_n}}\right) ,$$

provided

$$m_n \leq n^\alpha \,, \qquad \alpha < \frac{2(br+1)}{2br+b+1} \,.$$

Recalling that $\sigma\sqrt{\frac{\lambda_n^{-1/b}}{n\lambda_n}} = R\lambda_n^r = \lambda_n^{-s}a_n$, we arrive at

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(\tilde{f}_D^{\lambda_n} - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}} \leq C_p\, a_n \,.$$

As a result, for any $1 \leq p$

$$\limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\bar{B}^s(\tilde{f}_D^{\lambda_n} - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}}}{a_n} \leq C_p \,,$$

for some $C_p < \infty$, not depending on the model parameter $(\sigma, M, R) \in \mathbb{R}_+^3$. □

Chapter 5

Adaptivity

5.1 Introduction

We recall that while tuning the regularization parameter is essential for spectral regularization to work well, an *a priori* choice of the regularization parameter is in general not feasible in statistical problems since the choice necessarily depends on unknown structural properties (e.g. smoothness of the target function or behavior of the statistical dimension). This imposes the need for data-driven *a-posteriori* choices of the regularization parameter, which hopefully are optimal in some well defined sense. An attractive approach is (some version of) the balancing principle going back to Lepskii's seminal paper [55] in the context of Gaussian white noise, having been elaborated by Lepskii himself in a series of papers and by other authors, see e.g. [56], [57], [41], [8], [63] and references therein.

We shall briefly describe the basic idea. Originally this method was developed in the framework of the Gaussian white noise model

$$Y_\varepsilon(dt) = f(t)dt + \varepsilon W(dt) \,, \qquad \varepsilon > 0 \,, \quad 0 \le t \le 1 \,,$$

where $f \in L^2([0,1])$ is an unknown function and W is a standard Brownian motion. It is supposed that f belongs to some nested class of functions $\{\mathcal{F}_\theta\}_\theta$ (a common class for all values of ϵ) and θ varies over some parameter space Θ, which is a bounded subset of $\mathbb{R}_+$. For recovering f, one has available a family of estimators $\{\hat{f}_{\varepsilon,\theta}\}_\theta$, based on the observations Y_ε and depending on the parameter θ. For a given estimator $\hat{f}_{\varepsilon,\theta}$, consider the *risk*

$$R_\varepsilon(\hat{f}_{\varepsilon,\theta}, \theta) = \sup_{f \in \mathcal{F}_\theta} \mathbb{E}_f[\, ||f - \hat{f}_{\varepsilon,\theta}|| \,] \,.$$

Given $\theta \in \Theta$, a function $\varphi_\theta(\varepsilon)$ is said to be a *minimax rate of convergence* on the set $\mathcal{F}_\theta$ if

$$\liminf_{\varepsilon \to 0} \inf_{\hat{f}} \; \varphi_\theta^{-1}(\varepsilon) \; R_\varepsilon(\hat{f}, \theta) \; > \; 0 \,, \tag{5.1.1}$$

where the infimum is taken over all possible estimators $\hat{f}$, and if, in addition, there exists an estimator

$\hat{f}_{\varepsilon,\theta}$ such that

$$\limsup_{\varepsilon \to 0} \ \varphi_\theta^{-1}(\varepsilon)\ R_\varepsilon(\hat{f}_{\varepsilon,\theta}, \theta) \ < \ \infty \ . \tag{5.1.2}$$

An estimator satisfying both (5.1.1) and (5.1.2) is called *(asymptotically) minimax*.
Assume that for any $\theta \in \Theta$, there exists a minimax rate of convergence on the set $\mathcal{F}_\theta$ for the signal f. In his papers [55], [56], [57], Lepskii provides answers to the following questions:

1. Given a family $\{\hat{f}_{\varepsilon,\theta}\}_{\theta\in\Theta}$ of asymptotically minimax estimators (with corresponding minimax rates $\{\varphi_\theta\}_\theta$), does there exists an estimator $\tilde{f}_\varepsilon$ for which (5.1.2) holds uniformly for any $\theta \in \Theta$, i.e.

$$\sup_{\theta\in\Theta} \limsup_{\varepsilon \to 0} \ \varphi_\theta^{-1}(\varepsilon)\ R_\varepsilon(\tilde{f}_\varepsilon, \theta) \ < \ \infty \ ? \tag{5.1.3}$$

2. How can one get *adaptation* over the family $\{\mathcal{F}_\theta\}_{\theta\in\Theta}$, i.e. how does one construct a new estimator $\tilde{f}_\varepsilon$ which is uniformly asymptotically minimax and therefore satisfies (5.1.3)?

Such an estimator $\tilde{f}_\varepsilon$, if it exists, is then said to be *optimal adaptive* with respect to the given family $\{\mathcal{F}_\theta\}_{\theta\in\Theta}$.
The adaptive estimation procedure is achieved by applying the following approach: For any $\varepsilon > 0$, choose a suitable finite discretization $\theta_1 < ... < \theta_{m_\varepsilon}$ of Θ. Given some sufficiently large positive constant C, let

$$j_0 = \inf\{\ j \leq m_\varepsilon \ : \ ||\hat{f}_{\varepsilon,\theta_j} - \hat{f}_{\varepsilon,\theta_k}|| \leq C\ R_\varepsilon(\hat{f}_{\varepsilon,\theta_k}, \theta_k)\ , \forall k \in (j, m_\varepsilon]\ \}\ .$$

The final optimal adaptive estimator is defined by setting $\tilde{f}_\varepsilon = \hat{f}_{\varepsilon,\theta_{j_0}}$.

In the context of Learning Theory, there is yet no strict analog of the above procedure, the main problem being in precisely mimicking the crucial condition (5.1.3) on the existence of a supremum over the "maximal" parameter space Θ. But the important algorithmic idea of balancing and discretizing the parameter space has been successfully generalized. We remark that in the context of Learning we shall consider higher dimensional parameter spaces Θ, but in addition we have the real-valued regularization parameter as a family over Θ. Thus the grid on Θ in the original Lepskii approach shall be replaced by a grid for λ which allows to use the ordering on the real line.

The first such version of the balancing principle (using the RKHS structure in an essential way) is in [25] which itself was inspired by [41]. The above strict notion of an asymptotically minimax estimator is in some sense replaced by a slightly more vague notion of finding a fully data-dependent estimator by an analogous Lepskii- procedure, which in some sense minimizes error bounds given by probabilistic estimates of the form

$$\mathbb{P}\left(||\hat{f}_{\mathbf{z}} - f_\rho|| \leq \epsilon(n,\eta)\right) \leq \eta.$$

Here one minimizes $\epsilon(n,\eta)$ by solving a trade-off between the unknown approximation error and the sample error, which can be expressed empirically. Intuitively, this gives a good choice (hopefully: the optimal choice) and one approximates this by balancing over a grid in the regularization parameter. For KRR the authors in [25] replace the above norm by the expected risk, but the paper also discusses general spectral regularization, KRR with convex loss and elastic net. We follow the above approach by studying the error for our norm $||\cdot|| = ||B^s \cdot ||_{\mathcal{H}_1}$. We emphasize that contrary to the other parts of this thesis,

we shall follow the original approach in [25] by staying in the framework of estimates in large probability which we shall not push to estimates in expectation. But concerning the optimality of the estimator obtained by balancing, a few more remarks are appropriate, especially in view of the fact that some part of the machine learning literature seems to use the term *adaptivity* in some very vague sense as a semantical equivalent of (fully) *data-dependent*, without attaching to it any precise notion of optimality. We strongly disagree and therefore we amplify.

The original paper [25] somewhat prophetically refers to the value $\lambda_{opt}(n)$ defined by the intersection of the graphs of sample error and approximation error as *the best choice* but very justly adds the cautionary remark: *However, the rate will be optimal in a minimal sense if the bound we started from is tight. We do not discuss this problem and we refer to [24, 28, 43, 87] for further discussion.* Clearly, the quoted references do not contain a proof of optimality for the balancing principle in the context of the above paper, but they do reference the proof of minimax optimal rates for KRR in the context of *fast* rates in [24], which does not perfectly fit with the slow upper rates of [25]. Furthermore, the article [28] and the book [43] both discuss lower rates of convergence for some explicit classical function spaces of Sobolev, Besov or Hölder type by use of entropy methods. This does not cover in full generality the distribution free approach corresponding to treating largely arbitrary covariance operators B.[1] There is no reference to minimax optimal rates for the other examples. As we are deeply convinced that clarifying somewhat subtle points by formal definitions is an essential part of the culture of mathematics, we could not resist giving a formal definition of a *minimax optimal adaptive estimator* over a parameter space Θ in Definition 5.3.1 below. We have tried to make this definition as general as possible (but staying inside the class of spectral regularization, as we have defined minimax optimality only within that framework in this thesis).

Coming back to [25], it is *assumed* in Assumption 1 of that paper that the estimates of [25] are *uniform with respect to the discretizing grid in the regularization parameter* λ. In fact, with some additional work it is even more or less obvious that this uniformity actually holds true in the context of slow rates. This seems to be connected to the use of an additive error decomposition and, unfortunately, is lost in our approach which is adapted to fast rates (and which, when applied to slow rates, does not give this uniformity). This uniformity implies that the balancing estimator of [25] actually is minimax optimal adaptive in the sense of our Definition 5.3.1, in the context of slow rates, *granted a full proof of lower bounds in the case of slow rates for source conditions induced by general covariance operators*, which everyone seems to expect to hold true. This justifies the wording *best choice* in [25], with hindsight. The class of data generating distributions (the class $\mathcal{M}$ in our Definition 5.3.1), which depends on the context and is different for slow versus fast rates or elastic net etc. is, however, not given explicitly in [25]. One could say that the precise relation between adaptivity and optimality remains vague in [25] for the uninitiated reader, at least in a technical sense. One cannot quite avoid the feeling that the authors of [25] do not fully convey to the reader all technical details they are aware of.

A proper assessment seems to be less clear concerning the recent paper [62]. A possibly minor point is that minimax optimal rates in the context of the very general source conditions of that paper are not precisely referenced (one could e.g. refer to the paper [73] which adapts our proofs in [17] to more general source conditions). But, furthermore, the probabilistic estimates in the context of fast rates (or in the even more general context of [62], where everything is parametrized by the effective dimension as the only

[1] One certainly hopes that there is a deep relation between the operator theoretic approach using spectral theory for a general covariance operator B and the entropy approach to approximation theory for the spaces or source conditions induced by B. This seems to work e.g. for classical Sobolev spaces, but we are not aware of any hard mathematical result establishing this relation in full generality.

free parameter describing the model) are *not* uniform with respect to the regularization parameter λ. This makes necessary a final union bound which is missing (at least in the first version of that paper). Taking into account that the proof of lower bounds on the convergence rates is only sketched and certainly does not cover the full generality of the considered upper bounds (it needs some explicit spectral conditions which are absent in the upper estimates depending only on the effective dimension) it is somewhat unclear what the final class $\mathcal{M}$ of data generating distributions over which adaptivity is proved will turn out to be.

Our analysis in this section faces similar problems. We do not think that it is in final form. [2] To clarify the situation we insist on a formal definition (see Definition 5.3.1). We try to thoroughly track the dependence on a grid in the regularization parameter covering $\lambda_{opt}(n)$ (which is the optimal $\lambda_{opt}(n)$ of [25]) and we do perform the final union bound, which in our case adds (for the estimator obtained by balancing) an additional $\log\log(n)$ to the minimax optimal rate. Thus we could shortly describe our final result as: The estimator obtained by balancing is minimax optimal adaptive up to $\log\log(n)$ over our model classes from Chapter 2 (the regular case) and from Chapter 3 (beyond the regular case) or just *adaptive* in the sense of our Discussion, point 3. below.

5.2 Empirical Effective Dimension

The main point of this subsection is a two-sided estimate on the effective dimension by its empirical approximation which is crucial for our entire approach. We recall the definition of the *effective dimension* and introduce its empirical approximation, the *empirical effective dimension*: For $\lambda \in (0,1]$ we set

$$\mathcal{N}(\lambda) = \mathrm{Tr}\left[\ (\bar{B}+\lambda)^{-1}\bar{B}\ \right]\ , \qquad \mathcal{N}_{\mathbf{x}}(\lambda) = \mathrm{Tr}\left[\ (\bar{B}_{\mathbf{x}}+\lambda)^{-1}\bar{B}_{\mathbf{x}}\ \right]\ , \tag{5.2.1}$$

where we introduce the shorthand notation $\bar{B}_x := \kappa^{-2}B_{\mathbf{x}}$ and similarly $\bar{B} := \kappa^{-2}B$. Here $\mathcal{N}(\lambda)$ depends on the marginal ν (through B), but is considered as deterministic, while $\mathcal{N}_{\mathbf{x}}(\lambda)$ is considered as a random variable (with $B_{\mathbf{x}}$ and B_x, for $x \in \mathcal{X}$ introduced in Chapter 1).
By S^1 we denote the Banach space of trace class operators with norm $||A||_1 = \mathrm{Tr}\,[|A|]$. Furthermore, S^2 denotes the Hilbert space of Hilbert-Schmidt operators with norm $||A||_2 = \mathrm{Tr}\,[A^*A]^{1/2}$. By $||A||$ we denote the operator norm.

Proposition 5.2.1. *For any $\eta \in (0,1)$, with probability at least $1-\eta$*

$$|\ \mathcal{N}(\lambda) - \mathcal{N}_{\mathbf{x}}(\lambda)\ |\ \le\ 2\log(4\eta^{-1})\big(1+\sqrt{\mathcal{N}_{\mathbf{x}}(\lambda)}\big)\left(\frac{2}{\lambda n}+\sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}}\right)\ , \tag{5.2.2}$$

for all $n \in \mathbb{N}^*$ and $\lambda \in (0,1]$.

Corollary 5.2.2. *For any $\eta \in (0,1)$, with probability at least $1-\eta$, one has*

$$\sqrt{\max(\mathcal{N}(\lambda),1)} \le (1+4\delta)\sqrt{\max(\mathcal{N}_{\mathbf{x}}(\lambda),1)}\,,$$

as well as

$$\sqrt{\max(\mathcal{N}_{\mathbf{x}}(\lambda),1)} \le (1+4(\sqrt{\delta}\vee\delta^2))\sqrt{\max(\mathcal{N}(\lambda),1)}\,,$$

[2] Finalizing it is a current joint research effort, see [15].

where $\delta := 2\log(4\eta^{-1})/\sqrt{n\lambda}$ *. In particular, if* $\delta \leq 1$*, with probability at least* $1-\eta$ *one has*

$$\frac{1}{5}\,\sqrt{\max(\mathcal{N}(\lambda),1)} \leq \sqrt{\max(\mathcal{N}_{\mathbf{x}}(\lambda),1)} \leq 5\sqrt{\max(\mathcal{N}(\lambda),1)}\,.$$

Proof of Proposition 5.2.1. We formulate in detail all preliminary results, although they are in principle well known. There are always some subtleties related to inequalities in trace norm. For a proof of the following results we e.g. refer to [74], [31]:

1. If $A \in S^1$ is non-negative, then $||A||_1 = \mathrm{Tr}\,[A]$.
2. $|\mathrm{Tr}\,[A]\,| \leq ||A||_1$.
3. If A is bounded and if $B \in S^1$ is self-adjoint and positive, then $|\mathrm{Tr}\,[AB]\,| \leq ||A||\;|\mathrm{Tr}\,[B]\,|$.
4. If $A, B \in S^2$, then $||AB||_1 \leq ||A||_2\;||B||_2$.
5. If $A \in S^1$, then $||A||_2^2 = |\mathrm{Tr}\,[A^*A]\,| = ||A^*A||_1 \leq ||A||\;||A||_1$.

Consider the algebraic equality

$$\begin{aligned}(\bar{B}+\lambda)^{-1}\bar{B} - (\bar{B}_{\mathbf{x}}+\lambda)^{-1}\bar{B}_{\mathbf{x}} &= (\bar{B}+\lambda)^{-1}(\bar{B}-\bar{B}_{\mathbf{x}}) + (\bar{B}+\lambda)^{-1}(\bar{B}-\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}}+\lambda)^{-1}\bar{B}_{\mathbf{x}} \\ &=: N_1(\lambda,\mathbf{x}) + N_2(\lambda,\mathbf{x})\,. \end{aligned} \tag{5.2.3}$$

Hence,

$$|\,\mathcal{N}(\lambda) - \mathcal{N}_{\mathbf{x}}(\lambda)\,| \leq |\mathrm{Tr}\,[N_1(\lambda,\mathbf{x})]\,| + |\mathrm{Tr}\,[N_2(\lambda,\mathbf{x})]\,|\,. \tag{5.2.4}$$

We want to estimate the first term in (5.2.4) by applying the (classical) Bernstein inequality, Proposition A.1.1 in Chapter A. Setting $\xi(x) = \mathrm{Tr}\left[(\bar{B}+\lambda)^{-1}\bar{B}_x\right]$, $x \in \mathcal{X}$, gives

$$\frac{1}{n}\sum_{j=1}^{n}\xi(x_j) = \mathrm{Tr}\left[(\bar{B}+\lambda)^{-1}\bar{B}_{\mathbf{x}}\right]\,, \qquad \mathbb{E}[\xi] = \mathrm{Tr}\left[(\bar{B}+\lambda)^{-1}\bar{B}\right]\,,$$

and thus

$$\left|\frac{1}{n}\sum_{j=1}^{n}\xi(x_j) - \mathbb{E}[\xi]\right| = |\mathrm{Tr}\,[N_1(\lambda,\mathbf{x})]\,|\,.$$

Recall that $\bar{B}_x$ is positive and $\mathrm{Tr}\left[\bar{B}_x\right] = \kappa^{-2}||S_x||_{HS}^2 \leq 1$. Using 3. leads to

$$|\xi(x)| \leq \left\|(\bar{B}+\lambda)^{-1}\right\|\mathrm{Tr}\left[\bar{B}_x\right] \leq \frac{1}{\lambda} \quad a.s.\,.$$

Note that

$$|\xi(x)| = |\mathrm{Tr}\left[(\bar{B}+\lambda)^{-1}\bar{B}_{\mathbf{x}}\right]\,| = |\mathrm{Tr}\left[\bar{S}_x(\bar{B}+\lambda)^{-1}\bar{S}_x^*\right]\,| = \mathrm{Tr}\,[AA^*]$$

with $A = \bar{S}_x(\bar{B}+\lambda)^{-1/2}$ and by 1. , since AA^* is non-negative. Furthermore, using $\mathbb{E}[\bar{B}_x] = \bar{B}$,

$$\mathbb{E}[|\xi|^2] \leq \frac{1}{\lambda}\mathbb{E}[|\xi|] \leq \frac{1}{\lambda}\mathbb{E}\left[\mathrm{Tr}\left[\bar{S}_x(\bar{B}+\lambda)^{-1}\bar{S}_x^*\right]\right] = \frac{1}{\lambda}\mathrm{Tr}\left[\mathbb{E}[(\bar{B}+\lambda)^{-1}\bar{B}_x]\right] = \frac{1}{\lambda}\mathcal{N}(\lambda)\,.$$

As a result, with probability at least $1 - \frac{\eta}{2}$

$$|\mathrm{Tr}\,[N_1(\lambda, \mathbf{x})]\,| \leq 2\log(4\eta^{-1}) \left(\frac{2}{\lambda n} + \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}} \right) . \tag{5.2.5}$$

Writing $H = (\bar{B}_{\mathbf{x}} + \lambda)^{-1} \bar{B}_{\mathbf{x}}$, we estimate the second term in (5.2.4) using 2. and 4. and obtain

$$|\mathrm{Tr}\,[N_2(\lambda, \mathbf{x})]\,| \leq ||N_2(\lambda, \mathbf{x})||_1 \leq ||N_1(\lambda, \mathbf{x})||_2 \, ||H||_2 \, .$$

From Proposition A.1.3, we have with probability at least $1 - \frac{\eta}{2}$,

$$\|N_1\|_2 = ||(\bar{B} + \lambda)^{-1}(\bar{B} - \bar{B}_{\mathbf{x}})||_2 \ \leq 2\log(4\eta^{-1}) \left(\frac{2}{n\lambda} + \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}} \right) .$$

Finally, recalling that $||H|| \leq 1$ we get from 5.

$$||H||_2 \ \leq \ ||H||^{1/2} \, ||H||_1^{1/2} \ \leq \ \sqrt{\mathcal{N}_{\mathbf{x}}(\lambda)} \quad a.s. \ ,$$

where we used that $\mathrm{Tr}\,[H] = \mathrm{Tr}\,[AA^*]$, with $A = S_x(\bar{B}_{\mathbf{x}} + \lambda)^{-1/2}$ and point 1. . Collecting all pieces gives the result. □

Proof of Corollary 5.2.2. Since $\log(4\eta^{-1}) \geq 1$, the inequality of Proposition 5.2.1 implies that with probability at least $1 - \eta$:

$$|\,\mathcal{N}(\lambda) - \mathcal{N}_{\mathbf{x}}(\lambda)\,| \ \leq \ \frac{2\log(4\eta^{-1})}{\sqrt{\lambda n}} \big(1 + \sqrt{\mathcal{N}_{\mathbf{x}}(\lambda)}\big) \left(\frac{2\log(4\eta^{-1})}{\sqrt{\lambda n}} + \sqrt{\mathcal{N}(\lambda)} \right) .$$

Put $A := \sqrt{\mathcal{N}(\lambda)}$, $B := \sqrt{\mathcal{N}_{\mathbf{x}}(\lambda)}$, and $\delta := \frac{2\log(4\eta^{-1})}{\sqrt{\lambda n}}$, then one can rewrite the above as $\left|A^2 - B^2\right| \leq \delta(1 + B)(\delta + A)$.

Consider the case $A \geq B$. Then the above inequality is $A^2 - A\delta(1 + B) - (B^2 + \delta^2(1 + B)) \leq 0$. Observe that the larger root x^+ of the quadratic equation $x^2 + bx + c$ (for $b, c \leq 0$) is bounded as

$$x^+ = \frac{-b + \sqrt{b^2 - 4c}}{2} \leq |b| + \sqrt{|c|}\,,$$

while the smaller root x^- is negative. Hence, for $x \geq 0$

$$(x - x^+)(x - x^-) \leq 0 \quad \Longrightarrow \quad x \leq x^+ \leq |b| + \sqrt{|c|}\,.$$

Applying this to the above quadratic inequality (solved in $A \geq 0$), we obtain

$$A \leq \delta(1 + B) + \sqrt{B^2 + \delta^2(1 + B)} \leq (1 + \delta)B + \delta + \delta + \delta\sqrt{B} \leq (1 + 2\delta)(B \vee 1) + 2\delta.$$

Similarly, if $B \geq A$, the initial inequality becomes $B^2 - B\delta(\delta + A) - (A^2 + \delta(\delta + A)) \leq 0$ solving this in B and bounding as above we get

$$B \leq \delta(\delta + A) + \sqrt{A^2 + \delta(\delta + A)} \leq (1 + \delta)A + \delta^2 + \delta + \sqrt{\delta A} \leq (1 + 2(\delta \vee \sqrt{\delta}))(A \vee 1) + 2(\delta^2 \vee \delta)\,.$$

The rest of the proof follows by observing that $1 \leq B \vee 1$, $1 \leq A \vee 1$ and

$$2(\delta \vee \sqrt{\delta}) + 2(\delta^2 \vee \delta) \leq 4(\sqrt{\delta} \vee \delta^2) \ .$$

□

5.3 Balancing Principle

In this section, we present the main ideas related to the *Balancing Principle* . Before we present our somewhat abstract approach, we shall explain the general idea in a specific example. Recall the setting of Chapter 2.5 . Combining methods from this chapter with A.2.4 we can establish the error decomposition

$$\left\|(\bar{B}_{\mathbf{x}} + \lambda)^s (f_\rho - f_{\mathbf{z}}^\lambda)\right\|_{\mathcal{H}_1} \ \leq \ C_s(\eta)\lambda^s \left(R\lambda^r + \frac{\sigma}{\sqrt{n}}\lambda^{-\frac{b+1}{2b}} + d(n,\lambda) \right) \ , \tag{5.3.1}$$

with probability at least $1-\eta$, for any $\eta \in (0,1)$, provided n is big enough and $\lambda_0 \leq \lambda \leq 1$, for some $\lambda_0 \in (0,1)$. Here, the function $\lambda \mapsto R\lambda^r$ is the leading order of an upper bound for the *approximation error* and $\lambda \mapsto \frac{\sigma}{\sqrt{n}}\lambda^{-\frac{b+1}{2b}}$ is the leading order of an upper bound for the *sample error*, while $d(n,\lambda)$ will be shown to be subleading (at least for a good choice of the regularization parameter λ). We recall that the optimal regularization parameter (as well as the rate of convergence) is determined by the source condition assumption $f_\rho \in \Omega_\nu(r,R)$, by an assumed power decay of the effective dimension $\mathcal{N}(\lambda) \leq C_b\lambda^{-1/b}$ with intrinsic dimensionality $b > 1$ and by the noise variance $\sigma^2 > 0$. We combine these parameters in a vector (γ,θ) with $\gamma = (\sigma, R) \in \Gamma = \mathbb{R}_+ \times \mathbb{R}_+$ and $\theta = (r,b) \in \Theta = \mathbb{R}_+ \times (1,\infty)$. We had chosen the optimal regularization parameter $\lambda_{n,(\gamma,\theta)}$ by balancing the two leading error terms, more precisely by choosing $\lambda_{n,(\gamma,\theta)}$ as the unique solution of

$$R\lambda^r = \sigma\lambda^{-\frac{b+1}{2b}} \ . \tag{5.3.2}$$

The resulting error estimate is

$$\left\| \bar{B}^s \left(f_\rho - f_{\mathbf{z}}^{\lambda_{n,(\gamma,\theta)}} \right) \right\|_{\mathcal{H}_1} \ \leq \ 2C_s(\eta)\lambda_{n,(\gamma,\theta)}^{s+r} \ ,$$

with probability at least $1-\eta$ and n sufficiently large (see (2.5.4)). The associated sequence of estimated solutions $(f_{\mathbf{z}}^{\lambda_{n,(\gamma,\theta)}})_{n\in\mathbb{N}}$, depending on the regularization parameters $(\lambda_{n,(\gamma,\theta)})_{(n,\gamma)\in\mathbb{N}\times\Gamma}$ was called weak/ strong minimax optimal over the model family $(\mathcal{M}_{(\gamma,\theta)})_{(\gamma,\theta)\in\Gamma\times\Theta}$ with rate of convergence given by $(a_{n,(\gamma,\theta)})_{(n,\gamma)\in\mathbb{N}\times\Gamma}$, **pointwisely for any fixed** $\theta \in \Theta$.

We can formulate (5.3.1) more generally, namely with probability at least $1-\eta$

$$\left\|(\bar{B}_{\mathbf{x}} + \lambda)^s (f_\rho - f_{\mathbf{z}}^\lambda)\right\|_{\mathcal{H}_1} \ \leq \ C_s(\eta)\lambda^s \left(\tilde{\mathcal{A}}(\lambda) + \tilde{\mathcal{S}}(n,\lambda) \right) \ ,$$

where $\mathcal{A}(\cdot)$ is a function upper bounding the approximation error and $\mathcal{S}(n,\cdot)$ is an upper bound for the sample error. However, if the functions $\mathcal{A}(\cdot)$ and $\mathcal{S}(n,\cdot)$ are unknown (e.g., if the value of r in the Hölder-type source condition or the intrinsic dimensionalty $b > 1$ is unknown), an *a priori* choice of the theoretically best value $\lambda_{n,(\gamma,\theta)}$ as in (5.3.2) is impossible. Therefore, it is necessary to use some *a posteriori* choice of λ, independent of the parameter $\theta = (r,b) \in \Theta$. Our aim is to construct an estimator $f_{\mathbf{z}}^{\hat{\lambda}_{n,\gamma}(\mathbf{z})}$, i.e. to find a sequence of regularization parameters $(\hat{\lambda}_{n,\gamma}(\mathbf{z}))_n$, without knowledge of $\theta \in \Theta$,

but depending on the data $\mathbf{z}$, the confidence level $\eta \in (0,1)$ and possibly on $\gamma \in \Gamma$, such that $f_{\mathbf{z}}^{\hat{\lambda}_n(\mathbf{z},\eta,\gamma)}$ is *(minimax) optimal adaptive* in the following sense:

Definition 5.3.1. *Let Γ, Θ be sets and let, for $(\gamma,\theta) \in \Gamma \times \Theta$, $\mathcal{M}_{(\gamma,\theta)}$ be a class of data generating distributions on $\mathcal{X} \times \mathcal{Y}$. For each $\lambda \in (0,1]$ let $(\mathcal{X} \times \mathcal{Y})^n \ni \mathbf{z} \longmapsto f_{\mathbf{z}}^{\lambda} \in \mathcal{H}_1$ be an algorithm. If there is a sequence $(a_{n,(\gamma,\theta)})_{n\in\mathbb{N}}$ $(\gamma,\theta) \in \Gamma \times \Theta$ and a parameter choice $(\hat{\lambda}_{n,\gamma,\tau}(\mathbf{z}))_{(n,\gamma)\in\mathbb{N}\times\Gamma}$ (not depending on $\theta \in \Theta$) such that*

$$\lim_{\tau\to\infty} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{(\gamma,\theta)}} \rho^{\otimes n}\left(\left\|\bar{B}^s(f_{\mathbf{z}}^{\hat{\lambda}_{n,\gamma,\tau}(\mathbf{z})} - f_\rho)\right\|_{\mathcal{H}_1} \geq \tau a_{n,(\gamma,\theta)}\right) = 0 \tag{5.3.3}$$

and

$$\lim_{\tau\to 0} \liminf_{n\to\infty} \inf_{\hat{f}} \sup_{\rho\in\mathcal{M}_{(\gamma,\theta)}} \rho^{\otimes n}\left(\left\|\bar{B}^s(\hat{f} - f_\rho)\right\|_{\mathcal{H}_1} \geq \tau a_{n,(\gamma,\theta)}\right) > 0, \tag{5.3.4}$$

where the infimum is taken over all estimators $\hat{f}$, then the sequence of estimators $(f_{\mathbf{z}}^{\hat{\lambda}_{n,\gamma,\eta}(\mathbf{z})})_{n\in\mathbb{N}}$ is called minimax optimal adaptive over Θ and the model family $(\mathcal{M}_{(\gamma,\theta)})_{(\gamma,\theta)\in\Gamma\times\Theta}$, with respect to the family of rates $(a_{n,(\gamma,\theta)})_{(n,\gamma)\in\mathbb{N}\times\Gamma}$, for the interpolation norm of parameter $s \in [0,\frac{1}{2}]$.

We remind the reader that analogously to our discussion in Chapter 2 and Chapter 3, upper estimates typically hold on a class $\mathcal{M}^{<}_{(\gamma,\theta)}$ and lower estimates hold on a possibly different class $\mathcal{M}^{>}_{(\gamma,\theta)}$, the model class $\mathcal{M}_{(\gamma,\theta)}$ in the above definition being the intersection of both.

This definition is different from our discussion in Chapters 2 and 3. Estimates in expectation would make the definition somewhat smoother. If one has equation (5.3.3) only with a_n multiplied by an exploding logarithmic factor $O(\log^k n)$ for some fixed k, we shall simply say that the sequence of estimators is *adaptive*. We emphasize that these definitions are not standard. The existing literature in learning theory rather prefers to be somewhat vague concerning the question with respect to which models the new estimator is optimal adaptive. But we think that our definition above adequately formalizes what has actually been done and seems to be accepted. [3]

One could generalize Definition 5.3.1 by allowing *"minimax optimal"* estimators which are not necessarily constructed via spectral regularizaton. Since throughout this thesis we have defined minimax optimality only within the framework of spectral regularization in Definition 2.3.3, we have refrained from doing so, but we are aware of the fact that in other contexts this might be natural.

To find such an adaptive estimator, we apply a method which is known in the statistical literature as *Balancing Principle*. Throughout this section we need

Assumption 5.3.2. *Let $\mathcal{M}$ be a class of models. We consider a discrete set of possible values for the regularization parameter*

$$\Lambda_m \;=\; \{\, \lambda_j \;:\; 0 < \lambda_0 < \lambda_1 < \ldots < \lambda_m \,\}\,.$$

for some $m \in \mathbb{N}$. Let $s \in [0,\frac{1}{2}]$ and $\eta \in (0,1]$. We assume to have the following error decomposition

[3] Definitions mark a subtle and important point of contact between written texts and the culture of mathematics. Perhaps no one has ever expressed this more passionately than Stendhal: *Je n'ai jamais trouvé qu'une idée dans ce diable de livre, et encore elle n'était pas de Cailhava, mais bien de Bacon. Mais n'est-ce rien qu'une idée, dans un livre? Il s'agit de la définition du rire. Ma cohabitation passionnée avec les mathématiques m'a laissé un amour fou pour les bonnes définitions, sans lesquelles il n'y a que des à peu près,* [Stendhal, *Vie de Henry Brulard* , p. 95]. It would be next to impossible to find similar words on the emotional need for clarity and precision in any classical German text, even up to modern times, and I am deeply thankful for this very French contribution to our universal culture.

uniformly over the grid Λ_m*:*

$$\left\|(\bar{B}_{\mathbf{x}}+\lambda)^s(f_\rho-f_{\mathbf{z}}^\lambda)\right\|_{\mathcal{H}_1} \;\le\; C_s(m,\eta)\;\lambda^s\left(\;\tilde{\mathcal{A}}(\lambda)+\tilde{\mathcal{S}}(n,\lambda)\;\right)\;, \tag{5.3.5}$$

where

$$C_s(m,\eta)=C_s\log^2(8|\Lambda_m|\eta^{-1})\;,\qquad C_s>0\;, \tag{5.3.6}$$

with probability at least $1-\eta$*, for all data generating distributions from* $\mathcal{M}$*. The bounds* $\tilde{\mathcal{A}}(\lambda)$ *and* $\tilde{\mathcal{S}}(n,\lambda)$ *are given by*

$$\tilde{\mathcal{S}}(n,\lambda)=\mathcal{S}(n,\lambda)+d_1(n,\lambda)\;,\quad \mathcal{S}(n,\lambda)=\sigma\sqrt{\frac{\tilde{\mathcal{N}}(\lambda)}{n\lambda}}\;,\qquad d_1(n,\lambda)=\frac{M}{n\lambda}\;,$$

with $\tilde{\mathcal{N}}(\lambda)=\max(\mathcal{N}(\lambda),1)$ *and*

$$\tilde{\mathcal{A}}(\lambda)=\mathcal{A}(\lambda)+d_2(n)\;,\quad d_2(n)=\frac{C}{\sqrt{n}}\;,$$

where $\mathcal{A}(\lambda)$ *is increasing, satisfying* $\lim_{\lambda\to 0}\mathcal{A}(\lambda)=0$ *and for some constant* $C<\infty$*. We further define* $d(n,\lambda):=d_1(n,\lambda)+d_2(n)$*.*

We remark that it is actually sufficient to assume (5.3.5) for $s=0$ and $s=\frac{1}{2}$. Interpolation via inequality $||B^sf||_{\mathcal{H}_1}\le||\sqrt{B}f||_{\mathcal{H}_1}^{2s}\;||f||_{\mathcal{H}_1}^{1-2s}$ implies validity of (5.3.5) for any $s\in[0,\frac{1}{2}]$.

Note that for any $s\in[0,\frac{1}{2}]$, the map $\lambda\mapsto\lambda^s\mathcal{S}(n,\lambda)$ as well as $\lambda\mapsto\lambda^s d_1(n,\lambda)$ are strictly decreasing in λ. Also, if n is sufficiently large and if λ is sufficiently small, $\tilde{\mathcal{A}}(\lambda)\le\tilde{\mathcal{S}}(n,\lambda)$.

We let

$$\lambda_{opt}(n):=\sup\{\lambda:\;\tilde{\mathcal{A}}(\lambda)\le\tilde{\mathcal{S}}(n,\lambda)\}\;.$$

In this definition we have replaced $\mathcal{A}(\lambda)$, $\mathcal{S}(n,\lambda)$ by $\tilde{\mathcal{A}}(\lambda)$ and $\tilde{\mathcal{S}}(n,\lambda)$, thus including the remainder terms $d_1(n,\lambda)$ and $d_2(n)$ into our definition of $\lambda_{opt}(n)$. It will emerge *a-posteriori*, that the definition of $\lambda_{opt}(n)$ is not affected, since the remainder terms are subleading. But *a priori*, this is not known. A correct proof of the crucial oracle inequality in Lemma 5.3.8 below is much easier with this definition of $\lambda_{opt}(n)$. It will then finally turn out that the remainder terms are really subleading.

The grid Λ_m has to be designed such that the optimal value $\lambda_{opt}(n)$ is contained in $[\lambda_0,\lambda_m]$ Note that this definition requires neither strict monotonicity of $\mathcal{A}$ nor continuity.

The best estimator for $\lambda_{opt}(n)$ within Λ_m belongs to the set

$$\mathcal{J}(\Lambda_m)\;=\;\left\{\lambda_j\in\Lambda_m\;:\;\tilde{\mathcal{A}}(\lambda_j)\le\tilde{\mathcal{S}}(n,\lambda_j)\right\}$$

and is given by

$$\lambda_*:=\max\;\mathcal{J}(\Lambda_m)\;. \tag{5.3.7}$$

In particular, since we assume that $\mathcal{J}(\Lambda_m)\neq\emptyset$ and $\Lambda_m\setminus\mathcal{J}(\Lambda_m)\neq\emptyset$, there is some $l\in\mathbb{N}$ such that $\lambda_l=\lambda_*\le\lambda_{opt}(n)\le\lambda_{l+1}$. Note also that the choice of the grid Λ_m has to depend on n.

Before we define the balancing principle estimate of $\lambda_{opt}(n)$, we give some intuition of its possible choice:

For any $\lambda \leq \lambda_{opt}(n)$, we have $\tilde{\mathcal{A}}(\lambda) \leq \tilde{\mathcal{S}}(n,\lambda)$. Moreover, Lemma A.1.6 yields for given $\lambda_1 \leq \lambda_2$

$$\left\|(\bar{B}_{\mathbf{x}} + \lambda_1)^s f\right\|_{\mathcal{H}_1} \leq \left\|(\bar{B}_{\mathbf{x}} + \lambda_2)^s f\right\|_{\mathcal{H}_1} .$$

Finally, since $\lambda \mapsto \lambda^s \tilde{\mathcal{S}}(n,\lambda)$ is decreasing, Assumption 5.3.2 gives for any two $\lambda, \lambda' \in \mathcal{J}(\Lambda_m)$ satisfying $\lambda' \leq \lambda$, with probability at least $1-\eta$

$$\begin{aligned}
\left\|(\bar{B}_{\mathbf{x}} + \lambda')^s (f_{\mathbf{z}}^{\lambda'} - f_{\mathbf{z}}^{\lambda})\right\|_{\mathcal{H}_1} &\leq \left\|(\bar{B}_{\mathbf{x}} + \lambda')^s (f_\rho - f_{\mathbf{z}}^{\lambda'})\right\|_{\mathcal{H}_1} + \left\|(\bar{B}_{\mathbf{x}} + \lambda')^s (f_\rho - f_{\mathbf{z}}^{\lambda})\right\|_{\mathcal{H}_1} \\
&\leq \left\|(\bar{B}_{\mathbf{x}} + \lambda')^s (f_\rho - f_{\mathbf{z}}^{\lambda'})\right\|_{\mathcal{H}_1} + \left\|(\bar{B}_{\mathbf{x}} + \lambda)^s (f_\rho - f_{\mathbf{z}}^{\lambda})\right\|_{\mathcal{H}_1} \\
&\leq C_s(m,\eta)\, \lambda'^s \left(\tilde{\mathcal{A}}(\lambda') + \tilde{\mathcal{S}}(n,\lambda') \right) + \\
&\qquad + \;\; C_s(m,\eta)\, \lambda^s \left(\tilde{\mathcal{A}}(\lambda) + \tilde{\mathcal{S}}(n,\lambda) \right) \\
&\leq 4 C_s(m,\eta)\, \lambda'^s \tilde{\mathcal{S}}(n,\lambda') . \qquad (5.3.8)
\end{aligned}$$

An essential step is to find an empirical approximation of the sample error. In view of Corollary 5.2.2 we define

$$\tilde{\mathcal{S}}_{\mathbf{x}}(n,\lambda) = \mathcal{S}_{\mathbf{x}}(n,\lambda) + d_1(n,\lambda) , \quad \mathcal{S}_{\mathbf{x}}(n,\lambda) = \sigma \sqrt{\frac{\tilde{\mathcal{N}}_{\mathbf{x}}(\lambda)}{n\lambda}} ,$$

with $\tilde{\mathcal{N}}_{\mathbf{x}}(\lambda) = \max(\mathcal{N}_{\mathbf{x}}(\lambda), 1)$ and $\mathcal{N}_{\mathbf{x}}(\lambda)$ the empirical effective dimension given in (5.2.1). Corollary 5.2.2 implies uniformly in $\lambda \in \Lambda_m$

$$\frac{1}{5}\tilde{\mathcal{S}}_{\mathbf{x}}(n,\lambda) \leq \tilde{\mathcal{S}}(n,\lambda) \leq 5\tilde{\mathcal{S}}_{\mathbf{x}}(n,\lambda) , \qquad (5.3.9)$$

with probability at least $1-\eta$, provided

$$n\lambda_0 \geq 2 , \qquad 2\log(4|\Lambda_m|\eta^{-1}) \leq \sqrt{n\lambda_0} . \qquad (5.3.10)$$

Substituting (5.3.9) into the rhs of the estimate (5.3.8) motivates our definition of the balancing principle estimate of $\lambda_{opt}(n)$ as follows:

Definition 5.3.3. *Given $s \in [0, \frac{1}{2}]$, $\eta \in (0,1]$ and $\mathbf{z} \in \mathbf{Z}^n$, we set*

$$\begin{aligned}
\mathcal{J}_{\mathbf{z}}^{+}(\Lambda_m) = \{\, \lambda \in \Lambda_m \;:\; &||(\bar{B}_{\mathbf{x}} + \lambda')^s (f_{\mathbf{z}}^{\lambda} - f_{\mathbf{z}}^{\lambda'})||_{\mathcal{H}_1} \;\leq\; 20 C_s(m,\eta/2)\, \lambda'^s\, \tilde{\mathcal{S}}_{\mathbf{x}}(n,\lambda') , \\
&\forall \lambda' \in \Lambda_m,\; \lambda' \leq \lambda \,\}
\end{aligned}$$

and define

$$\hat{\lambda}_s(\mathbf{z}) := \max\, \mathcal{J}_{\mathbf{z}}^{+}(\Lambda_m) . \qquad (5.3.11)$$

Notice that $\mathcal{J}_{\mathbf{z}}^{+}(\Lambda_m)$ as well as $\hat{\lambda}_s(\mathbf{z})$ depend on the confidence level $\eta \in (0,1]$.

For the analysis it will be important that the grid Λ_m has a certain regularity. We summarize all requirements needed in

Assumption 5.3.4. *(on the grid)*

1. *Assume that $\mathcal{J}(\Lambda_m) \neq \emptyset$ and $\Lambda_m \setminus \mathcal{J}(\Lambda_m) \neq \emptyset$.*
2. *(Regularity of the grid) There is some $q > 1$ such that the elements in the grid obey $1 < \lambda_{j+1}/\lambda_j \leq q$, $j = 0, ..., m$.*

3. *Choose $\lambda_0 = \lambda_0(n)$ as the unique solution of $n\lambda = \mathcal{N}(\lambda)$. We require that n is sufficiently large, such that $\mathcal{N}(\lambda_0(n)) \geq 1$ (so that the maximum in the definition of $\tilde{\mathcal{N}}(\lambda)$ can be dropped). We further assume that $n\lambda_0 \geq 2$.*

Note that $\lambda_0(n) \to 0$ as $n \to \infty$. Then, since $\mathcal{N}(\lambda) \to \infty$ as $\lambda \to 0$, we get that this $\lambda_0 = \lambda_0(n)$ satisfies $\lambda_0 n = \mathcal{N}(\lambda_0) \to \infty$. Furthermore, a short argument shows that the optimal value $\lambda_{opt}(n)$ indeed satisfies $\lambda_0 \leq \lambda_{opt}(n)$, if n is big enough. Since $\mathcal{A}(\lambda) \to 0$ as $\lambda \to 0$, we get $\tilde{\mathcal{A}}(\lambda_0(n)) \to 0$ as $n \to \infty$. Since $\tilde{\mathcal{S}}(n, \lambda_0(n)) = 1 + \frac{M}{n\lambda_0(n)}$ by definition, it follows $\tilde{\mathcal{A}}(\lambda_0(n)) \leq \tilde{\mathcal{S}}(n, \lambda_0(n))$ for n big enough. From the definition of $\lambda_{opt}(n)$ as a supremum, we actually have $\lambda_0(n) \leq \lambda_{opt}(n)$, for n sufficiently large.

Under the regularity assumption, we find that

$$\tilde{S}(n, \lambda_j) < q\tilde{S}(n, \lambda_{j+1}) \,, \quad j = 0, ..., m \,. \tag{5.3.12}$$

Indeed, while the effective dimension $\lambda \to \mathcal{N}(\lambda)$ is decreasing, the related function $\lambda \to \lambda\mathcal{N}(\lambda)$ is non-decreasing. Hence we find that

$$q^{-1}\mathcal{N}(\lambda) = (q\lambda)^{-1}\lambda\mathcal{N}(\lambda) < (q\lambda)^{-1}(q\lambda)\mathcal{N}(q\lambda) = \mathcal{N}(q\lambda)$$

and since $q > 1$

$$q^{-1}\tilde{\mathcal{N}}(\lambda) = \max\left(q^{-1}\mathcal{N}(\lambda), q^{-1}\right) < \max(\mathcal{N}(q\lambda), 1) = \tilde{\mathcal{N}}(q\lambda) \,.$$

Therefore

$$q^{-1}\mathcal{S}(n, \lambda_j) = \sigma\sqrt{\frac{q^{-1}\tilde{\mathcal{N}}(\lambda_j)}{nq\lambda_j}} < \sigma\sqrt{\frac{\tilde{\mathcal{N}}(\lambda_{j+1})}{n\lambda_{j+1}}} = \mathcal{S}(n, \lambda_{j+1}) \,.$$

One also easily verifies that

$$d_1(n, \lambda_j) = \frac{M}{n\lambda_j} \leq \frac{qM}{n\lambda_{j+1}} = qd_1(n, \lambda_{j+1}) \,,$$

implying (5.3.12).

Remark 5.3.5. *The typical case for Assumption 5.3.4 to hold is given when the parameters λ_j follow a geometric progression, i.e., for some $q > 1$ we let $\lambda_j := \lambda_0 q^j$, $j = 1, ..., m$ and with $\lambda_m = 1$. In this case we are able to upper bounding the total number of grid points $|\Lambda_m|$ in terms of $\log(n)$. In fact, since $\lambda_m = 1 = \lambda_0 q^m$, simple calculations lead to*

$$|\Lambda_m| = m + 1 = 1 - \frac{\log(\lambda_0)}{\log(q)} \,.$$

Recall that the starting point λ_0 is required to obey $\mathcal{N}(\lambda_0) = n\lambda_0 \geq 2$ if n is sufficiently large, implying $-\log(\lambda_0) \leq -\log\left(\frac{2}{n}\right) \leq \log(n)$. Finally, we obtain for n sufficiently large

$$|\Lambda_m| \leq C_q \log(n) \,, \tag{5.3.13}$$

with $C_q = \log(q)^{-1} + 1$.

We shall need an additional assumption on the effective dimension:

Assumption 5.3.6. *1. For some $\gamma_1 \in (0,1]$ and for any λ sufficiently small*

$$\mathcal{N}(\lambda) \geq C_1 \lambda^{-\gamma_1} ,$$

for some $C_1 > 0$.

2. For some $\gamma_2 \in (0,1]$ and for any λ sufficiently small

$$\mathcal{N}(\lambda) \leq C_2 \lambda^{-\gamma_2} ,$$

for some $C_2 > 0$.

Note that such an additional assumption restricts the class of admissible marginals and shrinks the class $\mathcal{M}$ in Assumption 5.3.2 to a subclass $\mathcal{M}'$. Such a lower and upper bound will hold in all examples which we encounter in Section 5.4.

We further remark that Assumption 5.3.6 ensures a precise asymptotic behaviour for $\lambda_0 = n^{-1}\mathcal{N}(\lambda_0)$ of the form

$$C_{\gamma_1} \left(\frac{1}{n}\right)^{\frac{1}{1+\gamma_1}} \leq \lambda_0(n) \leq C_{\gamma_2} \left(\frac{1}{n}\right)^{\frac{1}{1+\gamma_2}} , \tag{5.3.14}$$

for some $C_{\gamma_1} > 0$, $C_{\gamma_2} > 0$.

Main Results

The first result is of preparatory character.

Proposition 5.3.7. *Let Assumption* 5.3.2 *be satisfied. Define λ_* as in* (5.3.7)*. Assume $n\lambda_0 \geq 2$. Then for any*

$$\eta \geq \eta_n := \min\left(1 \, , \; 4|\Lambda_m| \exp\left(-\frac{1}{2}\sqrt{\mathcal{N}(\lambda_0(n))}\right)\right) ,$$

uniformly over $\mathcal{M}$, with probability at least $1-\eta$

$$\left\| (\bar{B}_{\mathbf{x}} + \lambda_*)^s (f_{\mathbf{z}}^{\hat{\lambda}_s(\mathbf{z})} - f_\rho) \right\|_{\mathcal{H}_1} \leq 102 C_s(m, \eta/2) \lambda_*^s \, \tilde{\mathcal{S}}(n, \lambda_*) .$$

We shall need

Lemma 5.3.8. *If Assumption* 5.3.4 *holds, then*

$$\lambda_*^s \, \tilde{\mathcal{S}}(n, \lambda_*) \leq q^{1-s} \min_{\lambda \in [\lambda_0, \lambda_m]} \{ \, \lambda^s \, (\tilde{\mathcal{A}}(\lambda) + \tilde{\mathcal{S}}(n, \lambda)) \, \} . \tag{5.3.15}$$

We immediately arrive at our first main result of this section:

Theorem 5.3.9. *Let Assumption* 5.3.2 *be satisfied and suppose the grid obeys Assumption* 5.3.4*. Then for any*

$$\eta \geq \eta_n := \min\left(1 \, , \; 4|\Lambda_m| \exp\left(-\frac{1}{2}\sqrt{\mathcal{N}(\lambda_0(n))}\right)\right) ,$$

uniformly over $\mathcal{M}$*, with probability at least* $1-\eta$

$$\left\|\bar{B}^s\left(f_{\mathbf{z}}^{\hat{\lambda}_s(\mathbf{z})}-f_\rho\right)\right\|_{\mathcal{H}_1} \;\leq\; q^{1-s}\, D_s(m,\eta) \min_{\lambda\in[\lambda_0,\lambda_m]}\{\,\lambda^s(\tilde{\mathcal{A}}(\lambda)+\tilde{\mathcal{S}}(n,\lambda))\,\}\,,$$

with

$$D_s(m,\eta)=C_s'\log^{2(s+1)}(16|\Lambda_m|\eta^{-1})\,,$$

for some $C_s'>0$*.*

In particular, choosing a geometric grid and assuming a lower and upper bound on the effective dimension, we obtain:

Corollary 5.3.10. *Let Assumption* 5.3.2*, Assumption 5.3.4 and Assumption 5.3.6 be satisfied. Suppose the grid is given by a geometric sequence* $\lambda_j=\lambda_0 q^j$*, with* $q>1$*,* $j=1,...,m$ *and with* $\lambda_m=1$*. Then for any*

$$\eta\geq\eta_n:=4C_q\log(n)\exp\left(-C_{\gamma_1,\gamma_2}n^{\frac{\gamma_1}{2(1+\gamma_2)}}\right)\,,$$

uniformly over $\mathcal{M}'$*, with probability at least* $1-\eta$

$$\left\|\bar{B}^s\left(f_{\mathbf{z}}^{\hat{\lambda}_s(\mathbf{z})}-f_\rho\right)\right\|_{\mathcal{H}_1} \;\leq\; \tilde{D}_{s,q}(n,\eta) \min_{\lambda\in[\lambda_0,1]}\{\,\lambda^s(\tilde{\mathcal{A}}(\lambda)+\tilde{\mathcal{S}}(n,\lambda))\,\}\,,$$

with

$$\tilde{D}_{s,q}(n,\eta)=C_{s,q}\log^{2(s+1)}(\log(n))\log^{2(s+1)}(16\eta^{-1})\,,$$

for some $C_{\gamma_1,\gamma_2}>0$ *and some* $C_{s,q}>0$*, provided* n *is sufficiently large.*

Note that $\eta_n\to 0$ as $n\to\infty$.

One for All: L^2-Balancing is sufficient !

This section is due to an idea suggested by P. Mathé (which itself was inspired by the work [11]) which we have worked out in detail. We define the $L^2(\nu)-$ balancing estimate $\hat{\lambda}_{1/2}(\mathbf{z})$ according to Definition 5.3.3 by explicitely choosing $s=\frac{1}{2}$ (in contrast to Theorem 5.3.9, where we choose $\hat{\lambda}_s(\mathbf{z})$ depending on the norm parameter s). Our main result states that balancing in the $L^2(\nu)-$ norm suffices to automatically give balancing in all other (stronger !) intermediate norms $||\cdot||_s$, for any $s\in[0,\frac{1}{2}]$.

Theorem 5.3.11. *Let Assumption* 5.3.2 *and Assumption 5.3.4 be satisfied and suppose the grid obeys Assumption* 5.3.4*. Then for any*

$$\eta\geq\eta_n:=\min\left(1\,,\;4|\Lambda_m|\exp\left(-\frac{1}{2}\sqrt{\mathcal{N}(\lambda_0(n))}\right)\right)\,,$$

uniformly over $\mathcal{M}$*, with probability at least* $1-\eta$

$$\left\|\bar{B}^s(f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})}-f_\rho)\right\|_{\mathcal{H}_1} \;\leq\; q^{1-s}\hat{D}_s(m,\eta) \min_{\lambda\in[\lambda_0,\lambda_m]}\{\,\lambda^s(\tilde{\mathcal{A}}(\lambda)+\tilde{\mathcal{S}}(n,\lambda))\,\}\,,$$

with

$$\hat{D}_s(m,\eta) = C'_s \log^{2(s+1)}(16|\Lambda_m|\eta^{-1}) \ ,$$

for some $C'_s > 0$.

In particular, choosing a geometric grid and assuming a lower and upper bound on the effective dimension, we obtain:

Corollary 5.3.12. *Let Assumption* 5.3.2*, Assumption 5.3.4 and Assumption 5.3.6 be satisfied. Suppose the grid is given by a geometric sequence* $\lambda_j = \lambda_0 q^j$*, with* $q > 1$*,* $j = 1, ..., m$ *and with* $\lambda_m = 1$*. Then, for* n *sufficiently large and for any*

$$\eta \geq \eta_n := 4C_q \log(n) \exp\left(-C_{\gamma_1,\gamma_2} n^{\frac{\gamma_1}{2(1+\gamma_2)}}\right) \ ,$$

uniformly over $\mathcal{M}'$*, with probability at least* $1 - \eta$

$$\left\| \bar{B}^s (f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})} - f_\rho) \right\|_{\mathcal{H}_1} \ \leq \ q^{1-s} \hat{D}_{s,q}(n,\eta) \min_{\lambda \in [\lambda_0, 1]} \{ \ \lambda^s (\tilde{\mathcal{A}}(\lambda) + \tilde{\mathcal{S}}(n,\lambda)) \ \} \ ,$$

with

$$\hat{D}_{s,q}(n,\eta) = C_{s,q} \log^{2(s+1)}(\log(n)) \log^{2(s+1)}(16\eta^{-1}) \ ,$$

for some $C_{\gamma_1,\gamma_2} > 0$ *and some* $C_{s,q} > 0$.

Note that $\eta_n \to 0$ as $n \to \infty$.

Remark 5.3.13. *Still, our choice for* λ_0 *is only a theoretical value which remains unknown as it depends on the unknown marginal* ν *through the effective dimension* $\mathcal{N}(\lambda)$*. Implementation requires a data driven choice. Heuristically, it seems resonable to proceed as follows. Let* $q > 1$ *and* $\tilde{\lambda}_j = q^{-j}$*,* $j = 0, 1, ...$ *(we are starting from the right and reverse the order). Define the stopping index*

$$\hat{j}_0 := \min\{ \ j \in \mathbb{N} : \ \mathcal{S}_{\mathbf{x}}(n, \tilde{\lambda}_j) \geq 5 \ \}$$

and let $\Lambda = \{\tilde{\lambda}_{\hat{j}_0} < ... < \tilde{\lambda}_0 = 1\}$*. Here,* $\mathcal{S}_{\mathbf{x}}(n, \tilde{\lambda}_j)$ *depends on the empirical effective dimension* $\mathcal{N}_{\mathbf{x}}(\lambda)$*, see* (5.2.1)*, which by Corollary 5.2.2 is close to the unknown effective dimension* $\mathcal{N}(\lambda)$*. Thus we think that the above choice of* λ_0 *is reasonable for implementing the dependence of* λ_0 *on the unknown marginal. A complete mathematical analysis is in development.*

5.4 Applications

We proceed by illustrating some applications of our method as described in the previous section. In view of our Theorem 5.3.11 and Corollary 5.3.12 it suffices to only consider balancing in $L^2(\nu)$. We always choose a geometric grid as in Remark 5.3.5, satisfying $\lambda_m = 1$. We shall treat fast rates for the regular case for Hölder type source conditions and general source conditions, and our class beyond the regular

case for Hölder type source conditions. This is achieved by choosing appropriate parameter spaces in our general theory above and using the estimates from Chapters 2 and 3.

Example 1: The regular case

We consider the setting of Chapter 2, where the eigenvalues of $\bar{B}$ decay polynomially (with parameter $b > 1$), the target function f_ρ satisfies a Hölder-type source condition

$$f_\rho \in \Omega_\nu(r, R) := \{ f \in \mathcal{H}_1 \ : \ f = \bar{B}_\nu^r h \ , \ ||h||_{\mathcal{H}_1} \leq R \}$$

and the noise satisfies a Bernstein-Assumption

$$\mathbb{E}[\left| Y - \bar{S}_X f_\rho \right|^m \ | \ X] \leq \frac{1}{2} m! \ \sigma^2 M^{m-2} \quad \nu - \text{a.s.} \ , \tag{5.4.1}$$

for any integer $m \geq 2$ and for some $\sigma > 0$ and $M > 0$. We combine all structural parameters in a vector (γ, θ), with $\gamma = (M, \sigma, R) \in \Gamma = \mathbb{R}_+^3$ and $\theta = (r, b) \in \Theta = (0, \infty) \times (1, \infty)$. We are interested in adaptivity over Θ. We warn the reader that (for purely historic reasons in writing this thesis) the new definition of γ is the old definition of θ in Chapter 2-4, while the dependence on the new parameter θ was suppressed in the notation of the previous chapters (where this parameter was kept fixed and varying it was no issue). We hope that this will not terribly confuse the reader.

Under the assumptions of Corollary 2.3.6 , the sequence of estimators $(f_{\mathbf{z}}^{\lambda_{n,\gamma,\theta}})_n$ as defined in (2.2.15) is minimax optimal for any $\theta \in \Theta$ over the class $\mathcal{M}_{(\gamma,\theta)} := \mathcal{M}(r, R, \mathcal{P}')$ defined precisely in (2.2.9), with $\mathcal{P}' = \mathcal{P}^<(b, \beta) \cap \mathcal{P}^>(b, \alpha)$, where $\beta \geq \alpha > 0$ and

$$\mathcal{P}^<(b, \beta) := \{\nu \in \mathcal{P} : \ \mu_j \leq \beta / j^b \ \ \forall j \geq 1\} \ , \tag{5.4.2}$$

$$\mathcal{P}^>(b, \alpha) := \{\nu \in \mathcal{P} : \ \mu_j \geq \alpha / j^b \ \ \forall j \geq 1\} \ , \tag{5.4.3}$$

defined in (2.2.6) and (2.2.7). The corresponding minimax optimal rate is given by

$$a_n = a_{n,\gamma,\theta} = R \lambda_{n,\gamma,\theta}^{r+s} = R \left(\frac{\sigma^2}{R^2 n} \right)^{\frac{b(r+s)}{2br+b+1}} \ .$$

We shall now check validity of our Assumption 5.3.2. In the following, we assume that the data generating distribution belongs to the class $\mathcal{M} = \mathcal{M}_{(\gamma,\theta)}$. Recall that we let $\lambda_0(n)$ be determined as the unique solution of $\mathcal{N}(\lambda) = n\lambda$. Then, combining the estimates (2.5.6) and (2.5.14) with the new inequality (A.2.4), we have uniformly for all data generating distributions from the class $\mathcal{M}$, with probability at least $1 - \eta$, for any $\lambda \in \Lambda_m$,

$$||(\bar{B}_{\mathbf{x}} + \lambda)^s (f_{\mathbf{z}}^\lambda - f_\rho)||_{\mathcal{H}_1} \leq C_s \log^2(8|\Lambda_m|\eta^{-1}) \ \lambda^s \left(\ \tilde{\mathcal{A}}(\lambda) + \tilde{\mathcal{S}}(n, \lambda) \ \right) \ ,$$

for n sufficiently large, with

$$\tilde{\mathcal{A}}(\lambda) = R\lambda^r + \frac{Rr}{\sqrt{n}} 1_{(1,\infty)}(r) \ , \quad \tilde{\mathcal{S}}(n, \lambda) = \sigma \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}} + \frac{M}{n\lambda} \ ,$$

where C_s does not depend on the parameters $(\gamma, \theta) \in \Gamma \times \Theta$. Remember that the optimal choice for the

regularization parameter λ_n is obtained by solving

$$\mathcal{A}(\lambda) = \sigma\sqrt{\frac{\lambda^{-1/b}}{n\lambda}}$$

and belongs to the interval $[\lambda_0(n), 1]$. This can be seen by the following argument: If n is sufficiently large

$$1 = \sqrt{\frac{\mathcal{N}(\lambda_0(n))}{n\lambda_0(n)}} \geq \sqrt{C_{\beta,b}}R\lambda_n^r = \sigma\sqrt{\frac{C_{\beta,b}\lambda_n^{-\frac{1}{b}}}{n\lambda_n}} \geq \sqrt{\frac{\mathcal{N}(\lambda_n)}{n\lambda_n}} ,$$

which is equivalent to $\mathcal{S}(n, \lambda_0(n)) \geq \mathcal{S}(n, \lambda_n)$. Since $\lambda \mapsto \mathcal{S}(n, \lambda)$ is strictly decreasing we conclude $\lambda_n \geq \lambda_0(n)$. Here we use the bound $\mathcal{N}(\lambda) \leq C_{\beta,b}\lambda^{-\frac{1}{b}}$.

Recall that we also have corresponding lower bound $\mathcal{N}(\lambda) \geq C_{\alpha,b}\lambda^{-\frac{1}{b}}$, since $\nu \in \mathcal{P}^{>}(b, \alpha)$, granting Assumption 5.3.6. This follows by combining Remark 3.1.1 with Lemma 3.4.3 and Lemma 3.4.4.

We adaptively choose the regularization parameter $\hat{\lambda}_{1/2}(\mathbf{z})$ according to Definition 5.3.3 by $L^2(\nu)-$ balancing (i.e. by choosing $s = \frac{1}{2}$) and independently from the parameters $b > 1$, $r > 0$. Corollary 5.3.12 gives for any $s \in [0, \frac{1}{2}]$, if n is sufficiently large, with probability at least $1 - \eta$ (uniformly over $\mathcal{M}$)

$$\left\| \bar{B}^s (f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})} - f_\rho) \right\|_{\mathcal{H}_1} \leq C'_{s,q} C_s(\eta) \ (\ a_n + \lambda_n^s d(n, \lambda_n) \ ) \ , \tag{5.4.4}$$

where

$$C_s(\eta) = \log^{2(s+1)}(\log(n)) \log^{2(s+1)}(16\eta^{-1}),$$

provided that $\eta \geq \eta_n = 4C_q \log(n) \exp\left(-Cn^{\frac{1}{2(b+1)}}\right)$, for some $C > 0$, depending on α, β and b. Recall that $\eta_n \to 0$ as $n \to \infty$.

In (5.4.4) we have used that

$$\begin{aligned} \min_{\lambda \in [\lambda_0(n), 1]} \{ \ \lambda^s (\tilde{\mathcal{A}}(\lambda) + \tilde{\mathcal{S}}(n, \lambda)) \ \} &\leq \lambda_n^s (\tilde{\mathcal{A}}(\lambda_n) + \tilde{\mathcal{S}}(n, \lambda_n)) \\ &= \lambda_n^s (\mathcal{A}(\lambda_n) + \mathcal{S}(n, \lambda_n) + d(n, \lambda_n)) \ . \end{aligned}$$

Then $\lambda_n^s \mathcal{A}(\lambda_n) \leq a_n$ and $\lambda_n^s \mathcal{S}(n, \lambda_n) \leq C_b a_n$ give equation (5.4.4).

It remains to show that for n sufficiently large, the remainder $\lambda_n^s d(n, \lambda_n)$ is of lower order than the rate a_n. This follows exactly by argumenting as in Section 2.5. One finds that

$$\frac{M}{n\lambda_n} = o\left(C_b \sqrt{\frac{1}{n}\lambda_n^{-\frac{b+1}{b}}} \right) , \quad \frac{r}{\sqrt{n}} = o(\lambda_n^r) \ .$$

Summarizing the above findings gives

Corollary 5.4.1 (from Corollary 5.3.12)**.** *Let $s \in [0, \frac{1}{2}]$. Choose the regularization parameter $\hat{\lambda}_{1/2}(\mathbf{z}) = \hat{\lambda}_{n,\gamma,\eta}(\mathbf{z})$ according to Definition 5.3.3 by choosing $s = \frac{1}{2}$. Then, if n is sufficiently large, for any*

$$\eta \geq \eta_n = 4C_q \log(n) \exp\left(-Cn^{\frac{1}{2(b+1)}}\right) ,$$

$(r,b) \in \mathbb{R}_+ \times (1,\infty)$, $(M,\sigma,R) \in \mathbb{R}_+^3$

$$\sup_{\rho \in \mathcal{M}} \rho^{\otimes n} \left(\left\| \bar{B}^s (f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})} - f_\rho) \right\|_{\mathcal{H}_1} \leq C'_{s,q} \log^{2(s+1)}(16\eta^{-1})\, b_n \right) \geq 1 - \eta \,,$$

with $b_n = \log^{2(s+1)}(\log(n))\, a_n$.

Now defining $\tau = C'_{s,q} \log^{2(s+1)}(16\eta^{-1})$ gives

$$\eta = 16 \exp\left(- \left(\frac{\tau}{C'_{s,q}} \right)^{1/2(s+1)} \right) \,,$$

implying (5.3.3).

Observing that the proof of Theorem 2.3.5 and Corollary 2.3.6 implies validity of the lower bound (5.3.4), this means:

Corollary 5.4.2. *In the sense of Definition 5.3.1 the sequence of estimators* $(f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})})_{n\in\mathbb{N}} = (f_{\mathbf{z}}^{\hat{\lambda}_{n,\gamma,\eta}(\mathbf{z})})_{n\in\mathbb{N}}$ *is adaptive over* Θ *(up to log-term) and the model family* $(\mathcal{M}_{(\gamma,\theta)})_{(\gamma,\theta)\in\Gamma\times\Theta}$ *with respect to the family of rates* $(a_{n,(\gamma,\theta)})_{(n,\gamma)\in\mathbb{N}\times\Gamma}$*, for all interpolation norms of parameter* $s \in [0, \frac{1}{2}]$.

Example 2: General Source Condition, polynomial decay of eigenvalues

Our approach also applies to the case where the smoothness is measured in terms of a *general source condition*, generated by some index function, that is,

$$f_\rho \in \Omega_\nu(\mathcal{A}) := \{\, f \in \mathcal{H}_1 :\ f = \mathcal{A}(\bar{B}_\nu)h,\ ||h||_{\mathcal{H}_1} \leq 1 \,\} \,,$$

where $\mathcal{A} : (0,1] \longrightarrow \mathbb{R}_+$ is a continuous non-decreasing function, satisfying $\lim_{t\to 0} \mathcal{A}(t) = 0$. We keep the noise condition (5.4.1) and we choose the parameter $\gamma = (M,\sigma) \in \Gamma = \mathbb{R}_+^2$, $\theta = (\mathcal{A}, b) \in \Theta = \mathcal{F} \times (1,\infty)$, where $\mathcal{F}$ denotes either the class of *operator monotone* functions or the class of functions decomposing into an operator monotone part and an *operator Lipschitz* part. For more details, we refer the interested reader to [4], [62].

We introduce the class of data-generating distributions

$$\mathcal{M}^{<}_{(\gamma,\theta)} = \{\rho(dx,dy) = \rho(dy|x)\nu(dx); \rho(\cdot|\cdot) \in \mathcal{K}(\Omega_\nu(\mathcal{A})), \nu \in \mathcal{P}^{<}(b,\beta)\} \,,$$
$$\mathcal{M}^{>}_{(\gamma,\theta)} = \{\rho(dx,dy) = \rho(dy|x)\nu(dx); \rho(\cdot|\cdot) \in \mathcal{K}(\Omega_\nu(\mathcal{A})), \nu \in \mathcal{P}^{>}(b,\alpha)\} \,,$$

where $\mathcal{P}^{<}(b,\beta)$ and $\mathcal{P}^{>}(b,\alpha)$ are defined in (2.2.6) and (2.2.7), respectively. Then $\mathcal{M} = \mathcal{M}_{(\gamma,\theta)}$ is defined as the intersection.

From [73] and [62] (in particular Proposition 4.3), combined with our Proposition A.1.5 one then gets that Assumption 5.3.2 is satisfied: Uniformly for all data generating distributions from the class $\mathcal{M}$, with probability at least $1-\eta$,

$$||(\bar{B}_{\mathbf{x}} + \lambda)^s (f_{\mathbf{z}}^\lambda - f_\rho)||_{\mathcal{H}_1} \leq C_s \log^2(8|\Lambda_m|\eta^{-1})\, \lambda^s \left(\tilde{\mathcal{A}}(\lambda) + \tilde{\mathcal{S}}(n,\lambda) \right) \,,$$

for n sufficiently large, with

$$\tilde{\mathcal{A}}(\lambda) = \mathcal{A}(\lambda) + \frac{C}{\sqrt{n}}\,, \quad \tilde{\mathcal{S}}(n,\lambda) = \sigma\sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}} + \frac{M}{n\lambda}$$

and

$$d(n,\lambda_n) = \frac{C}{\sqrt{n}} + \frac{M}{n\lambda}\,.$$

Assuming $\mathcal{N}(\lambda) \leq C_{\beta,b}\lambda^{-1/b}$, which as above is implied by polynomial asymptotics of the eigenvalues of the covariance operator $\bar{B}$ specified by the exponent b, the sequence of estimators $(f_z^{\lambda_{n,\mathcal{A},b}})_n$ (defined via some spectral regularization having prescribed qualification) using the parameter choice

$$\lambda_n := \lambda_{n,\mathcal{A},b} := \psi_{\mathcal{A},b}^{-1}\left(\frac{1}{\sqrt{n}}\right)\,, \qquad \psi_{\mathcal{A},b}(t) := \mathcal{A}(t)t^{\frac{1}{2}\left(\frac{1}{b}+1\right)}\,, \tag{5.4.5}$$

is then minimax optimal, in both $\mathcal{H}_1-$norm ($s=0$) and $L^2(\nu)-$norm ($s=1/2$) (see [73], [62]), with rate

$$a_n := a_{n,\mathcal{A},b} := \lambda_{n,\mathcal{A},b}^s\,\mathcal{A}\left(\lambda_{n,\mathcal{A},b}\right)\,. \tag{5.4.6}$$

This holds pointwisely for any $(\mathcal{A},b) \in \Theta = \mathcal{F}\times(1,\infty)$. The crucial observation is that equation (5.4.6) is precisely the result obtained by balancing the leading order terms for sample and approximation error.

Arguments similar to those in the previous example show that $\lambda_n \in [\lambda_0(n),1]$. Recall that $\mathcal{N}(\lambda) \leq C_{\beta,b}\lambda^{-\frac{1}{b}}$ and that $\mathcal{A}(\lambda)\to 0$ as $\lambda\to 0$. Thus, if n is big enough

$$1 = \sqrt{\frac{\mathcal{N}(\lambda_0(n))}{n\lambda_0(n)}} \geq \sqrt{C_{\beta,b}}\,\mathcal{A}(\lambda_n) = \sqrt{C_{\beta,b}}\psi(\lambda_n)\lambda_n^{-\frac{1}{2}\left(\frac{1}{b}+1\right)} \geq \sqrt{\frac{\mathcal{N}(\lambda_n)}{n\lambda_n}}\,,$$

which is equivalent to $\mathcal{S}(n,\lambda_0(n)) \geq \mathcal{S}(n,\lambda_n)$. Since $\lambda \mapsto \mathcal{S}(n,\lambda)$ is strictly decreasing, we conclude that $\lambda_n \geq \lambda_0(n)$.

Recall that we also have corresponding lower bound $\mathcal{N}(\lambda) \geq C_{\alpha,b}\lambda^{-\frac{1}{b}}$, since $\nu \in \mathcal{P}^{>}(b,\alpha)$, granting Assumption 5.3.6. This follows by combining Remark 3.1.1 with Lemma 3.4.3 and Lemma 3.4.4.

We again adaptively choose the regularization parameter $\hat{\lambda}_{1/2}(\mathbf{z})$ according to Definition 5.3.3 by $L^2(\nu)-$balancing (i.e. by choosing $s=\frac{1}{2}$) and independently from the parameters $b>1$, $r>0$. Corollary 5.3.12 gives for any $s\in[0,\frac{1}{2}]$, if n is sufficiently large, with probability at least $1-\eta$ (uniformly over $\mathcal{M}$)

$$\left\|\bar{B}^s(f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})} - f_\rho)\right\|_{\mathcal{H}_1} \leq\; C'_{s,q}C_s(\eta)\;\left(\,a_n + \lambda_n^s d(n,\lambda_n)\,\right)\,, \tag{5.4.7}$$

where

$$C_s(\eta) = \log^{2(s+1)}(\log(n))\log^{2(s+1)}(16\eta^{-1})\,,$$

provided that

$$\eta \geq \eta_n = 4C_q\log(n)\exp\left(-Cn^{\frac{1}{2(b+1)}}\right)\,,$$

for some $C>0$, depending on α,β and b.

One readily verifies also in this case that the remainder term $d(n, \lambda_n)$ is indeed subleading:

$$n^{-1/2} = \psi_{\mathcal{A},b}(\lambda_n) = \lambda_n^{\frac{1}{2}(1+\frac{1}{b})}\mathcal{A}(\lambda_n) = o\left(\mathcal{A}(\lambda_n)\right),$$

and moreover

$$\frac{M}{n\lambda_n} = o\left(C_b\sqrt{\frac{1}{n}\lambda_n^{-\frac{b+1}{b}}}\right) .$$

From Theorem 3.12 in [73] one then obtains the lower bound (5.3.4).

Thus, we have proved:

Corollary 5.4.3 (from Corollary 5.3.12)**.** *Let $s \in [0, \frac{1}{2}]$. Choose the regularization parameter $\hat{\lambda}_{1/2}(\mathbf{z}) = \lambda_{n,\gamma,\eta}(\mathbf{z})$ according to Definition 5.3.3 by $L^2(\nu)-$ balancing. Then, if n is sufficiently large, for any*

$$\eta \geq \eta_n = 4C_q \log(n) \exp\left(-Cn^{\frac{1}{2(b+1)}}\right) ,$$

$\mathcal{A} \in \mathcal{F}$, $b > 1$ and $(M, \sigma, R) \in \mathbb{R}^3_+$ one has

$$\sup_{\rho\in\mathcal{M}_{(\gamma,\theta)}} \rho^{\otimes n}\left(\left\|\bar{B}^s(f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})} - f_\rho)\right\|_{\mathcal{H}_1} \leq C'_{s,q}\log^{2(s+1)}(16\eta^{-1})\; b_n\right) \geq 1-\eta ,$$

with

$$b_n = \log^{2(s+1)}(\log(n))\; a_n .$$

This means that in the sense of Definition 5.3.1 the sequence of estimators $(f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})})_{n\in\mathbb{N}} = (f_{\mathbf{z}}^{\hat{\lambda}_{n,\gamma,\eta}(\mathbf{z})})_{n\in\mathbb{N}}$ is adaptive over Θ (up to log-term) and the model family $(\mathcal{M}_{(\gamma,\theta)})_{(\gamma,\theta)\in\Gamma\times\Theta}$ with respect to the family of rates $(a_{n,\gamma,\theta})_{(n,\gamma)\in\mathbb{N}\times\Gamma}$ from (5.4.6), for all interpolation norms of parameter $s \in [0, \frac{1}{2}]$.

Example 3: Beyond the regular case

Recall the class of models considered in Chapter 3 : Let $\gamma = (M, \sigma, R) \in \Gamma = \mathbb{R}^3_+$, $\Theta = \{(r, \nu^*, \nu_*) \in \mathbb{R}_+ \times (1,\infty)^2; \nu^* \leq \nu_*\}$ and set

$$\mathcal{M}^<_{(\gamma,\theta)} := \{\, \rho(dx,dy) = \rho(dy|x)\nu(dx) \;:\; \rho(\cdot|\cdot) \in \mathcal{K}(\Omega_\nu(r,R)),\; \nu \in \mathcal{P}^<(\nu^*) \,\} , \tag{5.4.8}$$

$$\mathcal{M}^>_{(\gamma,\theta)} := \{\, \rho(dx,dy) = \rho(dy|x)\nu(dx) \;:\; \rho(\cdot|\cdot) \in \mathcal{K}(\Omega_\nu(r,R)),\; \nu \in \mathcal{P}^>(\nu_*) \,\} , \tag{5.4.9}$$

and denote by $\mathcal{M} = \mathcal{M}_{(\gamma,\theta)}$ the intersection.

We shall verify validity of our Assumption 5.3.2. In the following, we assume that the data generating distribution belongs to the class $\mathcal{M}$. Then, combining the bound (3.4.4) with the new inequality (A.2.4), we have uniformly for all data generating distributions from the class $\mathcal{M}$, with probability at least $1-\eta$, for any $\lambda \in \Lambda_m$,

$$||(\bar{B}_{\mathbf{x}} + \lambda)^s(f_{\mathbf{z}}^\lambda - f_\rho)||_{\mathcal{H}_1} \leq C_{s,\nu^*}\log^2(8|\Lambda_m|\eta^{-1})\;\lambda^s\left(\tilde{\mathcal{A}}(\lambda) + \tilde{\mathcal{S}}(n,\lambda)\right) ,$$

with

$$\tilde{\mathcal{A}}(\lambda) = R\lambda^r + \frac{Rr}{\sqrt{n}} 1_{(1,\infty)}(r) \ , \quad \tilde{\mathcal{S}}(n,\lambda) = \sigma \sqrt{\frac{\lambda_n^{2r}}{n\mathcal{G}(\lambda)}} + \frac{M}{n\lambda} \ .$$

As usual, we shall investigate adaptivity on the parameter space Θ.

We upper bound the effective dimension by applying Lemma 3.4.2, using the counting function $\mathcal{F}(\lambda)$ defined in equation (3.1.1). We obtain

$$\mathcal{N}(\lambda) \leq C_{\nu^*} \mathcal{F}(\lambda) \ ,$$

for any λ sufficiently small. We now follow the discussion in Example 1 above, with $\mathcal{A}(\lambda)$, $\mathcal{S}(n,\lambda)$, $d_1(n)$, $d_2(n,\lambda)$ remaining unchanged. We shall only use the new upper bound on $\mathcal{S}(n,\lambda)$ defined by

$$\mathcal{S}_+(n,\lambda) = \sigma \sqrt{\frac{\mathcal{F}(\lambda)}{n\lambda}} = \sigma \sqrt{\frac{\lambda^{2r}}{n\mathcal{G}(\lambda)}} \ .$$

This gives, equating $R\lambda^r = \mathcal{S}_+(n,\lambda)$, for n sufficiently large (see equation (3.2.3))

$$\lambda_n = \lambda_{n,\theta} = \mathcal{G}^{-1}\left(\frac{\sigma^2}{R^2 n}\right) \ .$$

Also in this case, λ_n can shown to fall in the interval $[\lambda_0(n), 1]$. Indeed, if n is sufficiently large

$$1 = \sqrt{\frac{\mathcal{N}(\lambda_0(n))}{n\lambda_0(n)}} \geq \sqrt{C_{\nu^*}} R\lambda_n^r = \sqrt{C_{\nu^*}}\sigma \sqrt{\frac{\mathcal{F}(\lambda_n)}{n\lambda_n}} \geq \sigma \sqrt{\frac{\mathcal{N}(\lambda_n)}{n\lambda_n}} \ ,$$

which is equivalent to $\mathcal{S}(n,\lambda_0(n)) \geq \mathcal{S}(n,\lambda_n)$. Since $\lambda \mapsto \mathcal{S}(n,\lambda)$ is strictly decreasing, we have $\lambda_0(n) \leq \lambda_n$, provided n is big enough.

More refined bounds for the effective dimension follow from Lemma 3.4.4 and Lemma 3.4.5. We have

$$C_{\nu_*} \lambda^{-\frac{1}{\nu_*}} \leq \mathcal{N}(\lambda) \leq C_{\nu^*} \lambda^{-\frac{1}{\nu^*}}$$

and Assumption 5.3.6 is satisfied.

We adaptively choose the regularization parameter $\hat{\lambda}_{1/2}(\mathbf{z})$ according to Definition 5.3.3 by $L^2(\nu)-$ balancing, i.e. by choosing $s = \frac{1}{2}$. Corollary 5.3.12 gives for any $s \in [0, \frac{1}{2}]$, if n is sufficiently large, with probability at least $1-\eta$ (uniformly over $\mathcal{M}$)

$$\left\| \bar{B}^s (f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})} - f_\rho) \right\|_{\mathcal{H}_1} \leq \ C'_{s,q} C_s(\eta) \ \left(\ a_n + \lambda_n^s d(n,\lambda_n) \ \right) \ , \tag{5.4.10}$$

where

$$C_s(\eta) = \log^{2(s+1)}(\log(n)) \log^{2(s+1)}(16\eta^{-1}),$$

provided that

$$\eta \geq \eta_n = 4C_q \log(n) \exp\left(-C_{\nu_*,\nu^*} n^{\frac{\nu^*}{2\nu_*(1+\nu^*)}}\right) \ .$$

In (5.4.10) we have used that $a_n = \lambda_n^{r+s}$ and

$$\min_{\lambda\in[\lambda_0(n),1]}\{\ \lambda^s(\tilde{\mathcal{A}}(\lambda)+\tilde{\mathcal{S}}(n,\lambda))\ \} \le \lambda_n^s(\tilde{\mathcal{A}}(\lambda_n)+\tilde{\mathcal{S}}(n,\lambda_n)) \\ = \lambda_n^s(\mathcal{A}(\lambda_n)+\mathcal{S}(n,\lambda_n)+d(n,\lambda_n))\ .$$

As above, one readily checks that that the subleading term $d(n,\lambda_n)$ is really subleading:

$$n^{-\frac{1}{2}} = o(\lambda_n^r)\ , \quad \frac{M}{n\lambda_n} = o\left(\sqrt{\frac{\lambda_n^{2r}}{n\mathcal{G}(\lambda_n)}}\right)\ .$$

Summarizing, we have proved

Corollary 5.4.4 (from Corollary 5.3.12)**.** *Let $s\in[0,\frac{1}{2}]$. Choose the regularization parameter $\hat{\lambda}_{1/2}(\mathbf{z}) = \lambda_{n,\gamma,\eta}(\mathbf{z})$ according to Definition 5.3.3 by choosing $s=\frac{1}{2}$. Then, if n is sufficiently large, for any*

$$\eta \ge \eta_n = 4C_q\log(n)\exp\left(-C_{\nu_*,\nu^*}n^{\frac{\nu^*}{2\nu_*(1+\nu^*)}}\right)\ .$$

for any $r>0$, $1<\nu^*\le\nu_*$, $(M,\sigma,R)\in\mathbb{R}_+^3$, one has

$$\sup_{\rho\in\mathcal{M}_{(\gamma,\theta)}}\rho^{\otimes n}\left(\left\|\bar{B}^s(f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})}-f_\rho)\right\|_{\mathcal{H}_1}\le C'_{s,q}\log^{2(s+1)}(16\eta^{-1})\ b_n\right)\ge 1-\eta\ ,$$

with

$$b_n = \log^{2(s+1)}(\log(n))\ a_n\ .$$

Observing that the proof of Theorem 3.2.2 and Corollary 3.2.3 implies validity of the lower bound (5.3.4)*, this means that in the sense of Definition 5.3.1 the sequence of estimators $(f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})})_{n\in\mathbb{N}} = (f_{\mathbf{z}}^{\hat{\lambda}_{n,\gamma,\eta}(\mathbf{z})})_{n\in\mathbb{N}}$ is adaptive over Θ (up to log-term) and the model family $(\mathcal{M}_{(\gamma,\theta)})_{(\gamma,\theta)\in\Gamma\times\Theta}$ with respect to the family of rates $(a_{n,\gamma,\theta})_{(n,\gamma)\in\mathbb{N}\times\Gamma}$, for all interpolation norms of parameter $s\in[0,\frac{1}{2}]$.*

5.5 Discussion

1. We have shown that it suffices to prove adaptivity only in $L^2(\nu)-$norm, which is the weakest of all our interpolating norms indexed by $s\in[0,1/2]$. Similar results of this type (an estimate in a weak norm suffices to establish the estimate in a stronger norm) have been obtained e.g. in [11] and also in the recent paper of Lepskii, see [54], in a much more general context.

2. We shall briefly discuss where and how the presentation of the balancing principle in our work improves the results in the existing literature on the subject. We recall from the introduction to this chapter that the first paper on the balancing principle for kernel methods, [25], did not yet introduce *fast rates*, i.e. rates depending on the intrinsic dimensionality b. Within this framework the results give - in the wording of the authors - *an optimal adaptive choice of the regularization parameter for the class of spectral regularization methods* as defined in Chapter 2. In the sense of our Definition 5.3.1 the obtained estimators are optimal adaptive - with hindsight, as amplified in our introduction - on the parameter space $\Theta=\mathbb{R}_+$ with respect to minimax optimal rates, which depend on r but not on b (or more general, not on the effective dimension $\mathcal{N}(\lambda)$). Technically, the authors

of [25] define their optimal adaptive estimator as the minimum of 2 estimators, corresponding to 2 different norms, namely, setting

$$\mathcal{J}^+_{\mathbf{z}}(\Lambda_m) \;=\; \left\{\lambda_i \in \Lambda_m \;:\; \left\|\bar{B}^s_{\mathbf{x}}(f^{\lambda_i}_{\mathbf{z}} - f^{\lambda_j}_{\mathbf{z}})\right\|_{\mathcal{H}_1} \;\leq\; 4C_s(\eta)\;\lambda_j^s\;\mathcal{S}(n,\lambda_j)\;,\; j=0,...,i-1\right\}$$

and defining $\tilde{\lambda}_s(\mathbf{z}) := \max\; \mathcal{J}^+_{\mathbf{z}}(\Lambda_m)$, their final estimator is given by

$$\hat{\lambda}_s(\mathbf{z}) := \min\{\tilde{\lambda}_s(\mathbf{z}), \tilde{\lambda}_0(\mathbf{z})\}\;. \tag{5.5.1}$$

We encourage the reader to directly compare this definition with our definition in (5.3.11). Using the minimum of two estimators in this way can be traced back to the use of an additive error estimate of the form

$$\left|\left\|\bar{B}^s f\right\|_{\mathcal{H}_1} - \left\|\bar{B}^s_{\mathbf{x}} f\right\|_{\mathcal{H}_1}\right| \;\leq\; \sqrt{6}\log(4/\eta)\;n^{-\frac{s}{2}}\;\|f\|_{\mathcal{H}_1}\;, \tag{5.5.2}$$

holding for any $f \in \mathcal{H}_1$, $s \in [0,1/2]$ and $\eta \in (0,1)$, with probability at least $1-\eta$. Here we have slightly generalized the original estimate in [25] to all values of $s \in [0,1/2]$.

In the setting of [25], where only slow rates are considered, the variance $\mathcal{S}(n,\lambda)$ is fully known. However, when considering fast rates (polynomial decay of eigenvalues), $\mathcal{S}(n,\lambda)$ additionally depends on the unknown parameter $b > 1$ and we have to replace the variance by its empirical approximation $\mathcal{S}_{\mathbf{x}}(n,\lambda)$. This can effectively achieved by our Corollary 5.2.2, where we provide a two sided bound

$$\frac{1}{5}\;\mathcal{S}_{\mathbf{x}}(n,\lambda) \leq \mathcal{S}(n,\lambda) \leq 5\;\mathcal{S}_{\mathbf{x}}(n,\lambda)\;.$$

Our bound (in a slightly weaker form) is also used in [62] for bounding the variance by its empirical approximation.

In the preprint [62] the authors independently present the balancing principle for fast rates. More precisely, in the case of Hölder-type source conditions, it covers the range Θ_{hs} of parameters (r,b) of *high smoothness* where $b > 1$ and $r \geq 1/2(1-1/b)$, which excludes the region of *low smoothness*. In addition, their results include more general types of source conditions. This work started independently from our work on the balancing principle. A crucial technical difference is that [62] is still based on using (5.5.2) in an essential way. This paper contains the new multiplicative error estimate of [42] (see Appendix A.2), which leads to Proposition A.2.1 and Corollary A.2.2. Both are crucial to extend the definition of the adaptive estimator to all values of the confidence level $\eta \in (0,1)$. However, the discussion proceeds essentially along the traditional lines of [25], using the above mentioned additive error estimates. This makes the region of low smoothness, i.e. $r < 1/2(1-1/b)$, much less accessible and leads to an estimator obtained by balancing only on the restricted parameter space Θ_{hs} (with respect to minimax optimal rates of convergence, which, however, are known on the larger parameter space $\Theta = \mathbb{R}_+ \times (0,\infty)$). As before, the final estimator is taken to be a minimum of 2 estimators corresponding to different norms.

Our approach also exploits the technical improvement contained in the new multiplicative error estimate which simplifies the derivation of probabilistic error estimates on the full range $\eta \in (0,1)$ of the confidence level. Furthermore, our modified definition of the estimator defined by balancing, avoiding the additive error estimate in equation (5.5.2), allows in the case of Hölder type source conditions to obtain an optimal adaptive estimator (up to $\log\log(n)$ term) on the parameter space

$\Theta = \mathbb{R}_+ \times (1, \infty)$. The final estimator is constructed somewhat more directly. It is not taken as a minimum of 2 separately constructed estimators and in our view the presentation is streamlined. Furthermore, our discussion in Example 5.4 shows how the more general results of [62] on source conditions different from Hölder -type can naturally be recovered in our approach.

3. Finally we want to emphasize that this notion of optimal adaptivity is *not* quite the original approach of Lepskii. The paper [8] contains an approach to the optimal adaptivity problem in the white noise framework which is closer to the original Lepskii approach and thus somewhat stronger than the weak approach described above, where the optimal adaptive estimator depends on the confidence level. It seems to be a wide open question how to adapt this original approach to the framework of kernel methods, i.e. constructing an estimator which is *optimal adaptive in Lepskii-sense* (independent of the confidence level η) and satisfies

$$\sup_{\theta\in\Theta} \sup_{\gamma\in\Gamma} \limsup_{n\to\infty} a_{n,(\gamma,\theta)}^{-1} \, R_n(\tilde{f}^{\lambda_{n,\gamma}(\mathbf{z})}, \gamma) \; < \; \infty \,, \tag{5.5.3}$$

with R_n being the risk

$$R_n(\tilde{f}^{\lambda_{n,(\gamma,\theta)}(\mathbf{z})}, \gamma) = \sup_{\rho\in\mathcal{M}_{(\gamma,\theta)}} \mathbb{E}_{\rho^{\otimes n}} \left[\|\bar{B}^s (f_\rho - \tilde{f}^{\lambda_{n,\gamma}(\mathbf{z})})\|_{\mathcal{H}_1}^p \right]^{\frac{1}{p}} , \quad p > 0 \, , s \in [0, 1/2] \, ,$$

and $a_{n,(\gamma,\theta)}$ being a minimax optimal rate.

Here we always want to take Θ as the maximal parameter space on which one has minimax optimal rates. For slow rates, i.e. $\Theta = \{r > 0\}$, the supremum over Θ in equation (5.5.3) exists. For fast rates, the boundary of the open set $\{b > 1\}$ poses problems at $b = 1$, since one looses the trace class condition on the covariance operator $\bar{B}$ (in which case minimax optimality as in this thesis is not even proved). We remark that, trying to only use the effective dimension and parametrizing it by

$$\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{b}}),$$

(thus redefining somewhat the meaning of b) possibly changes the nature of the boundary at $b = 1$ and might give existence of the sup. We leave this question for future research. Furthermore we remark that a rigorous proof of non-existence of the sup for our (spectral) meaning of b requires a suitable lower bound exploding as $b \downarrow 1$, similar to the example in [55].

A similar type of difficulty (related to the non-existence of the sup) has already been systematically investigated in [55] and [57]. In such a case Lepskii has introduced the weaker notion of *the adaptive minimax order of exactness* and he also discusses additional log terms. Such estimators (which are not optimally adaptive) are called simply *adaptive*. This is related to the situation which we encounter in this section. It is known that e.g. for point estimators, additional log terms are indispensable. Our situation, however, is different and one could expect to prove optimal adaptivity in future research.

5.6 Proofs

Lemma 5.6.1. *For any $s \in [0, \frac{1}{2}]$ and $\eta \in (0, 1]$, with probability at least $1 - \eta$ we have $\lambda_* \leq \hat{\lambda}_s(\mathbf{z})$, provided $2\log(4|\Lambda_m|\eta^{-1}) \leq \sqrt{n\lambda_0}$ and $n\lambda_0 \geq 2$.*

Proof of Lemma 5.6.1. Let $\lambda \in \Lambda_m$ satisfy $\lambda \leq \lambda_*$. We consider the decomposition

$$\left\|(\bar{B}_{\mathbf{x}}+\lambda)^s(f_{\mathbf{z}}^{\lambda}-f_{\mathbf{z}}^{\lambda_*})\right\|_{\mathcal{H}_1} \leq \left\|(\bar{B}_{\mathbf{x}}+\lambda)^s(f_{\mathbf{z}}^{\lambda}-f_{\rho})\right\|_{\mathcal{H}_1} + \left\|(\bar{B}_{\mathbf{x}}+\lambda)^s(f_{\mathbf{z}}^{\lambda_*}-f_{\rho})\right\|_{\mathcal{H}_1} .$$

From Assumption 5.3.2 and since $\lambda \leq \lambda_*$ we have

$$\begin{aligned}\left\|(\bar{B}_{\mathbf{x}}+\lambda)^s(f_{\mathbf{z}}^{\lambda}-f_{\rho})\right\|_{\mathcal{H}_1} &\leq C_s(m,\eta)\ \lambda^s\ (\tilde{\mathcal{A}}(\lambda)+\tilde{\mathcal{S}}(n,\lambda)) \\ &\leq 2C_s(m,\eta)\lambda^s\ \tilde{\mathcal{S}}(n,\lambda)\ ,\end{aligned}$$

with probability at least $1-\eta$.
Since $\lambda \leq \lambda_*$, applying Lemma A.1.6 and Assumption 5.3.2 give, recalling the definition of λ_* and that $\lambda \mapsto \lambda^s\tilde{\mathcal{S}}(n,\lambda)$ is decreasing

$$\begin{aligned}\left\|(\bar{B}_{\mathbf{x}}+\lambda)^s(f_{\mathbf{z}}^{\lambda_*}-f_{\rho})\right\|_{\mathcal{H}_1} &\leq \left\|(\bar{B}_{\mathbf{x}}+\lambda_*)^s(f_{\mathbf{z}}^{\lambda_*}-f_{\rho})\right\|_{\mathcal{H}_1} \\ &\leq C_s(m,\eta)\ \lambda_*^s\ (\tilde{\mathcal{A}}(\lambda_*)+\tilde{\mathcal{S}}(n,\lambda_*)) \\ &\leq 2C_s(m,\eta)\lambda_*^s\ \tilde{\mathcal{S}}(n,\lambda_*) \\ &\leq 2C_s(m,\eta)\lambda^s\ \tilde{\mathcal{S}}(n,\lambda)\ ,\end{aligned}$$

with probability at least $1-\eta$. As a result, using 5.3.9, if $2\log(4|\Lambda_m|\eta^{-1}) \leq \sqrt{n\lambda_0}$ and $n\lambda_0 \geq 2$, with probability at least $1-\eta$

$$\left\|(\bar{B}_{\mathbf{x}}+\lambda)^s(f_{\mathbf{z}}^{\lambda}-f_{\mathbf{z}}^{\lambda_*})\right\|_{\mathcal{H}_1} \leq 20C_s(m,\eta/2)\ \lambda^s\ \tilde{\mathcal{S}}_{\mathbf{x}}(n,\lambda)\ ,$$

with $C_s(m,\eta/2) = C_s\log^2(16|\Lambda_m|\eta^{-1})$. Finally, from the definition (5.3.11) of $\hat{\lambda}_s(\mathbf{z})$ as a maximum, one has $\lambda_* \leq \hat{\lambda}_s(\mathbf{z})$ with probability at least $1-\eta$.

□

Proof of Proposition 5.3.7. Let Assumption 5.3.2 be satisfied. Define λ_* as in (5.3.7). is implied by the sufficient condition We write

$$\left\|(\bar{B}_{\mathbf{x}}+\lambda_*)^s(f_{\mathbf{z}}^{\hat{\lambda}_s(\mathbf{z})}-f_{\rho})\right\|_{\mathcal{H}_1} \leq \left\|(\bar{B}_{\mathbf{x}}+\lambda_*)^s(f_{\mathbf{z}}^{\hat{\lambda}_s(\mathbf{z})}-f_{\mathbf{z}}^{\lambda_*})\right\|_{\mathcal{H}_1} + \left\|(\bar{B}_{\mathbf{x}}+\lambda_*)^s(f_{\mathbf{z}}^{\lambda_*}-f_{\rho})\right\|_{\mathcal{H}_1}$$

and bound each term separately. By definition (5.3.11) of $\hat{\lambda}_s(\mathbf{z})$, by Lemma 5.6.1 and by (5.3.9), with probability at least $1-\frac{\eta}{2}$

$$\begin{aligned}\left\|(\bar{B}_{\mathbf{x}}+\lambda_*)^s(f_{\mathbf{z}}^{\hat{\lambda}_s(\mathbf{z})}-f_{\mathbf{z}}^{\lambda_*})\right\|_{\mathcal{H}_1} &\leq 20C_s(m,\eta/2)\lambda_*^s\tilde{S}_{\mathbf{x}}(n,\lambda_*) \\ &\leq 100C_s(m,\eta/2)\lambda_*^s\tilde{S}(n,\lambda_*)\ .\end{aligned}$$

By Assumption 5.3.2 and recalling the definition of λ_* in (5.3.7) gives for the second term with probability at least $1-\frac{\eta}{2}$

$$\begin{aligned}\left\|(\bar{B}_{\mathbf{x}}+\lambda_*)^s(f_{\mathbf{z}}^{\lambda_*}-f_{\rho})\right\|_{\mathcal{H}_1} &\leq C_s(m,\eta/2)\ \lambda_*^s\ (\ \tilde{\mathcal{A}}(\lambda_*)+\tilde{S}(n,\lambda_*)\ ) \\ &\leq 2C_s(m,\eta/2)\ \lambda_*^s\ \tilde{S}(n,\lambda_*)\ .\end{aligned}$$

The result follows from collecting the previous estimates.

□

Proof of Lemma 5.3.8. Let Assumption 5.3.4, point 1. and 2. be satisfied. We distinguish between the following cases:

Case 1: $\lambda \geq q\lambda_*$
Since $\lambda \to \tilde{\mathcal{A}}(\lambda)$ is increasing and by (5.3.12)

$$\begin{aligned}\lambda^s \ (\tilde{\mathcal{A}}(\lambda) + \tilde{\mathcal{S}}(n,\lambda)) &\geq \lambda^s \ \tilde{\mathcal{A}}(\lambda) \geq (q\lambda_*)^s \ \tilde{\mathcal{A}}(q\lambda_*) \\ &\geq (q\lambda_*)^s \ \tilde{\mathcal{S}}(n, q\lambda_*) \geq q^{s-1}\lambda_*^s \ \tilde{\mathcal{S}}(n,\lambda_*) \ .\end{aligned}$$

Case 2: $\lambda \leq q\lambda_*$
Again, since $\lambda \to \lambda^s \tilde{\mathcal{S}}(n,\lambda)$ is decreasing and by (5.3.12) we have

$$\lambda^s \ (\tilde{\mathcal{A}}(\lambda) + \tilde{\mathcal{S}}(n,\lambda)) \geq \lambda^s \ \tilde{\mathcal{S}}(n,\lambda) \geq (q\lambda_*)^s \ \tilde{\mathcal{S}}(n, q\lambda_*) \geq q^{s-1}\lambda_*^s \ \tilde{\mathcal{S}}(n,\lambda_*) \ .$$

The result follows.

□

Proof of Theorem 5.3.9. Since $\lambda_0(n) \leq \lambda_*$, we may apply estimate (A.2.4) . From Proposition 5.3.7 we have

$$\begin{aligned}\left\| \bar{B}^s (f_\rho - f_{\mathbf{z}}^{\hat{\lambda}_s(\mathbf{z})}) \right\|_{\mathcal{H}_1} &\leq 15 \log^{2s}(4|\Lambda_m|\eta^{-1}) \left\| (\bar{B}_{\mathbf{x}} + \lambda_*)^s (f_\rho - f_{\mathbf{z}}^{\hat{\lambda}_s(\mathbf{z})}) \right\|_{\mathcal{H}_1} \\ &\leq D_s(m,\eta) \ \lambda_*^s \ \tilde{\mathcal{S}}(n,\lambda_*) \ ,\end{aligned}$$

with probability at least $1-\eta$, provided

$$\eta \geq \eta_n := \min\left(1 \ , \ 4|\Lambda_m| \exp\left(-\frac{1}{2}\sqrt{\mathcal{N}(\lambda_0(n))}\right)\right)$$

and where $D_s(m,\eta) = C_s' \log^{2(s+1)}(16|\Lambda_m|\eta^{-1})$. The result follows by applying Lemma 5.3.8 .

□

Proof of Corollary 5.3.10. The proof follows from Theorem 5.3.9, by applying (5.3.13) and by using the lower bound from Assumption 5.3.6. More precisely, the condition

$$\eta \geq \eta_n := \min\left(1 \ , \ 4|\Lambda_m| \exp\left(-\frac{1}{2}\sqrt{\mathcal{N}(\lambda_0(n))}\right)\right)$$

is implied by the sufficient condition

$$\eta \geq \eta_n := \min\left(1, 4C_q \log(n) \exp\left(-\frac{\sqrt{C}_1}{2}\lambda_0(n)^{-\frac{\gamma_1}{2}}\right)\right) \ ,$$

which itself is implied by

$$\eta \geq \eta_n := C_q \log(n) \exp\left(-C_{\gamma_1,\gamma_2} n^{\frac{\gamma_1}{2(1+\gamma_2)}}\right) ,$$

by using (5.3.14), provided n is sufficiently large and with $C_{\gamma_1,\gamma_2} = \frac{\sqrt{C_1}}{2} C_{\gamma_2}^{-\frac{\gamma_1}{2}}$.

Moreover, using $1 \leq \log(16\eta^{-1})$ for any $\eta \in (0,1]$, we obtain

$$\begin{aligned} q^{1-s}\, D_s(m,\eta) &= q^{1-s}\, C'_s \log^{2(s+1)}(16|\Lambda_m|\eta^{-1}) \\ &\leq q^{1-s}\, C'_s \left(\log(C_q \log(n)) + \log(16\eta^{-1})\right)^{2(s+1)} \\ &\leq q^{1-s}\, C'_s \left(\log(C_q \log(n)) + 1\right)^{2(s+1)} \log^{2(s+1)}(16\eta^{-1}) . \end{aligned}$$

Moreover, if n is sufficiently large, we have

$$\log(C_q \log(n)) \leq \log(C_q) + \log(n) \leq (1 + \log(C_q)) \log(n)$$

and thus

$$q^{1-s}\, D_s(m,\eta) \leq C_{s,q} \log^{2(s+1)}(\log(n)) \log^{2(s+1)}(16\eta^{-1}) =: \tilde{D}_{s,q}(n,\eta) ,$$

with $C_{s,q} = q^{1-s}\, C'_s (1 + \log(C_q))^{2(s+1)}$. □

Lemma 5.6.2. *Assume $n\lambda_0 \geq 2$. With probability at least $1 - \eta$*

$$||f_{\mathbf{z}}^{\hat{\lambda}_0(\mathbf{z})} - f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})}||_{\mathcal{H}_1} \;\leq\; D(m,\eta)\, \tilde{\mathcal{S}}(n,\lambda_*) ,$$

provided

$$\eta \geq \eta_n := \min\left(1 \,,\; 4|\Lambda_m| \exp\left(-\frac{1}{2}\sqrt{\mathcal{N}(\lambda_0(n))}\right)\right)$$

and with $D(m,\eta) = 200 \max(C_{1/2}, C_0) \log^2(16|\Lambda_m|\eta^{-1})$.

Proof of Lemma 5.6.2. Recall the definition of λ_* in (5.3.7) and write

$$||f_{\mathbf{z}}^{\hat{\lambda}_0(\mathbf{z})} - f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})}||_{\mathcal{H}_1} \leq ||f_{\mathbf{z}}^{\hat{\lambda}_0(\mathbf{z})} - f_{\mathbf{z}}^{\lambda_*}||_{\mathcal{H}_1} + ||f_{\mathbf{z}}^{\lambda_*} - f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})}||_{\mathcal{H}_1} . \tag{5.6.1}$$

By definition of $\hat{\lambda}_0(\mathbf{z})$, Lemma 5.6.1 and applying (5.3.9) gives with probability at least $1 - \frac{\eta}{2}$

$$\begin{aligned} ||f_{\mathbf{z}}^{\hat{\lambda}_0(\mathbf{z})} - f_{\mathbf{z}}^{\lambda_*}||_{\mathcal{H}_1} &\leq 20 C_0(m,\eta/2) \tilde{\mathcal{S}}_{\mathbf{x}}(n,\lambda_*) \\ &\leq 100 C_0(m,\eta/2) \tilde{\mathcal{S}}(n,\lambda_*) . \end{aligned} \tag{5.6.2}$$

Using $||f||_{\mathcal{H}_1} \leq \lambda_*^{-\frac{1}{2}} ||(\bar{B}_{\mathbf{x}} + \lambda_*)^{\frac{1}{2}} f||_{\mathcal{H}_1}$, Lemma 5.6.1 and the definition of $\hat{\lambda}_{1/2}(\mathbf{z})$ yields with probability

at least $1 - \frac{\eta}{2}$

$$\begin{aligned} ||f_{\mathbf{z}}^{\lambda_*} - f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})}||_{\mathcal{H}_1} &\leq \lambda_*^{-\frac{1}{2}} ||(\bar{B}_{\mathbf{x}} + \lambda_*)^{\frac{1}{2}} (f_{\mathbf{z}}^{\lambda_*} - f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})})||_{\mathcal{H}_1} \\ &\leq 20 C_{1/2}(m, \eta/2) \tilde{\mathcal{S}}_{\mathbf{x}}(n, \lambda_*) \\ &\leq 100 C_{1/2}(m, \eta/2) \tilde{\mathcal{S}}(n, \lambda_*) \ . \end{aligned} \tag{5.6.3}$$

In the last step we applied (5.3.9) once more. Combining (5.6.2) and (5.6.3) with (5.6.1) gives the result.

□

Proof of Theorem 5.3.11. Assume n is sufficiently large and

$$\eta \geq \eta_n = \min\left(1, 4|\Lambda_m| \exp\left(-\frac{1}{2}\sqrt{\mathcal{N}(\lambda_0(n))}\right)\right) \ .$$

Recall that $C_s(m, \eta) = C_s \log^2(8|\Lambda_m|\eta^{-1})$. We firstly show the result for the case where $s = 0$ and get the final one from interpolation. We write

$$||f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})} - f_\rho||_{\mathcal{H}_1} \leq ||f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})} - f_{\mathbf{z}}^{\hat{\lambda}_0(\mathbf{z})}||_{\mathcal{H}_1} + ||f_{\mathbf{z}}^{\hat{\lambda}_0(\mathbf{z})} - f_\rho||_{\mathcal{H}_1}$$

and bound each term separately. From Proposition 5.3.7, with probability at least $1 - \frac{\eta}{2}$

$$||f_{\mathbf{z}}^{\hat{\lambda}_0(\mathbf{z})} - f_\rho||_{\mathcal{H}_1} \leq 102 C_0 \log^2(16|\Lambda_m|\eta^{-1}) \tilde{\mathcal{S}}(n, \lambda_*) \ .$$

Applying Lemma 5.6.2 yields with probability at least $1 - \frac{\eta}{2}$

$$||f_{\mathbf{z}}^{\hat{\lambda}_0(\mathbf{z})} - f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})}||_{\mathcal{H}_1} \leq D(m, \eta) \tilde{\mathcal{S}}(n, \lambda_*) \ ,$$

with $D(m, \eta) = 200 \max(C_0, C_{1/2}) \log^2(16|\Lambda_m|\eta^{-1})$. Collecting both pieces leads to

$$||f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})} - f_\rho||_{\mathcal{H}_1} \ \leq \ D'(m, \eta) \tilde{\mathcal{S}}(n, \lambda_*) \ , \tag{5.6.4}$$

with probability at least $1 - \eta$, where $D'(m, \eta) = C \log^2(16|\Lambda_m|\eta^{-1})$, $C = 302 \max(C_0, C_{1/2})$.

Using $||\bar{B}^s f||_{\mathcal{H}_1} \leq ||\sqrt{\bar{B}} f||_{\mathcal{H}_1}^{2s} \ ||f||_{\mathcal{H}_1}^{1-2s}$ for any $s \in [0, \frac{1}{2}]$, applying (A.2.4), Proposition 5.3.7 and (5.6.4) gives with probability at least $1 - \eta$

$$\begin{aligned} \left\| \bar{B}^s (f_{\mathbf{z}}^{\hat{\lambda}_{1/2}(\mathbf{z})} - f_\rho) \right\|_{\mathcal{H}_1} &\leq \tilde{C}^{2s} \left(\log^3(16|\Lambda_m|\eta^{-1}) \ \sqrt{\lambda_*} \ \tilde{\mathcal{S}}(n, \lambda_*) \right)^{2s} \\ &\qquad C^{1-2s} \ \left(\log^2(16|\Lambda_m|\eta^{-1}) \ \tilde{\mathcal{S}}(n, \lambda_*) \right)^{1-2s} \\ &\leq C_s' \log^{2(s+1)}(16|\Lambda_m|\eta^{-1}) \lambda_*^s \ \tilde{\mathcal{S}}(n, \lambda_*) \ , \end{aligned}$$

for some $C_s' > 0$. Finally, the result follows by applying Lemma 5.3.8. □

Proof of Corollary 5.3.12. The proof follows by combining Theorem 5.3.11 and the argumentation in the proof of Corollary 5.3.10. □

Chapter 6

Future Research

6.1 Asymptotics of Effective Dimension

The effective dimension $\mathcal{N}(\lambda) = \mathrm{Tr}\left[B(B+\lambda)^{-1}\right]$ has emerged as a crucial quantity parame-trizing the impact of the covariance operator B on the learning problem. We have treated 2 cases of eigenvalue distribution in detail (the regular case and the case beyond the regular case), where we have derived the asymptotic behavior of the effective dimension as $\lambda \to 0$. But, in some sense, our approach has been naive: Surely, the trace contains averaged information, and we have derived the asymptotics of this average by imposing much more detailed conditions on the asymptotics of individual eigenvalues (in the regular case), which we then have relaxed (in the case beyond the regular case), but still staying in a framework of imposing conditions on eigenvalues individually.

It is not clear how one could do better in a completely distribution free context, where the input space $\mathcal{X}$ and the associated reproducing kernel is arbitrary. But since a more thorough understanding of the effective dimension undoubtedly is of importance, we think that it is of interest to use tools of microlocal analysis to analyze the effective dimension more closely in some classical cases (related but not identical to classical Sobolev type conditions). As a first step, one could consider the simplest case $\mathcal{X} = \mathbb{R}^d$ (which allows to use Fourier transform) and take B as a pseudodifferential operator of trace class. [1]

We recall a few basic facts (see e.g. [49], [31]). By the Schwarz kernel theorem, any linear continuous operator B from the Schwarz space $\mathcal{S}(\mathbb{R}^d)$ to its dual $\mathcal{S}'(\mathbb{R}^d)$ is represented by a distributional kernel K_B in $\mathcal{S}'(\mathbb{R}^d \times \mathbb{R}^d)$, which by Fourier transform induces a symbol b_t in $\mathcal{S}'(\mathbb{R}^d \times \mathbb{R}^d)$, formally given by

$$b_t(x,\xi) = \int e^{-iy\cdot\xi} K_B(x+(1-t)y, x-ty)dy, \qquad 0 \le t \le 1. \tag{6.1.1}$$

We emphasize that the above "integral" is not in the Lebesgues sense; it is merely a convenient notation for the Fourier transform of distributions (well defined by duality), and neither $K_B(x,y)$ nor $b_t(x,\xi)$ have pointwise sense on this level of generality. The symbol b induces a (so called pseudodifferential) operator

[1] The use of pseudodifferential operators is closely related to my diploma thesis in Analysis/Mathematical Physics. In addition I acknowledge helpful discussions with Markus Klein on the Weyl estimate and the theory of Gevrey spaces. Without those this section could not have been written as it stands.

$\mathrm{Op}_t(b)$, for any $t \in [0,1]$, formally given by the quantization rule

$$\mathrm{Op}_t(b)u(x) = (2\pi)^{-d} \int e^{i(x-y)\cdot\eta} b(tx + (1-t)y, \eta) u(y) dy d\eta. \tag{6.1.2}$$

We have defined the so called t quantisation. Loosely speaking, for $t = 1$ *all derivatives have been put to the right of the multiplication operators*, while for $t = 0$ derivatives are on the left. For the Weyl quantisation $t = 1/2$, a symmetric compromise, real symbols induce symmetric operators (and this seems to be most appropriate in our context, reducing bother with subleading terms). For details we refer to the literature. It then follows that $B = \mathrm{Op}_t(b_t)$, which establishes a general correspondence between kernel induced linear operators and (very general) pseudodifferential operators.

Most applications of pseudodifferential operators require that the symbol belongs to a much more specific symbol class, which typically is a Fréchet space of C^∞ functions with Fréchet seminorms specifying bounds on the derivatives (of all orders). A classical example would be the space $S^{m_1,m_2}_{\delta_1,\delta_2}(\mathbb{R}^{2d})$ consisting of functions $b \in \mathcal{C}^\infty(\mathbb{R}^d)$ satisfying

$$|\partial_x^\alpha \partial_\xi^\beta b(x,\xi)| \leq C_{\alpha,\beta} \langle x \rangle^{m_1 - \delta_1|\alpha|} \langle \xi \rangle^{m_2 - \delta_2|\beta|}, \qquad \langle x \rangle := (1 + |x|^2)^{1/2}. \tag{6.1.3}$$

There is a huge amount of different symbol classes in the literature (the most general are the Hörmander classes $S(m,g)$, see [49], for a slowly varying Riemannian metric g on $T^*\mathbb{R}^d$ and a corresponding $g-$ continuous order function $m(x,\xi)$), but (6.1.3) should suffice for a start. We remark that the limiting case $\delta_1 = \delta_2 = 0$ corresponds to a calculus *without gain* (at least if one does not consider a semiclassical situation as in [31]), while at least one $\delta_j > 0$ corresponds to a calculus *with gain*. Only in the latter case we expect good control of the effective dimension.

We recall that for symbols a in such a class there are reasonable sufficient conditions for $A = \mathrm{Op}_t(a)$ to be of trace class, namely

$$\sum_{|\alpha| \leq 2d+1} ||\partial^\alpha_{x,\xi} a||_{L^1} < \infty, \tag{6.1.4}$$

which might be slightly weakened. As usual, a sharp characterization of A being trace class, corresponding to a necessary condition, is unknown. But under condition (6.1.4), one simply has

$$\mathrm{Tr}\,[\mathrm{Op}(a)] = (2\pi)^{-n} \int a(x,\xi) dx d\xi, \tag{6.1.5}$$

both for the $t = 1$ and the Weyl quantization.

To compute the effective dimension efficiently via this formula, 2 problems have to be adressed: Firstly, obtain a representation of the resolvent $(\lambda + B)^{-1}$ as a pseudodifferential operator and an expansion of its symbol, and, secondly, to have a sharp parameter dependent control of the error term in the symbolic calculus. The first problem can in principle be solved via Beals characterisation of pseudodifferential operators (based on this idea, an asymptotic expansion of the resolvent has in the semiclassical case been worked out in [31] which hopefully can be adapted to the present slightly different setting). The starting point is to use the symbol $(\lambda + b(x,\xi))^{-1}$ to define an appropriate parametrix, and then continue via the symbolic calculus (which is central for the second problem also). The construction crucially depends on some sort of ellipticity of the covariance operator B and its symbol. As a first step, possibly well adapted

to the symbol space $S^{m_1,m_2}_{\delta_1,\delta_2}(\mathbb{R}^{2d})$, the (strong) assumption

$$b(x,\xi) \geq C\langle x\rangle^{-M_1}\langle \xi\rangle^{-M_2}, \tag{6.1.6}$$

for some $M_1, M_2 > 0$, might be appropriate. The second problem depends on the evaluation of the composition formula in the so called symbolic calculus.

We recall that, if the symbols a, b belong to the Hörmander class $S(m_1, g), S(m_2, g)$ respectively, then $\mathrm{Op}_t(a)\mathrm{Op}_t(b) = \mathrm{Op}_t(c)$, where $c \in S(m_1 m_2, g)$ is given by

$$c(x,\xi) = e^{iq_t(D_x,D_\xi;D_y,D_\eta)}a(x,\xi)b(y,\eta)|_{y=x,\eta=\xi}, \quad D_x = \frac{1}{i}\partial_x, \ldots, \tag{6.1.7}$$

where $q_t(D_x, D_\xi; D_y, D_\eta)$ is a certain quadratic form (for $t = 1$ it is simply $D_\xi \cdot D_\eta$, for the Weyl quantisation it is $\frac{1}{2}\sigma(D_x, D_\xi; D_y, D_\eta)$, with σ being the standard symplectic form in the cotangent bundle $T^*\mathbb{R}^d$, isomorphic to $\mathbb{R}^{2d}$). At least formally, this gives an asymptotic expansion

$$c(x,\xi) \sim \sum_0^\infty \frac{1}{k!} q_t(D_x, D_\xi; D_y, D_\eta)^k a(x,\xi)b(y,\eta)|_{y=x,\eta=\xi}. \tag{6.1.8}$$

As usual, the main problem are good remainder estimates, including control on additional parameters. Carefully checking all remainder terms and controlling the spectral parameter hopefully gives a result of the type

$$\mathcal{N}(\lambda) = (2\pi)^{-d}\int \frac{b(x,\xi)}{\lambda + b(x,\xi)} dx d\xi (1 + o(1)), \quad \text{as } \lambda \downarrow 0, \tag{6.1.9}$$

if appropriate lower bounds on the symbol $b(x, \xi)$ as e.g. in (6.1.6) are imposed. Clearly, in a lot of cases this integral can be evaluated to extract the asymptotic behavior as $\lambda \downarrow 0$. However, we already mention here that not all naive choices of the symbol $b(x, \xi)$ are legal in our context. In particular, the symbol is not allowed to be of compact support: By the easy part of the Paley Wiener theorem, compact support in ξ implies analyticity of the Fourier transform, thus by (6.1.1) analyticity of the kernel $\mathcal{K}_B(x, y)$ in the second variable. Since B is assumed to be self-adjoint, the kernel is symmetric. Thus compact support in the first variable also implies vanishing of the kernel, by analyticity. More general, one should keep in mind that a simple formula as (6.1.9) requires at least some form of ellipticity.

Of course, formulated in this way, these are problems in analysis. Since they are very close in spirit to my diploma thesis, see [67], on functional calculus for pseudodifferential operators (which is a starting point for the classical Weyl estimates), it seems natural to me to apply these techniques also to obtain additional insight into the statistically significant object of the effective dimension.

In addition, using functional calculus, I would like to investigate in which cases an estimate analogous to Chapter 3 will hold, i.e.

$$\frac{1}{C}\mathrm{Tr}\left[1_{[\lambda,\infty)}(B)\right] \leq \mathcal{N}(\lambda) \leq C\mathrm{Tr}\left[1_{[\lambda,\infty)}(B)\right], \tag{6.1.10}$$

for some constant $C > 0$. If B is the inverse of a positive elliptic pseudodifferential operator P, this

compares the effective dimension to the number of eigenvalues of P below λ^{-1}. By the Weyl estimate this is asymptotic to the volume of $p^{-1}([0,\lambda^{-1}])$ in $T^*\mathbb{R}^d$, if p denotes the symbol of P.

Furthermore, I would like to have precise estimates relating the asymptotics of the effective dimension to a wide range of smoothness classes of the associated kernel (starting with $\mathcal{X} = \mathbb{R}^d$). Since via Fourier transform smoothness of the kernel transforms into decay in the covariable ξ, the symbol classes defined in (6.1.3) are appropriate for Sobolev-type smoothness corresponding to existence of finitely many derivatives. But if the kernel is much more regular, e.g. analytic (as an entire function, or in a strip $|\Im y| < \beta$), the corresponding symbol decays exponentially, by the Paley-Wiener theorem. Clearly, this leads to a much smaller effective dimension, which should be evaluated asymptotically. Here it is essential to relax the ellipticity condition (6.1.6) while still keeping control on the trace. Similar relations should hold for regularity of the kernel in a Gevrey - s class (for $s > 1$), which by Fourier transform leads to decay of the symbol as a stretched exponential, i.e.

$$b(x,\xi) \leq C \exp(-c|\xi|^{1/s}). \tag{6.1.11}$$

For the sake of the reader, we recall that Gevrey spaces were introduced in [39]. A standard textbook is [75]. A finer notion of Gevrey spaces (appropriate for slightly more refined estimates on the Fourier transform) was used in [52]: For $s \geq 1, b > 0$, the Gevrey space $\Gamma^{s,b}$ is the space of all $f \in \mathcal{C}^\infty(\mathbb{R}^d)$ such that

$$||\partial^\alpha f||_\infty \leq c_0(f)(|\alpha|+1)^{c_0(f)} b^{|\alpha|} |\alpha|!^s. \tag{6.1.12}$$

The global Gevrey space Γ^s and the small Gevrey space γ^s (see [48]) then correspond to union and intersection with respect to the parameter b. We recall that Gevrey-s regularity is intermediate between analyticity and smoothness in the $\mathcal{C}^\infty-$ sense. For $s = 1$ it reduces to analyticity, but for $s > 1$ it is not a quasi-analytic class in the sense of the Denjoy-Carlemann theorem, i.e. it contains non-trivial functions of compact support. A version of the Paley Wiener theorem then states that for $f \in \Gamma^{s,b} \cap L^1(\mathbb{R}^d)$ - recall that, since B is trace class, it is reasonable to consider kernels $K(x,\cdot)$ in $L^1(\mathbb{R}^d)$ - the Fourier transforms $\hat{f}(\xi)$ satisfies the estimate (6.1.11), for all constants $c < sb^{-\frac{1}{s}}$. To the best of our knowledge, this estimate in the case of an L^1 condition - which is natural in our context - is not to be found in the literature, since the classical Paley Wiener theorem is concerned with the Fourier transform of functions of compact support only (for Gevrey spaces it is due to Komatsu, see [53], in its work on ultradistributions, which form the dual space; for analytic functions see [47]). But the proof in [52] via almost analytic extensions applies (and gives a characterization of $f \in \Gamma^{s,b}$ in terms of the estimate (6.1.11), with the relation between s, b, c as given above).

It is not clear if even in these cases the asymptotics of the effective dimension are described by the naive leading term (6.1.9), since ellipticity is pretty much lost at infinity. It has to be checked if there is still a useful expansion of the resolvent and a useful symbolic calculus, possibly utilizing the full range of the H"ormander spaces $S(m,g)$.

In any case, if the symbol $b(x,\xi)$ can be bounded by a product of weightfunctions in x and ξ and there is some sort of ellipticity, we expect an upper bound of the form

$$\mathcal{N}(\lambda) \leq F_1(\lambda) F_2(\lambda), \tag{6.1.13}$$

where $\mathcal{F}_j(\lambda) \uparrow \infty$ as $\lambda \downarrow 0$, $F_j(\lambda)$ (and $F_1(\lambda)F_2(\lambda)$) are bounded by λ^{-1} and

$$F_j(\lambda) \leq C(\log \lambda^{-1})^{sd}$$

corresponds to decay of $b(x,\xi)$ as in (6.1.11), where decay in x gives an estimate on $F_1(\lambda)$, and decay in ξ on $F_2(\lambda)$. Clearly, since the effective dimension depends symmetrically on x and ξ, decay in only one of the variables may be completely masked by the reverse asymptotics in the other variable. In particular, it is not true that smooth kernels automatically give small (or even logarithmically bounded) effective dimension.

Furthermore we remark that, although general marginals ν which do not have a smooth density with respect to Lebesgues measure, are not easily accessible to a pseudodifferential approach, measures concentrated on a submanifold of lower dimension are accessible via covariance operators which are pseudodifferential on an appropriate submanifold of $\mathbb{R}^n$. Such examples are classical in the context of Weyl estimates. We recall that in all these cases the sharpest results have been obtained by an additional use of evolution equations, using Fourier integral operators. A stationary approach, even using the best pseudodifferential calculus available (see e.g. [46], has not yet reproduced the known sharp estimates on the error term. It is wide open if a precise analysis of the effective dimension will show similar phenomena.

Finally we remark that, at least in the semiclassical context, Weyl type estimates have been extended to cases where a complete asymptotic expansion of the resolvent is not available, e.g. pseudodifferential operator boundary values. This goes back to work of Ivrii, see [50], using hyperbolic energy estimates, and has been transformed to a more stationary approach by Dimassi and Sjöstrand in [30], by systematic use of almost analytic extensions. A complete version is contained in the book [31]. In the distribution free philosophy, it is certainly interesting to check if these ideas can be applied or adapted to a more general study of the trace in the definition of the effective dimension.

6.2 Non-Linear Inverse Problems

We present some thoughts on how to use the methods of this thesis for the problem of solving a fully non-linear inverse regression problem. These thoughts are preliminary and might possibly change substantially upon working them out in detail. We remark that there is already a vast literature on the deterministic non-linear inverse problem, see e.g. [35], [68], [85], [51], but there are much less results in the stochastic setting, see [69], [59], [9]. In particular, to the best of our knowledge, there is no implementation of kernel based methods for a large class of general spectral regularization schemes.

We want to formulate the regression problem and our kernel based approach in a way similar to the approach of this thesis for the linear problem. Thus we consider the non-linear inverse regression problem

$$Y_i = g(X_i) + \epsilon_i, \qquad g = A(f), \qquad i = 1 \dots n, \tag{6.2.1}$$

where $A : \mathcal{D}(A) \to \mathcal{H}_2$ is a known non-linear operator on a domain $\mathcal{D}(A) \subset \mathcal{H}_1$ in some Hilbert space $\mathcal{H}_1$. We take $\mathcal{D}(A)$ convex and weakly sequentially closed and we let $\mathcal{H}_2$ be a space of real valued functions on some input space $\mathcal{X}$, taken to be standard Borel. Thus, for simplicity, the output space $\mathcal{Y}$, containing the oucomes Y_i, is the real line, but as for the rest of this thesis we expect that this could be generalized

to any separable Hilbert space with very little additional effort.

As in the linear case, we assume that the observed data $(X_i, Y_i)_{1\le i\le n} \subset (\mathcal{X}\times\mathcal{Y})^n$ are iid random variables, drawn according to a data generating distribution ρ on $\mathcal{X}\times\mathcal{Y}$, with $\mathbb{E}[Y_i|X_i] = g(X_i)$, so that the distribution of ϵ_i may depend on X_i but satisfies $\mathbb{E}[\epsilon_i|X_i] = 0$. As above, we furthermore assume that the map $(f,x)\mapsto (A(f))(x)$ is measurable in x and continuous in f (which requires to endow $\mathcal{H}_2$ with some a priori topological structure, e.g. by assuming $\mathcal{H}_2$ to be a Hilbert space). Furthermore, we shall take A to be at least $\mathcal{C}^1$ in Fréchet- sense on the open interior of $\mathcal{D}(A)$ (with a continuous extension of DA to the boundary of $\mathcal{D}(A)$), with $DA : \mathcal{D}(A)\to\mathcal{L}(\mathcal{H}_1,\mathcal{H}_2)$ being (at least) locally Lipschitz. Possibly, it is useful to require $A\in\mathcal{C}^k(\mathcal{D}(A),\mathcal{H}_2)$ for some $k>1$ (in the sense that the derivatives D^jA, for $0\le j\le k$, have continuous extensions from the interior to the boundary of $\mathcal{D}(A)$).

Then the goal is to approximate the target function $f\in\mathcal{D}(A)$ by an estimator $\widehat{f}_n$, depending on the data (X_i,Y_i) in $(\mathcal{X}\times\mathcal{Y})^n$ in some optimal way. As in any non-linear problem, any iterative algorithm to solve the non-linear inverse regression problem (6.2.1) requires the choice of some good starting point $f_0\in\mathcal{H}_1$, sufficiently close to the solution f. How to get this f_0 (e.g. by some probabilistic search algorithm) is a separate problem which we shall not discuss here. We assume f_0 sufficiently close to f, as being given.

Setting $A_0 := DA|_{f_0}$ and replacing $A(f)$ by its linearization $A(f_0)+A_0h_0$, $h_0 := f-f_0$, at f_0, we obtain the first linearized inverse regression model (indexed by zero)

$$Y_i^0 = g_0(X_i)+\epsilon_i,\qquad Y_i^0 := Y_i - A(f_0)(X_i),\qquad g_0 = A_0h_0. \tag{6.2.2}$$

We shall assume that for $x\in\mathcal{X}, f\in\mathcal{D}(A)$ all evaluation functionals

$$S_{x,f}:\mathcal{H}_1\to\mathbb{R},\qquad S_{x,f}h := (DA|_f - f_0h)(x)$$

are uniformly bounded in x:

$$|S_{x,f}h|\le\kappa||h||_{\mathcal{H}_1}\qquad (x\in\mathcal{X}),$$

possibly taking κ to be uniform also w.r.t. $f\in\mathcal{D}(A)$. Then, by Riesz, there is $F_{x,f}\in\mathcal{H}_1$ with

$$(DA|_fh)(x) = \langle h, F_{x,f}\rangle_{\mathcal{H}_2},\qquad h\in\mathcal{H}_1,$$

defining a base point dependent feature map

$$\mathcal{X}\ni x\mapsto F_{x,f}\in\mathcal{H}_1$$

and a p.s.d. base point dependent kernel

$$K_f:\mathcal{X}\times\mathcal{X}\to\mathbb{R},\qquad K_f(x_1,x_2) := \langle F_{x_1,f}, F_{x_2,f}\rangle_{\mathcal{H}_1}.$$

This defines for each base point $f\in\mathcal{D}(A)$ an RKHS structure on $\mathrm{Im}(DA|_f) =: \mathcal{H}_{K_f}\subset\mathcal{H}_2$, a space of bounded real-valued functions. Assuming, as above, measurability of all functions in $\mathcal{H}_{K_f}$, when f varies through $\mathcal{D}(A)$, defines a continuous embedding $\mathcal{H}_{K_f}\to L^2(\mathcal{X},d\nu)$, where ν is the marginal distribution of ρ describing the law of X_i. Furthermore, we obtain the covariance operator (at the base point f)

$$B_{\nu,f} := \int_{\mathcal{X}} F_{x,f}\otimes F_{x,f}^*\,\nu(dx).$$

Having all this at our disposal, we may apply the kernel methods developed in this thesis to approximately solve the linear regression problem (6.2.2):

For any sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^n$ we define the empirical sampling operator (for fixed base point $f \in \mathcal{D}(A)$)

$$S_{\mathbf{x},f} : \mathcal{H}_1 \to \mathbb{R}^n, \qquad (S_{x,f}h)_j := \langle h, F_{x_j,f} \rangle_{\mathcal{H}_1}.$$

Then the empirical inner product on $\mathbb{R}^n$ defines the adjoint

$$S^*_{x,f} : \mathbb{R}^n \to \mathcal{H}_1, \qquad S^*_{x,f}\mathbf{y} = \frac{1}{n}\sum_{j=1}^{n} y_j F_{x_j,f}$$

and the empirical covariance operator

$$B_{x,f} = S^*_{x,f}S_{x,f} = \frac{1}{n}\sum_{j=1}^{n} F_{x_j,f} \otimes F_{x_j,f}.$$

Introducing, as in Chapter 2, the normalized operators $\bar{B}_{x,f}$ and $\bar{S}_{x,f}$ and choosing a regularizing function g_λ, the regularized approximate solution of (6.2.2) is defined by

$$h^\lambda_{\mathbf{z},0} := g_\lambda(\bar{B}_{x,f_0})S^*_{x,f_0}\mathbf{y}. \tag{6.2.3}$$

Note that, for any linearized problem with fixed base point, we may consider source conditions and classes of marginals as explained in Chapter 2 and 3. However, to cover the non-linear case, some globalization procedure is needed, e.g. by taking an intersection over all base points varying in some appropriate subset of $\mathcal{D}(A)$. Furthermore, the most naive approach is in using *all* data $\mathbf{Z} \in (\mathcal{X} \times \mathcal{Y})^n$, in every single linearization step. But, certainly, any efficient algorithm will ultimately depend on using in each step just an appropriate fraction of data. This requires to define an appropriate map

$$\mathbb{N} \ni j \mapsto m_j \leq n$$

which associates to each iteration step j a reasonable cardinality $M_j \leq n$ of an appropriate subset of data to be used in iteration step j. We expect this to be crucial for defining a reasonable iterative solution algorithm for the non-linear regression problem (6.2.1).

Coming back to the approximate regularized solution $h^\lambda_{\mathbf{z},0}$ of (6.2.2), we obtain an approximate solution $f_1 \in \mathcal{D}(A)$ of (6.2.1), provided f_0 was chosen luckily, by

$$f_1 := f_0 + h^{\lambda_{m_0}}_{\mathbf{z},0},$$

where m_0 is the cardinality of data $\mathbf{z}$ used in the first iteration step. Assuming $f_1 \in \mathcal{D}(A)$, which is a first basic requirement of any *good* starting point f_0, and similarly for all iterates f_j, one may set inductively

$$f_{j+1} = f_j + h^{\lambda_{m_j}}_{\mathbf{z},j}, \qquad (j \leq J(n)), \tag{6.2.4}$$

where $J(n)$ is the number of iterations used for data of cardinality n (which has to be determined by an appropriate stopping rule) and

$$h^{\lambda_{m_j}}_{\mathbf{z},j} = g_{\lambda_{m_j}}(\bar{B}_{x,f_j})\bar{S}^*_{x,f_j}\mathbf{y}$$

is the approximate regularized solution of the j-th linearized regression problem

$$Y_i^j = g_j(X_i) + \epsilon_i, \qquad i = 1, \dots m_j, \tag{6.2.5}$$

where

$$g_j := A_j h_j, \quad A_j := DA|_{f_j}, \qquad h_j := f - f_j, \quad Y_i^j := Y_i - A(f_j)(Xi).$$

Here the data $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^{m_j}$ depend on the iteration step j (and have to be extracted from the data in $(\mathcal{X} \times \mathcal{Y})^n$ by some well defined procedure), but for simplicity of notation we have suppressed an additional index in our notation.

Clearly, for data of size n and an appropriate stopping rule, bounding the number of iterations by $j \leq J(n)$, we obtain a last approximation

$$\widehat{f}_n := f^n_{J(n)+1} = f_0 + \sum_{j=0}^{J(n)} h_{\mathbf{z},j}^{\lambda_{m_j}} \tag{6.2.6}$$

for which one would like to investigate rates of convergence and their optimality. Clearly, the ultimate goal is to define in a purely data driven way a stopping criterion defining the appropriate number of iterations $J(n)$ and a sequence of cardinalities of data m_j, $1 \leq j \leq J(n)$, where m_j is somewhat minimal but still sufficient to give optimal rate of convergence for the etimator $\widehat{f}_n$.

As a preliminary step, one might in analogy to Chapter 2 and 3 analyze rates of convergence for a known *a-priori* smoothness of f and fixed assumptions on the spectral properties of all covariance operators $B_{\nu,f}$ for ν in a fixed class of marginals and $f \in \mathcal{D}(A)$, but finally one aims at a truly adaptive estimator. It is clear that proceeding in this manner requires a chain of clarifying definitions adapted to the non-linear case, which in particular will specify *optimality* and *adaptivity*. It is also clear that any convergence analysis has to compare (and finally balance) the errors in the non-linear iteration scheme (which one might call a stochastic Newton method by the obvious analogy with the deterministic Newton method) with the errors arising in the approximate solution of the linearized regression problems, which have been analyzed in this thesis. This will be at the heart of the non-linear regression problem and requires, beyond the results of this thesis, reasonable error estimates for the stochastic Newton method, e.g. by some form of the discrepancy principle, popular in the deterministic case.

We also expect that the number of iterations in the Newton method serves as additional regularization for the linearized regression problems (somewhat similar to out findings for DL in Chapter 4), which might make underregularization attractive. Furthermore, any numerically efficient implementation should use appropriate combinations of speeding up the solution of the linearized problems.

Clearly, in all these problems there is a lot of overlap with the methods and results of this thesis. This puts a fairly complete analysis of the stochastic Newton method high up on my priority list for natural future research topics.

6.3 LocalNysation: Combining Localized Kernel Regression and Nyström Subsampling

It is clear from our discussion in Chapter 4 that at least some of the different approaches to speed up computation for large data sets should be combined to produce a twofold beneficial effect on computation time. Instead of describing this metatheoretically, we did - after the rest of this thesis had been finished - proceed by way of example, combining the partitioning approach and plain Nyström Subsampling. The corresponding paper has been submitted to a peer-reviewd conference and we reproduce it here as Appendix B.

Appendices

Appendix A

Tools

A.1 Concentration Inequalities

Proposition A.1.1. *Let $(Z, \mathcal{B}, \mathbb{P})$ be a probability space and ξ a random variable on Z with values in a real separable Hilbert space $\mathcal{H}$. Assume that there are two positive constants L and σ such that for any $m \geq 2$*

$$\mathbb{E}\big[\, \|\xi - \mathbb{E}[\xi]\|_{\mathcal{H}_1}^m \,\big] \leq \frac{1}{2} m! \sigma^2 L^{m-2}. \tag{A.1.1}$$

If the sample $z_1, ..., z_n$ is drawn i.i.d. from Z according to $\mathbb{P}$, then, for any $0 < \eta < 1$, with probability greater than $1 - \eta$

$$\Big\| \frac{1}{n} \sum_{j=1}^{n} \xi(z_j) - \mathbb{E}[\xi] \Big\|_{\mathcal{H}} \leq 2 \log(2\eta^{-1}) \left(\frac{L}{n} + \frac{\sigma}{\sqrt{n}} \right) . \tag{A.1.2}$$

In particular, (A.1.1) holds if

$$\begin{aligned} \|\xi(z)\|_{\mathcal{H}_1} &\leq \frac{L}{2} \qquad a.s. \,, \\ \mathbb{E}\big[\, \|\xi\|_{\mathcal{H}_1}^2 \,\big] &\leq \sigma^2. \end{aligned}$$

Proof. See [20, 21], from the original result of [71] (Corollary 1) . □

The following propositions summarize important concentration properties of the empirical quantities involved.

Proposition A.1.2. *For $n \in \mathbb{N}$, $\lambda \in (0, 1]$ and $\eta \in (0, 1]$, it holds with probability at least $1 - \eta$:*

$$\big\| (\bar{B} + \lambda)^{-\frac{1}{2}} \, \big(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y} \big) \big\|_{\mathcal{H}_1} \leq 2 \log(2\eta^{-1}) \left(\frac{M}{n\sqrt{\lambda}} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n}} \right) .$$

Also, it holds with probability at least $1 - \eta$:

$$\big\| \bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y} \big\|_{\mathcal{H}_1} \leq 2 \log(2\eta^{-1}) \left(\frac{M}{n} + \sqrt{\frac{\sigma^2}{n}} \right) .$$

Proof of Proposition A.1.2. Define $\xi_1 : \mathcal{X} \times \mathbb{R} \longrightarrow \mathcal{H}_1$ by

$$\begin{aligned}\xi_1(x,y) &:= (\bar{B}+\lambda)^{-1/2}(y - \bar{S}_x f_\rho)\bar{F}_x \\ &= (\bar{B}+\lambda)^{-1/2}\bar{S}_x^\star(y - \bar{S}_x f_\rho)\,.\end{aligned}$$

Abusing notation we also denote ξ_1 the random variable $\xi_1(X,Y)$ where $(X,Y) \sim \rho$. The model assumption (2.2.4) implies

$$\begin{aligned}\mathbb{E}[\xi_1] &= (\bar{B}+\lambda)^{-1/2} \int_{\mathcal{X}} \bar{F}_x \int_{\mathbb{R}} (y - \bar{S}_x f_\rho)\, \rho(dy|x)\nu(dx) \\ &= (\bar{B}+\lambda)^{-1/2} \int_{\mathcal{X}} \bar{F}_x(\bar{S}_x f_\rho - \bar{S}_x f_\rho)\, \nu(dx) \\ &= 0\,,\end{aligned}$$

and therefore

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^n \xi_1(x_j,y_j) - \mathbb{E}[\xi_1] &= \frac{1}{n}\sum_{j=1}^n (\bar{B}+\lambda)^{-1/2}(y_j - \bar{S}_{x_j} f_\rho)\bar{F}_{x_j} \\ &= (\bar{B}+\lambda)^{-1/2}\bar{S}_{\mathbf{x}}^\star\ \left(\mathbf{y} - \bar{S}_{\mathbf{x}} f_\rho\right)\ . \\ &= (\bar{B}+\lambda)^{-1/2}\ \left(\bar{S}_{\mathbf{x}}^\star \mathbf{y} - \bar{B}_{\mathbf{x}} f_\rho\right)\ .\end{aligned}$$

Moreover, by assumption (2.2.5), for $m \geq 2$:

$$\begin{aligned}\mathbb{E}[\ \|\xi_1\|_{\mathcal{H}_1}^m\ ] &= \int_{\mathcal{X}\times\mathbb{R}} \left\|(\bar{B}+\lambda)^{-1/2}\bar{S}_x^\star(y - \bar{S}_x f_\rho)\right\|_{\mathcal{H}_1}^m \rho(dx,dy) \\ &= \int_{\mathcal{X}\times\mathbb{R}} |\langle \bar{S}_x(\bar{B}+\lambda)^{-1}\bar{S}_x^\star(y - \bar{S}_x f_\rho), (y - \bar{S}_x f_\rho)\rangle_{\mathbb{R}}|^{\frac{m}{2}} \rho(dx,dy) \\ &\leq \int_{\mathcal{X}} \left\|\bar{S}_x(\bar{B}+\lambda)^{-1}\bar{S}_x^\star\right\|^{\frac{m}{2}} \int_{\mathbb{R}} \left|y - \bar{S}_{\mathbf{x}} f_\rho\right|^m\ \rho(dy|x)\nu(dx) \\ &\leq \frac{1}{2}m!\sigma^2 M^{m-2} \int_{\mathcal{X}} \left\|\bar{S}_x(\bar{B}+\lambda)^{-1}\bar{S}_x^\star\right\|^{\frac{m}{2}}\ \nu(dx)\,.\end{aligned}$$

Setting $A := \bar{S}_x(\bar{B}+\lambda)^{-1/2}$, we have using $\|\cdot\| \leq \mathrm{Tr}\,[\cdot]$ for positive operators

$$\begin{aligned}\left\|\bar{S}_x(\bar{B}+\lambda)^{-1}\bar{S}_x^\star\right\|^{\frac{m}{2}} &= \|AA^\star\|^{\frac{m}{2}} \leq \|AA^\star\|^{\frac{m}{2}-1}\,\mathrm{Tr}\,[AA^\star] \\ &= \|AA^\star\|^{\frac{m}{2}-1}\,\mathrm{Tr}\,[A^\star A]\ .\end{aligned}$$

Firstly, observe that

$$\|AA^\star\|^{\frac{m}{2}-1} \leq \left(\frac{1}{\lambda}\right)^{\frac{m}{2}-1}\ ,$$

since our main assumption 2.2.1 implies $\left\|\bar{S}_x\right\| \leq 1$. Secondly, by linearity of $\mathrm{Tr}\,[\cdot]$

$$\int_{\mathcal{X}} \mathrm{Tr}\,[A^\star A]\,\nu(dx) = \int_{\mathcal{X}} \mathrm{Tr}\left[(\bar{B}+\lambda)^{-1/2}\bar{B}_x(\bar{B}+\lambda)^{-1/2}\right]\nu(dx) = \mathcal{N}(\lambda)\,.$$

Thus,

$$\mathbb{E}[\ \|\xi_1\|_{\mathcal{H}_1}^m\ ] = \frac{1}{2}m!(\sigma\sqrt{\mathcal{N}(\lambda)})^2\left(\frac{M}{\sqrt{\lambda}}\right)^{m-2}\ .$$

As a result, Proposition $A.1.1$ implies with probability at least $1-\eta$

$$\left\|(\bar{B}+\lambda)^{-1/2}(\bar{B}_{\mathbf{x}}f_\rho - \bar{S}^\star_{\mathbf{x}}\mathbf{y})\right\|_{\mathcal{H}_1} \leq \delta_1\left(n,\eta\right) \ ,$$

where

$$\delta_1(n,\eta) = 2\log(2\eta^{-1})\left(\frac{M}{n\sqrt{\lambda}} + \frac{\sigma}{\sqrt{n}}\sqrt{\mathcal{N}(\lambda)}\right).$$

For the second part of the proposition, we introduce similarly

$$\xi_1'(x,y) := (y - \bar{S}_x f_\rho)\bar{F}_x = \bar{S}_x^\star(y - \bar{S}_x f_\rho) \ ,$$

which satisfies

$$\mathbb{E}[\xi_1'] = 0\,; \qquad \frac{1}{n}\sum_{j=1}^n \xi_1'(x_j,y_j) - \mathbb{E}[\xi_1'] = \bar{S}^\star_{\mathbf{x}}\mathbf{y} - \bar{B}_{\mathbf{x}}f_\rho\,,$$

and

$$\mathbb{E}\big[\,\|\xi_1'\|^m_{\mathcal{H}_1}\,\big] \leq \mathbb{E}\big[\,\big|y - \bar{S}_x f_\rho\big|^m\,\big] \leq \frac{1}{2}m!\sigma^2 M^{m-2}\,.$$

Applying Proposition $A.1.1$ yields the result.

□

Proposition A.1.3. *For any $n \in \mathbb{N}$, $\lambda \in (0,1]$ and $\eta \in (0,1)$, it holds with probability at least $1-\eta$:*

$$\left\|(\bar{B}+\lambda)^{-1}(\bar{B} - \bar{B}_{\mathbf{x}})\right\|_{\mathrm{HS}} \ \leq 2\log(2\eta^{-1})\left(\frac{2}{n\lambda} + \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}}\right) \ .$$

Proof of Proposition A.1.3. We proceed as above by defining $\xi_2 : \mathcal{X} \longrightarrow \mathrm{HS}(\mathcal{H}_1)$ (the latter denoting the space of Hilbert-Schmidt operators on $\mathcal{H}_1$) by

$$\xi_2(x) := (\bar{B}+\lambda)^{-1}\bar{B}_x\,,$$

where $\bar{B}_x := \bar{F}_x \otimes \bar{F}_x^\star$. We also use the same notation ξ_2 for the random variable $\xi_2(X)$ with $X \sim \nu$. Then,

$$\mathbb{E}[\xi_2] = (\bar{B}+\lambda)^{-1}\int_{\mathcal{X}} \bar{B}_x\ \nu(dx) = (\bar{B}+\lambda)^{-1}\bar{B}\,,$$

and therefore

$$\frac{1}{n}\sum_{j=1}^n \xi_2(x_j) - \mathbb{E}[\xi_2] = (\bar{B}+\lambda)^{-1}(\bar{B} - \bar{B}_{\mathbf{x}})\ .$$

Furthermore, since $\bar{B}_x$ is of trace class and $(\bar{B}+\lambda)^{-1}$ is bounded, we have using Assumption 2.2.1

$$\|\xi_2(x)\|_{\mathrm{HS}} \leq \left\|(\bar{B}+\lambda)^{-1}\right\|\left\|\bar{B}_x\right\|_{\mathrm{HS}} \leq \lambda^{-1} =: L_2/2\ ,$$

uniformly for any $x \in \mathcal{X}$. Moreover,

$$\begin{aligned}\mathbb{E}\big[\,\|\xi_2\|_{\mathrm{HS}}^2\,\big] &= \int_{\mathcal{X}} \mathrm{Tr}\,\big[\,\bar{B}_x(\bar{B}+\lambda)^{-2}\bar{B}_x\,\big]\,\nu(dx)\\ &\le \big\|(\bar{B}+\lambda)^{-1}\big\| \int_{\mathcal{X}} \big\|\bar{B}_x\big\|\ \mathrm{Tr}\,\big[\,(\bar{B}+\lambda)^{-1}\bar{B}_x\,\big]\,\nu(dx)\\ &\le \frac{\mathcal{N}(\lambda)}{\lambda} =: \sigma_2^2\,.\end{aligned}$$

Thus, Proposition $A.1.1$ applies and gives with probability at least $1-\eta$

$$\big\|(\bar{B}+\lambda)^{-1}(\bar{B}-\bar{B}_{\mathbf{x}})\big\|_{\mathrm{HS}} \ \le\ \delta_2(n,\eta)$$

with

$$\delta_2(n,\eta) = 2\log(2\eta^{-1})\left(\frac{2}{n\lambda}+\sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}}\right)\,.$$

□

Proposition A.1.4. *Let $\eta \in (0,1)$. Assume that $\lambda \in (0,1]$ satisfies*

$$\sqrt{n\lambda} \ge 8\log(2\eta^{-1})\sqrt{\max(\mathcal{N}(\lambda),1)}\,. \tag{A.1.3}$$

Then, with probability at least $1-\eta$:

$$\big\|(\bar{B}_{\mathbf{x}}+\lambda)^{-1}(\bar{B}+\lambda)\big\| \le 2\,. \tag{A.1.4}$$

Proof of Proposition A.1.4. We write the Neumann series identity

$$(\bar{B}_{\mathbf{x}}+\lambda)^{-1}(\bar{B}+\lambda) = (I-H_{\mathbf{x}}(\lambda))^{-1} = \sum_{j=0}^{\infty} H_{\mathbf{x}}(\lambda)^j(\lambda)\,,$$

with

$$H_{\mathbf{x}}(\lambda) = (\bar{B}+\lambda)^{-1}(\bar{B}-\bar{B}_{\mathbf{x}})\,.$$

It is well known that the series converges in norm provided that $\|H_{\mathbf{x}}(\lambda)\| < 1$. In fact, applying Proposition A.1.3 gives with probability at least $1-\eta$:

$$\|H_{\mathbf{x}}(\lambda)\| \ \le\ 2\log(2\eta^{-1})\left(\frac{2}{n\lambda}+\sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}}\right)\,.$$

Put $C_\eta := 2\log(2\eta^{-1}) > 1$ for any $\eta \in (0,1)$. Assumption (A.1.3) reads $\sqrt{n\lambda} \ge 4C_\eta\sqrt{\max(\mathcal{N}(\lambda),1)}$, implying $\sqrt{n\lambda} \ge 4C_\eta \ge 4$ and therefore $\frac{2}{n\lambda} \le \frac{1}{2\sqrt{n\lambda}} \le \frac{1}{8C_\eta}$, hence

$$C_\eta\left(\frac{2}{n\lambda}+\sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}}\right) \le C_\eta\left(\frac{1}{8C_\eta}+\frac{1}{4C_\eta}\right) < \frac{1}{2}\,.$$

Thus, with probability at least $1-\eta$:

$$\big\|(\bar{B}_{\mathbf{x}}+\lambda)^{-1}(\bar{B}+\lambda)\big\| \le 2\,.$$

□

Proposition A.1.5. *For any $n \in \mathbb{N}$ and $0 < \eta < 1$ one has with probability at least $1 - \eta$:*

$$\left\|\bar{B} - \bar{B}_{\mathbf{x}}\right\|_{\mathrm{HS}} \;\leq\; 6\log(2\eta^{-1})\,\frac{1}{\sqrt{n}}\,.$$

Proof of Proposition A.1.5. Defining $\xi_3 : \mathcal{X} \longrightarrow \mathrm{HS}(\mathcal{H}_1)$ by

$$\xi_3(x) := \bar{F}_x \otimes \bar{F}_x^{\star} = \bar{B}_x$$

and denoting also, as before, ξ_3 for the random variable $\xi_3(X)$ (with $X \sim \nu$), we have $\mathbb{E}[\xi_3] = \bar{B}$ and therefore

$$\frac{1}{n}\sum_{j=1}^{n}\xi_3(x_j) - \mathbb{E}[\xi_3] = (\bar{B}_{\mathbf{x}} - \bar{B})\,.$$

Furthermore, by Assumption 2.2.1

$$\|\xi_3(x)\|_{\mathrm{HS}} = \left\|\bar{F}_x\right\|^2 \leq 1 =: \frac{L_3}{2} \qquad \text{a.s.}\,,$$

also leading to $\mathbb{E}[\|\xi_2\|_{\mathrm{HS}}^2] \leq 1 =: \sigma_3^2$. Thus, Proposition $A.1.1$ applies and gives with probability at least $1 - \eta$

$$\left\|\bar{B} - \bar{B}_{\mathbf{x}}\right\|_{\mathrm{HS}} \;\leq\; 6\log(2\eta^{-1})\,\frac{1}{\sqrt{n}}\,.$$

□

Lemma A.1.6. *Let $s \in [0, \frac{1}{2}]$ and $f \in \mathcal{H}_1$. If $0 < \lambda_1 \leq \lambda_2$, then*

$$\left\|(\bar{B}_{\mathbf{x}} + \lambda_1)^s f\right\|_{\mathcal{H}_1} \leq \left\|(\bar{B}_{\mathbf{x}} + \lambda_2)^s f\right\|_{\mathcal{H}_1}$$

Proof of Lemma A.1.6. Applying Proposition A.4.2 in [17] gives

$$\begin{aligned}\left\|(\bar{B}_{\mathbf{x}} + \lambda_1)^s f\right\|_{\mathcal{H}_1} &= \left\|(\bar{B}_{\mathbf{x}} + \lambda_1)^s(\bar{B}_{\mathbf{x}} + \lambda_2)^{-s}(\bar{B}_{\mathbf{x}} + \lambda_2)^s f\right\|_{\mathcal{H}_1} \\ &\leq ||(\bar{B}_{\mathbf{x}} + \lambda_1)(\bar{B}_{\mathbf{x}} + \lambda_2)^{-1}||^s \left\|(\bar{B}_{\mathbf{x}} + \lambda_2)^s f\right\|_{\mathcal{H}_1}\,.\end{aligned}$$

From the spectral Theorem we get, since $0 < \lambda_1 \leq \lambda_2$

$$||(\bar{B}_{\mathbf{x}} + \lambda_1)(\bar{B}_{\mathbf{x}} + \lambda_2)^{-1}|| \leq \sup_{0<t\leq 1}|(t + \lambda_1)(t + \lambda_2)^{-1}| \leq 1\,,$$

which immediately implies the result. □

A.2 A New Useful Inequality

After we had finished writing Chapters 2 and 3 of this thesis, a new bound for the operator product $(\bar{B}_{\mathbf{x}} + \lambda)^{-1}(\bar{B} + \lambda)$ was presented in [42], (with a proof not yet published at the time of submission of this thesis). We will use this result in Chapter 5 when presententing an adaptive estimator for the unkown target function, and we shall take the precise form of this estimate from the recent paper [62].

Proposition A.2.1 ([62]). *Let $x_1, ..., x_n$ be an iid sample, drawn according to ν on $\mathcal{X}$. Define*

$$\mathcal{B}_n(\lambda) := \left[1 + \left(\frac{2}{n\lambda} + \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}}\right)^2\right] \tag{A.2.1}$$

For any $\lambda > 0$, $\eta \in (0, 1]$, with probability at least $1 - \eta$ one has

$$\left\|(\bar{B}_{\mathbf{x}} + \lambda)^{-1}(\bar{B} + \lambda)\right\| \leq 8 \log^2(2\eta^{-1})\mathcal{B}_n(\lambda) \; . \tag{A.2.2}$$

Corollary A.2.2. *Let $\eta \in (0, 1)$. For $n \in \mathbb{N}$ let $\tilde{\lambda}_n$ be implicitly defined as the unique solution of $\mathcal{N}(\tilde{\lambda}_n) = n\tilde{\lambda}_n$. Then for any $\tilde{\lambda}_n \leq \lambda \leq 1$ one has*

$$\mathcal{B}_n(\lambda) \leq 26 \; .$$

In particular,

$$\left\|(\bar{B}_{\mathbf{x}} + \lambda)^{-1}(\bar{B} + \lambda)\right\| \leq 208 \log^2(2\eta^{-1}) \; ,$$

with probability at least $1 - \eta$.

Proof of Corollary A.2.2. Let $\tilde{\lambda}_n$ be defined via $\mathcal{N}(\tilde{\lambda}_n) = n\tilde{\lambda}_n$. Since $\mathcal{N}(\lambda)/\lambda$ is decreasing, we have for any $\lambda \geq \tilde{\lambda}_n$

$$\sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}} \leq \sqrt{\frac{\mathcal{N}(\tilde{\lambda}_n)}{n\tilde{\lambda}_n}} = 1 \; .$$

Furthermore, by Lemma 2.2.13, the effective dimesion is lower bounded by $\frac{1}{2}$, so by the inequality above

$$1 \geq \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}} \geq \frac{1}{2n\lambda} \quad \Longrightarrow \quad \frac{1}{n\lambda} \leq 2$$

for any $\lambda \geq \tilde{\lambda}_n$. Inserting these bounds into A.2.2 and noticing that $1 \leq 2\log(2\eta^{-1})$ for any $\eta \in (0, 1)$ leads to the conclusion. □

We shortly illustrate how Corollary A.2.2 and Proposition A.4.2 will be used to simplify previous bounds. Let $u \in [0, 1]$, $\tilde{\lambda}_n \leq \lambda$ as above and $f \in \mathcal{H}_1$. We have

$$\begin{aligned}
\left\|\bar{B}^u f\right\|_{\mathcal{H}_1} &= \left\|\bar{B}^u(\bar{B} + \lambda)^{-u}(\bar{B} + \lambda)^u(\bar{B}_{\mathbf{x}} + \lambda)^{-u}(B_{\mathbf{x}} + \lambda)^u f\right\|_{\mathcal{H}_1} \\
&\leq \left\|\bar{B}^u(\bar{B} + \lambda)^{-u}\right\| \left\|(\bar{B} + \lambda)^u(\bar{B}_{\mathbf{x}} + \lambda)^{-u}\right\| \left\|(\bar{B}_{\mathbf{x}} + \lambda)^u f\right\|_{\mathcal{H}_1} \\
&\leq 8 \log^{2u}(2\eta^{-1})\mathcal{B}_n(\lambda)^u \left\|(\bar{B}_{\mathbf{x}} + \lambda)^u f\right\|_{\mathcal{H}_1} \; ,
\end{aligned} \tag{A.2.3}$$

with probability at least $1 - \eta$, for any $\eta \in (0, 1)$. In particular, for any $\tilde{\lambda}_n \leq \lambda$ (with $\tilde{\lambda}_n$ as in Corollary A.2.2)

$$\left\|\bar{B}^u f\right\|_{\mathcal{H}_1} \leq 208^u \log^{2u}(2\eta^{-1}) \left\|(\bar{B}_{\mathbf{x}} + \lambda)^u f\right\|_{\mathcal{H}_1} \; , \tag{A.2.4}$$

with probability at least $1 - \eta$.

We remark that Corollary A.2.2 could be used to streamline our discussion in Chapters 2 and 3, since it improves our Proposition A.1.4. Instead of completely rewriting the presentation of our previous results we shall systematically demonstrate the power of this new approach only in Chapter 5.

A.3 Auxiliary Technical Lemmata

Lemma A.3.1. *Let X be a nonnegative real random variable such that the following holds:*

$$\mathbb{P}[X > F(t)] \leq t\,, \text{ for all } t \in (0,1]\,, \tag{A.3.1}$$

where F is a monotone non-increasing function $(0,1] \to \mathbb{R}_+$. Then

$$\mathbb{E}[X] \leq \int_0^1 F(u)du\,.$$

Proof. An intuitive, non-rigorous proof is as follows. Let G be the tail distribution function of X, then it is well known that $\mathbb{E}[X] = \int_{\mathbb{R}_+} G$. Now it seems clear that $\int_{\mathbb{R}_+} G = \int_0^1 G^{-1}$, where G^{-1} is the upper quantile function for X. Finally, F is an upper bound on G^{-1}.

Now for a rigorous proof, we can assume without loss of generality that F is left continuous: Replacing F by its left limit in all points of $(0,1]$ can only make it larger since it is non-increasing, hence (A.3.1) is still satisfied. Moreover, since a monotone function has an at most countable number of discontinuity points, this operation does not change the value of the integral $\int_0^1 F$. Define the following pseudo-inverse for $x \in \mathbb{R}_+$:

$$F^\dagger(x) := \inf\{t \in (0,1] : F(t) < x\}\,,$$

with the convention $\inf \emptyset = 1$. Denote $\widetilde{U} := F^\dagger(X)$. From the definition of $F^\dagger$ and the monotonicity of F it holds that $F^\dagger(x) < t \Rightarrow x > F(t)$ for all $(x,t) \in \mathbb{R}_+ \times (0,1]$. Hence, for any $t \in (0,1]$

$$\mathbb{P}[\widetilde{U} < t] \leq \mathbb{P}[X > F(t)] \leq t\,,$$

implying that for all $t \in [0,1]$, $\mathbb{P}[\widetilde{U} \leq t] \leq t$, i.e., $\widetilde{U}$ is stochastically larger than a uniform variable on $[0,1]$. Furthermore, by left continuity of F, one can readily check that $F(F^\dagger(x)) \geq x$ if $x \leq F(0)$. Since $\mathbb{P}[X > F(0)] = 0$, we can replace X by $\widetilde{X} := \min(X, F(0))$ without changing its distribution (nor that of $\widetilde{U}$). With this modification, it then holds that $F(\widetilde{U}) = F(F^\dagger(\widetilde{X})) \geq \widetilde{X}$. Hence,

$$\mathbb{E}[X] = \mathbb{E}[\widetilde{X}] \leq \mathbb{E}[F(\widetilde{U})] \leq \mathbb{E}[F(U)] = \int_0^1 F(u)du\,,$$

where U is a uniform variable on $[0,1]$, and the second equality holds since F is non-increasing. □

Corollary A.3.2. *Let X be a nonnegative random variable and $t_0 \in (0,1)$ such that the following holds:*

$$\mathbb{P}[X > a + b\log t^{-1}] \leq t\,, \text{ for all } t \in (t_0, 1]\,, \text{ and} \tag{A.3.2}$$

$$\mathbb{P}[X > a' + b'\log t^{-1}] \leq t\,, \text{ for all } t \in (0,1]\,, \tag{A.3.3}$$

where a, b, a', b' are nonnegative numbers. Then for any $1 \leq p \leq \frac{1}{2}\log t_0^{-1}$:

$$\mathbb{E}[X^p] \leq C_p\left(a^p + b^p\Gamma(p+1) + t_0\left((a')^p + 2(b'\log t_0^{-1})^p\right)\right)\,,$$

with $C_p := \max(2^{p-1}, 1)$.

Proof. Let $F(t) := \mathbf{1}\{t \in (t_0, 1]\}(a + b \log t^{-1}) + \mathbf{1}\{t \in (0, t_0]\}(a' + b' \log t^{-1})$. Then F is nonnegative, non-increasing on $(0, 1]$ and

$$\mathbb{P}[X^p > F^p(t)] \leq t$$

for all $t \in (0, 1]$. Applying Lemma A.3.1, we find

$$\mathbb{E}[X^p] \leq \int_0^{t_0} (a' + b' \log t^{-1})^p dt + \int_{t_0}^1 (a + b \log t^{-1})^p dt \,. \tag{A.3.4}$$

Using $(x+y)^p \leq C_p(x^p + y^p)$ for $x, y \geq 0$, where $C_p = \max(2^{p-1}, 1)$, we upper bound the second integral in (A.3.4) via

$$\int_{t_0}^1 (a + b \log t^{-1})^p dt \leq C_p \left(a^p + b^p \int_0^1 (\log t^{-1})^p dt \right) = C_p \left(a^p + b^p \Gamma(p+1) \right) .$$

Concerning the first integral in (A.3.4), we write similarly

$$\int_0^{t_0} (a' + b' \log t^{-1})^p dt \leq C_p \left(t_0 (a')^p + (b')^p \int_0^{t_0} (\log t^{-1})^p dt \right)$$
$$= C_p \left(t_0 (a')^p + (b')^p \Gamma(p+1, \log t_0^{-1}) \right) ,$$

by the change of variable $u = \log t^{-1}$, where Γ is the incomplete gamma function. We use the following coarse bound: It can be checked that $t \mapsto t^p e^{-\frac{t}{2}}$ is decreasing for $t \geq 2p$. Hence, putting $x := \log t_0^{-1}$,

$$\Gamma(p+1, x) = \int_x^\infty t^p e^{-t} dt \leq x^p e^{-\frac{x}{2}} \int_x^\infty e^{-\frac{t}{2}} dt = 2x^p e^{-x} = 2t_0 (\log t_0^{-1})^p ,$$

provided $x = \log t_0^{-1} \geq 2p$. Collecting all the above pieces we get the conclusion. □

Lemma A.3.3. *Let X be a nonnegative random variable with $\mathbb{P}[X > C \log^u(k\eta^{-1})] < \eta$ for any $\eta \in (0, 1]$. Then $\mathbb{E}[X] \leq \frac{C}{k} u \Gamma(u)$.*

Proof. Apply $\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > t] dt$. □

A.4 Some Operator Perturbation Inequalities

The estimate of the following proposition is crucial for proving the upper bound in case the source condition is of Hölder type r with $r > 1$. We remark that for $r > 1$ the function $t \mapsto t^r$ is not operator monotone. One might naively expect estimate $(A.4.1)$ to hold for a constant C given by the Lipschitz constant of the scalar function t^r. As shown in [6], this is false even for finite-dimensional positive matrices. The point of Proposition A.4.1 is that $(A.4.1)$ still holds for some larger constant depending on r and the upper bound of the spectrum. We do not expect this result to be particularly novel, but tracking down a proof in the literature proved elusive, not to mention that occasionally erroneous statements about related issues can be found. For this reason we here provide a self-contained proof for completeness sake.

Proposition A.4.1. *Let B_1, B_2 be two nonnegative self-adjoint operators on some Hilbert space with $||B_j|| \leq a$, $j = 1, 2$, for some $a > 0$. Assume B_1 and B_2 belong to the Schatten class $\mathcal{S}^p$ for $1 \leq p \leq \infty$.*

If $1 < r$, then

$$||B_1^r - B_2^r||_p \;\leq\; rCa^{r-1}\,||B_1 - B_2||_p \;\;, \tag{A.4.1}$$

where C is given in (A.4.2)*. This inequality also holds in operator norm for non-compact bounded (nonnegative and self-adjoint) B_j.*

Proof. We extend the proof of [37], given there in the case $r = 3/2$ in operator norm. We also restrict ourselves to the case $a = 1$. On $\mathcal{D} := \{z : \; |z| \leq 1\}$, we consider the functions $f(z) = (1-z)^r$ and $g(z) = (1-z)^{r-1}$. The proof is based on the power series expansions

$$f(z) = \sum_{n\geq 0} b_n z^n \quad \text{and} \quad g(z) = \sum_{n\geq 0} c_n z^n \;,$$

which converge absolutely on $\mathcal{D}$. To ensure absolute convergence on the boundary $|z| = 1$, notice that

$$c_n = \frac{1}{n!} g^{(n)}(0) = \frac{(-1)^n}{n!} \prod_{j=1}^{n} (r-j)\,,$$

so that all coefficients c_n for $n \geq r$ have the same sign $s := (-1)^{\lfloor r \rfloor}$ (if r is an integer these coefficients vanish without altering the argument below) implying for any $N > r$:

$$\begin{aligned}
\sum_{n=0}^{N} |c_n| &= \sum_{n=0}^{\lfloor r \rfloor} |c_n| + s \sum_{n=\lfloor r \rfloor + 1}^{N} c_n = \sum_{n=0}^{\lfloor r \rfloor} |c_n| + s \lim_{z \nearrow 1} \sum_{n=\lfloor r \rfloor + 1}^{N} c_n z^n \\
&\leq \sum_{n=0}^{\lfloor r \rfloor} |c_n| + s \lim_{z \nearrow 1} \Big(g(z) - \sum_{n=0}^{\lfloor r \rfloor} c_n \Big) \\
&= 2 \sum_{i=0}^{\lfloor r/2 \rfloor} |c_{\lfloor r \rfloor - 2i}| \;.
\end{aligned}$$

A bound for $\sum_n |b_n|$ can be derived analogously. Since $f(1 - B_j) = B_j^r$, we obtain

$$\|B_1^r - B_2^r\|_p \leq \sum_{n=0}^{\infty} |b_n| \; \|(I - B_1)^n - (I - B_2)^n\|_p \;.$$

Using the algebraic identity $T^{n+1} - S^{n+1} = T(T^n - S^n) + (T - S)S^n$, the triangle inequality and making use of $\|TS\|_p \leq \|T\| \; \|S\|_p$ for $S \in \mathcal{S}^p$, T bounded, the reader can easily convince himself by induction that

- for $j = 1, 2$, $B_j \in \mathcal{S}^p$ imply $(I - B_1)^n - (I - B_2)^n \in \mathcal{S}^p$ and
- $\|(I - B_1)^n - (I - B_2)^n\|_p \leq n \|B_1 - B_2\|_p$.

From $f'(z) = -rg(z)$ we have the relation $|b_n| = \frac{r}{n}|c_{n-1}|$, $n \geq 1$. Collecting all pieces leads to

$$\|B_1^r - B_2^r\|_p \leq \|B_1 - B_2\|_p \sum_{n=0}^{\infty} n|b_n| = rC\|B_1 - B_2\|_p \;,$$

with

$$C = \sum_{n=0}^{\infty} |c_n| \tag{A.4.2}$$

□

Proposition A.4.2 (Cordes Inequality,[5], Theorem IX.2.1-2)**.** *Let A, B be to self-adjoint, positive operators on a Hilbert space. Then for any $s \in [0,1]$:*

$$\|A^s B^s\| \leq \|AB\|^s \ . \tag{A.4.3}$$

Note: this result is stated for positive matrices in [5], but it is easy to check that the proof applies as well to positive operators on a Hilbert space.

A.5 General Reduction Scheme

Consider a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probability measures on a measurable space $(\mathbf{Z}, \mathcal{A})$, indexed by Θ. Additionally, let $d : \Theta \times \Theta \longrightarrow [0, \infty)$ be a (semi-) distance.

For two probability measures P_1, P_2 on some common measurable space $(\mathbf{Z}, \mathcal{A})$, we recall the definition of the *Kullback-Leibler divergence* between P_1 and P_2

$$\mathcal{K}(P_1, P_2) := \int_{\mathcal{X}} \log\left(\frac{dP_1}{dP_2}\right) dP_1 \ ,$$

if P_1 is absolutely continuous with respect to P_2. If P_1 is not absolutely continuous with respect to P_2, then $\mathcal{K}(P_1, P_2) := \infty$. One easily observes that

$$\mathcal{K}(P_1^{\otimes n}, P_2^{\otimes n}) \ = \ n\, \mathcal{K}(P_1, P_2) \ .$$

In order to obtain minimax lower bounds we briefly recall the general reduction scheme, presented in Chapter 2 of [87]. The main idea is to find N_ε parameters $\theta_1, ..., \theta_{N_\varepsilon} \in \Theta$, depending on $\varepsilon < \varepsilon_0$ for some $\varepsilon_0 > 0$, with $N_\varepsilon \to \infty$ as $\varepsilon \to 0$, such that any two of these parameters are ε-separated with respect to the distance d, but that the associated distributions $P_{\theta_j} =: P_j \in \mathcal{P}$ have small Kullback-Leibler divergence to each other and are therefore statistically close. It is then clear that

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathcal{P}} \mathbb{E}_\theta[d^p(\hat{\theta}, \theta)]^{\frac{1}{p}} \geq \varepsilon \inf_{\hat{\theta}} \sup_{\theta \in \mathcal{P}} \mathbb{P}_\theta[d(\hat{\theta}, \theta) \geq \varepsilon] \geq \varepsilon \inf_{\hat{\theta}} \max_{1 \leq j \leq N_\varepsilon} \mathbb{P}_j[d(\hat{\theta}, \theta_j) \geq \varepsilon] \ , \tag{A.5.1}$$

where the infimum is taken over all estimators $\hat{\theta}$ of θ. The above RHS is then lower bounded through the following proposition which is a consequence of Fano's lemma, see [87], Theorem 2.5:

Proposition A.5.1. *Assume that $N \geq 2$ and suppose that Θ contains N+1 elements $\theta_0, ..., \theta_N$ such that:*

(i) For some $\varepsilon > 0$, and for any $0 \leq i < j \leq N$, $d(\theta_i, \theta_j) \geq 2\varepsilon$;

(ii) For any $j = 1, ..., N$, P_j is absolutely continuous with respect to P_0, and

$$\frac{1}{N} \sum_{j=1}^{N} \mathcal{K}(P_j, P_0) \leq \omega \ \log(N) \ , \tag{A.5.2}$$

for some $0 < \omega < 1/8$.

Then

$$\inf_{\hat{\theta}} \max_{1 \leq j \leq N} P_j(\ d(\hat{\theta}, \theta_j) \geq \varepsilon \ ) \ \geq \ \frac{\sqrt{N}}{1 + \sqrt{N}} \left(1 - 2\omega - \sqrt{\frac{2\omega}{\log(N)}} \right) > 0 \ ,$$

where the infimum is taken over all estimators $\hat{\theta}$ of θ.

Appendix B

LocalNysation: Combining Localized Kernel Regression and Nyström Subsampling

Abstract

We consider a localized approach in the well-established setting of reproducing kernel learning under random design. The input space $\mathcal{X}$ is partitioned into local disjoint subsets $\mathcal{X}_j$ $(j = 1, ..., m)$ equipped with a local reproducing kernel K_j. It is then straightforward to define local KRR estimates. Our first main contribution is in showing that minimax optimal rates of convergence are preserved if the number m of partitions grows sufficiently slowly with the sample size, under locally different degrees on smoothness assumptions on the regression function. As a byproduct, we show that low smoothness on exceptional sets of small probability does not contribute, leading to a faster rate of convergence. Our second contribution lies in showing that the partitioning approach for KRR can be efficiently combined with local Nyström subsampling, improving computational cost twofold. If the number of locally subsampled inputs grows sufficiently fast with the sample size, minimax optimal rates of convergence are maintained.

B.1 Introduction and Motivation

B.1.1 Kernel Regression

We are concerned with the classical regression learning problem, where we observe training data $\mathcal{D} := (X_i, Y_i)_{i=1,...,n}$, assumed to be an i.i.d. sample from a distribution ρ over $\mathcal{X} \times \mathbb{R}$ (ν will denote the marginal distribution of ρ), and the goal is to estimate the regression function $f^*(x) := \mathbb{E}[Y|X = x]$. We consider the well-established setting of (reproducing) kernel learning: we assume a positive semi-definite kernel $K(.,.)$ has been defined on $\mathcal{X}$, with associated canonical feature mapping $\Phi : \mathcal{X} \to \mathcal{H}$ into a corresponding reproducing kernel Hilbert space $\mathcal{H}$. A classical approach is kernel ridge regression, depending on a regularization parameter $\lambda > 0$, giving rise to the estimate $\widehat{f}^{\lambda}$. In this paper, we shall focus only on KRR, although our results could be extended to a much larger class of general spectral regularization methods.

A common goal of learning theory is to give upper bounds for the convergence of $\widehat{f}^{\lambda_n}$ to f^* (where the regularization parameter is tuned according to sample size), and derive rates of convergence as $n \to \infty$ under appropriate assumptions on the "regularity" of f^*. In this paper, the notion of convergence we will consider is the usual squared $L^2(\nu)$ distance with respect to the sampling distribution, which is equal to the excess risk with respect to Bayes when using the squared loss, i.e.

$$\|\widehat{f}_{\lambda} - f^*\|_{2,\nu}^2 = \mathbb{E}\big[\big(Y - \widehat{f}_{\lambda}(X)\big)^2\big] - \min_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}\big[\big(Y - f(X)\big)^2\big] .$$

More precisely, we are interested in bounding the averaged above error over the draw of the training data (this is also called Mean Integrated Squared Error or MISE in the statistics literature).

B.1.2 Upper bounds on rates of convergence and optimality

We shall derive upper rates of convergence for algorithms speeding up the more usual single machine version of KRR. We recall that defining such a rate requires to fix a set $\mathcal{M}$ of data generating distributions. We shall not work in a traditional framework where regularity of the target function is assumed to be of Sobolev type, $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and the marginal ν is assumed to have a bounded density with respect to Lebesgue measure. As is well known, this is too strong a restriction in a general statistical learning context. In a distribution free spirit, all these highly specific features are replaced by properties of the (uncentered) covariance operator $T_\nu = \mathbb{E}_\nu[\Phi(X) \otimes \Phi(X)^*]$ of the kernel feature mapping $\Phi(X)$. Letting $(\mu_{\nu,i}, \psi_{\nu,i})_{i\geq 1}$ be an eigendecomposition of T_ν , one introduces, for $r, R > 0$, the class

$$\Omega_\nu(r, R) := \left\{ f \in \mathcal{H} : \sum_{i\geq 1} \mu_{\nu,i}^r f_i^2 \leq R^2 \right\} = T_\nu^r \mathcal{B}_{\mathcal{H}}(R) \,, \tag{B.1.1}$$

where $\mathcal{B}_{\mathcal{H}}(R)$ denotes the ball of radius R in $\mathcal{H}$ and $f_i := \langle f, \psi_{\nu,i} \rangle$ are the coefficients of f in the eigenbasis. In the classical Sobolev type setting, $T = T_\nu$ basically is the inverse of the Laplacian (obeying, e.g., Dirichlet boundary conditions on $\partial\mathcal{X}$) and the condition in (2.5.8) becomes a condition on the decay of classical Fourier coefficients of f, which measures smoothness in a classical sense.

In the above more general form, (B.1.1) encodes the properties of the distribution of the feature map $\Phi(X)$. If the target function f^* is well approximated in the eigenbasis in the sense that its coefficients decay fast enough in comparison to the eigenvalues, it is considered as regular in this geometry (higher regularity corresponds to higher values of r and/or lower values of R). This type of regularity class, also called *source condition*, has been considered in a statistical learning context by [23], and [24] have established upper bounds for the performance of kernel ridge regression $\widehat{f}^\lambda$ over such classes; this has been extended to other types of kernel regularization methods by [17, 20, 29]. These bounds rely on tools introduced in the seminal work of [94], and depend in particular on the notion of *effective dimension* of the data with respect to the regularization parameter λ , defined as

$$\mathcal{N}_\nu(T, \lambda) := \mathcal{N}(T_\nu, \lambda) := \mathrm{Tr}\left[(T_\nu + \lambda)^{-1} T_\nu\right] \,. \tag{B.1.2}$$

In this paper we shall consider a model class $\mathcal{M} := \mathcal{M}(\theta, r, b)$ (see Section 2, Assumption B.2.1, B.2.2 and also B.2.4 for its precise definition) depending on a fixed degree r of regularity in the above general sense and the asymptotic behavior of the effective dimension, parametrized by a number $b \geq 1$ via $\mathcal{N}(T_\nu, \lambda) \asymp \lambda^{-1/b}$, where $\asymp$ stands for upper and lower bounded up to a constant.

We recall that a sequence a_n (tending to zero as $n \to \infty$) is called *an upper rate of convergence* for the sequence of estimated solutions $(\hat{f}_{\mathcal{D}}^{\lambda_n})_n$ over the family of data generating distributions $\mathcal{M}$, iff

$$\limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\| f^* - \hat{f}_{\mathcal{D}}^{\lambda_n} \right\|_{L^2(\nu)}^2 \right]^{\frac{1}{2}}}{a_n} < C \,. \tag{B.1.3}$$

If there is also a corresponding lower bound, the sequence a_n is said to be minimax optimal. We recall that, for the single machine problem, minimax optimal rates have been obtained for similar looking model classes $\mathcal{M}$. The first comprehensive result in this direction was established by [24]; [17] gives a sharp estimate of the convergence rate in this case including the dependence on the parameters R and noise variance σ, namely $\mathcal{O}\Big(R^2\big(\frac{\sigma^2}{R^2 n}\big)^{\frac{2r+1}{2r+1+\frac{1}{b}}}\Big)$. There is, however, a small caveat. While upper rates of convergence are known to only depend on the asymptotics of the effective dimension, lower bounds, to the best of our knowledge, have up to now only been established under an additional assumption on the eigenvalues of the covariance operator, namely $\mu_{\nu,i} \asymp i^{-b}$, for some $b > 1$. While this implies the estimate on the effective dimension assumed in this paper, the converse implication is obviously wrong. Therefore, strictly speaking, minimax optimal rates are not known in the case considered in this paper. Furthermore, the partitioning approach considered in the present paper imposes additional constraints on the data generating distribution via Assumption B.2.4, as we shall amplify. For this reason we shall focus on upper rates of convergence only and leave the important and interesting question of minimax optimality for a longer version of this paper.

B.1.3 Large Scale Problems: Localization and Subsampling

Kernel-based methods for solving non-parametric regression problems are attractive because they attain asymptotically minimax optimal rates of convergence. But it is well known that these methods scale poorly when massive datasets are involved. Large training sets give rise to large computational and storage costs. For example, computing a kernel ridge regression estimate needs inversion of a $n \times n$-matrix, with n the sample size. This requires $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory, which becomes prohibitive for large sample sizes. For this reason, various methods have been developed for saving computation time and memory requirements. Among them are e.g. *distributed learning* [16], [95], [58] and *early-stopping* [93],[72], [12], [13]. During the last years, a huge amount of research effort was devoted to finding *low-rank approximations* of the kernel matrix, both from an algorithmic and an inferential perspective (providing statistical guarantees). A popular instance is Nyström sampling see e.g. [90], [2], [77] , which we shall shortly review in Section 3.

The common feature of all these methods is to replace the theoretically optimal approximation obtained by a spectral decomposition (which requires time at least $\mathcal{O}(n^2)$) by a less ambitious suitable low rank approximation of the kernel matrix via column sampling, reducing run time to $\mathcal{O}(np^2)$ where p denotes the rank of the approximation. Clearly the rules of the game are to choose p as small as possible while maintaining minimax optimality of convergence rates and to explicitely determine this p as a function of the sample size n, keeping track of the source condition and the rate of eigenvalue decay, entering the estimate via the effective dimension.

Another line of research with computational benefits is devoted to so called *partition-based* or *localized* approaches [65, 84]: Based on a (disjoint) partition of the input space $\mathcal{X} = \bigcup_{j=1}^m \mathcal{X}_j$, the sample $\mathcal{D}$ is split according to this partition into m subsamples $\mathcal{D}_1, \cdots \mathcal{D}_m$. On each local set $\mathcal{X}_j$, a separate reproducing kernel K_j is defined, giving rise to a local RKHS $\mathcal{H}_j$. It is then straightforward to define local KRR estimates $\widehat{f}_j^{\lambda}$ (based on only using a fraction n_j of samples), resulting in a global estimate $\widehat{f}_{\mathcal{D}}^{\lambda} = \oplus_{j=1}^m \widehat{f}_j^{\lambda}$, belonging to the RKHS with kernel K, which is constructed from the K_j and is adapted to the direct sum decomposition $\mathcal{H} = \oplus_{j=1}^m \mathcal{H}_j$. It is well established that, using this approach, prediction for a new input $x \in \mathcal{X}$ is much faster, because one only has to identify the local space $\mathcal{X}_j$ to which x belongs and

to use the local estimator $\widehat{f}_j^{\lambda}$.

A more subtle point concerns the regularity assumption for the unknown $f^* = \oplus_{j=1}^m f_j^*$ which we require to belong to $\mathcal{H}$. Assuming local regularities $f_j^* \in T_j^{r_j}\mathcal{B}_{\mathcal{H}_j}(R)$, $r_1 \leq ... \leq r_m$, where T_j denotes the local covariance operator associated with K_j, the global smoothness (and thus the rate of convergence) is determined by only the lowest one:

$$f^* = \oplus_{j=1}^m T_j^{r_j} g_j = \oplus_{j=1}^m T_j^{r_1} \tilde{g}_j = T^{r_1}\tilde{g} \ , \quad \tilde{g} = \oplus \tilde{g}_j, \quad \tilde{g}_j = T_j^{r_j - r_1} g_j \ , \tag{B.1.4}$$

meaning that $f^* \in T^{r_1}\mathcal{B}_{\mathcal{H}}(R)$, since $T = \oplus_{j=1}^m T_j$.

Basic problems are the the choice of the regularization parameter λ on the subsamples (depending on the global sample size n) and, most importantly, the proper choice of m since choosing m too large gives a suboptimal rate of convergence in the limit $n \to \infty$.

First results establishing learning rates using a KRR partition-based approach for smoothness parameter $r = 0$ and polynomially decaying eigenvalues are given in [84]. The authors claim to prove optimal rates requiring the existence of sufficiently high moments (in $L^2(\nu)$) of the eigenfunctions of their local covariance operators, uniformly over all subsets, in the limit $n \to \infty$. This is a strong assumption. Moreover, while the decay rate of the eigenvalues can be determined by the smoothness of K (see e.g. [36] and references therein) it is a widely open question which (general) properties of the kernel imply such assumptions on the eigenfunctions.

The paper [65] considers localized SVMs, localized tuned Gaussian kernels and a corresponding direct sum decomposition, where a global smoothness assumption is introduced in terms of a scale of Besov spaces. Instead of using the effective dimension $\mathcal{N}(T_\nu, \lambda)$ - recall that in [65] the regularization parameter is not a spectral parameter in the resolvent $(T + \lambda)^{-1}$ but rather a coupling constant for the penalty term - the authors of [65] use entropy numbers, obtaining minimax optimal rates up to a small error (concerning the prefactor).

B.1.4 Contribution

Our main contribution is in showing that the partitioning approach for KRR can be efficiently combined with Nyström subsampling, improving computational cost twofold. On the way we improve results from [84], [65] and [77]. We somewhat amplify the result in [77] by adding an explicit asymptotic result on the number l_n of subsampled points and providing an estimate in expectation. Compared with [84], we remove the assumptions on the eigenfunctions of the covariance operators which are difficult to prove, and we allow locally different degrees of smoothness. Our results on upper rates of convergence only depend on the effective dimension. Compared with [65], our more general smoothness assumptions are distribution free in spirit, and we go beyond Gaussian kernels (and allow more general input spaces than open subsets of $\mathbb{R}^n$). If the number of subsets is not too large ($m = n^\alpha$ with an explicit bound on $\alpha < \frac{1}{2}$), we obtain an asymptotic result on upper rates of convergence which we expect to be minimax optimal over appropriate model classes of marginals.

An important aspect of our approach is the observation that under appropriate conditions on the proba-

bility of subsamples - which come quite naturally in the partitioning approach - one can actually do *better* than the naively expected minimax optimal rate allows: If low smoothness r_l only occurs on subsets of low probability (while most subsets have larger smoothness r_h), then an upper rate of convergence only depends on r_h and not on r_l. At first sight this seems to contradict equation (B.1.4) which sets the degree of global smoothness equal to r_l (it does not, see our Discussion). To the best of our knowledge this effect of having local smoothness available has never been analysed before.

B.2 The Partitioning Approach

We say that a family $\{\mathcal{X}_1, ..., \mathcal{X}_m\}$ of nonempty disjoint subsets of $\mathcal{X}$ is a partition of $\mathcal{X}$, if $\mathcal{X} = \bigcup_{j=1}^m \mathcal{X}_j$. Given a probability measure ν on $\mathcal{X}$, let $p_j = \nu(\mathcal{X}_j)$. We endow each $\mathcal{X}_j$ with a probability measure by restricting the conditional probability $\nu_j(A) := \nu(A|\mathcal{X}_j) = p_j^{-1}\nu(A \cap \mathcal{X}_j)$ to the Borel sigma algebra on $\mathcal{X}_j$.

We further assume that $\mathcal{H}_j$ is a (separable) RKHS, equipped with a measurable positive semi-definite real-valued kernel K_j on each $\mathcal{X}_j$, bounded by κ_j. We extend a function $f \in \mathcal{H}_j$ to a function $\hat{f} : \mathcal{X} \longrightarrow \mathbb{R}$ by setting $\hat{f}(x) = f(x)$ for any $x \in \mathcal{X}_j$ and $\hat{f}(x) = 0$ else. In particular, $\hat{K}_j$ denotes the kernel extended to $\mathcal{X}$, explicitely given by $\hat{K}_j(x, x') = K_j(x, x')$ for any $x, x' \in \mathcal{X}_j$ and zero else. Then the space $\hat{\mathcal{H}}_j := \{\hat{f} : f \in \mathcal{H}_j\}$ equipped with the norm $||\hat{f}||_{\hat{\mathcal{H}}_j} = ||f||_{\mathcal{H}_j}$ is an RKHS of functions on $\mathcal{X}$ with kernel $\hat{K}_j$. Finally, for $p_1, ..., p_m \in \mathbb{R}_+$, the direct sum

$$\mathcal{H} := \bigoplus_{j=1}^m \hat{\mathcal{H}}_j = \{\ \hat{f} = \sum_{j=1}^m \hat{f}_j\ :\ \hat{f}_j \in \hat{\mathcal{H}}_j\ \}$$

with norm $||\hat{f}||_{\mathcal{H}}^2 = \sum_{j=1}^m p_j ||\hat{f}_j||_{\hat{\mathcal{H}}_j}^2$ is again an RKHS for which

$$K(x, x') = \sum_{j=1}^m p_j^{-1} \hat{K}_j(x, x')\ , \qquad x, x' \in \mathcal{X}\ ,$$

is the reproducing kernel, see [65].

Given training data $\mathcal{D} = \{x_i, y_i\}_{i \in [n]}$, we let $I_j = \{i \in [n] : x_i \in \mathcal{X}_j\}$, with $|I_j| = n_j$. We split $\mathcal{D}$ according to the above partition, i.e. we let $\mathcal{D}_j = \{x_i, y_i\}_{i \in I_j}$. We further let $\mathbf{x}_j = (x_i)_{i \in I_j}$, $\mathbf{y}_j = (y_i)_{i \in I_j}$.

Fixing a regularization parameter $\lambda > 0$, we compute for each $\mathcal{D}_j$ a local estimator

$$\hat{f}_{\mathcal{D}_j}^{\lambda} := \frac{1}{n_j} \sum_{i \in I_j} \alpha_j^{(i)}(\lambda) \hat{K}_j(x_i, \cdot) \in \hat{\mathcal{H}}_j\ ,$$

where $\alpha_j = (\alpha_j^{(1)}, ..., \alpha_j^{(n_j)}) \in \mathbb{R}^{n_j}$ is given by $\alpha_j = (\frac{1}{n_j}\mathbb{K}_j + \lambda)^{-1}\mathbf{y}_j$ and with $\mathbb{K}_j$ the kernel matrix associated to $\mathcal{D}_j$. Finally, the overall estimator is defined by

$$\hat{f}_{\mathcal{D}}^{\lambda} := \sum_{j=1}^m \hat{f}_{\mathcal{D}_j}^{\lambda}\ , \tag{B.2.1}$$

which by construction belongs $\mathcal{H}$ and decomposes according to the direct sum $\hat{\mathcal{H}}_1 \oplus ... \oplus \hat{\mathcal{H}}_m$.

Assumption B.2.1. *We assume:*

1. *The regression function f^* belongs to $\mathcal{H}$ and thus has a unique representation $f^* = \oplus_{j=1}^m f_j^*$, with $f_j^* \in \hat{\mathcal{H}}_j$.*

2. *The sampling is random i.i.d., where each observation point (X_i, Y_i) follows the model $Y = f(X)+\varepsilon$, and the noise satisfies the following Bernstein-type assumption: For any integer $k \geq 2$ and some $\sigma > 0$ and $M > 0$:*

$$\mathbb{E}[\, \varepsilon^k \mid X \,] \leq \frac{1}{2} k!\, \sigma^2 M^{k-2}\ \nu - \text{a.s.}\ . \qquad (\ \textbf{Bernstein}(M, \sigma))$$

3. *The local effective dimensions obey*

$$m \sum_{j=1}^m p_j \mathcal{N}_{\nu_j}(\bar{T}_j, \lambda) = \mathcal{O}(\mathcal{N}_\nu(\bar{T}, \lambda))\ .$$

4. *The global effective dimension satisfies $\mathcal{N}_\nu(\bar{T}, \lambda) \lesssim \lambda^{-1/b}$ for some $b \geq 1$.*

Assumption B.2.2. *We assume*

1. *The global regularity of the regression function is measured in terms of a source condition:*

$$f^* \in \Omega_\nu(r, R)\ , \qquad 0 < r \leq \frac{1}{2}\ , R < \infty\ , \qquad (\ \textbf{SC}(r, R))$$

   where $\Omega_\nu(r, R)$ is defined in (2.5.8).

2. *Given $\theta = (M, \sigma, R) \in \mathbb{R}_+^3$, the class $\mathcal{M} := \mathcal{M}(\theta, r, b)$ consists of all distributions ρ with X-marginal ν and conditional distribution of Y given X satisfying* $\textbf{Bernstein}(M, \sigma)$ *for the deviations and* ($\textbf{SC}(r, R)$) *for the mean, with ν satisfying Assumption B.2.1, 3. and 4. .*

B.2.1 Error Bounds

Granted Assumptions B.2.1 and B.2.2, we establish that the estimator $\hat{f}_{\mathcal{D}}^{\lambda}$ given in (B.2.1) satisfies an upper rate of convergence which we expect to be minimax optimal, provided that the cardinality $m = m_n$ of partitions grows sufficiently slowly with n.

Theorem B.2.3. *Assume the number $m = m_n$ of partitions satisfies*

$$m_n \leq n^\alpha\ , \quad \alpha < \frac{2br}{2br + b + 1} \qquad (B.2.2)$$

and $n_j = \lfloor \frac{n}{m_n} \rfloor$. If $r \in (0, \frac{1}{2}]$, then there exists a choice $(\lambda_n)_n$ such that the sequence

$$a_n := R \left(\frac{\sigma^2}{R^2 n} \right)^{\frac{b(r+\frac{1}{2})}{2br+b+1}}\ . \qquad (B.2.3)$$

is an upper rate of convergence for the sequence of estimated solutions $(\hat{f}_{\mathcal{D}}^{\lambda_n})_n$ *over the family of models* $\mathcal{M}$ *- defined by requiring Assumption B.2.2 for all* $m = m_n$ *for each* n *sufficiently large -, i.e.*

$$\limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|f^* - \hat{f}_{\mathcal{D}}^{\lambda_n}\right\|_{L^2(\nu)}^2\right]^{\frac{1}{2}}}{a_n} < C\,, \tag{B.2.4}$$

where C *does not depend on the model parameters* σ, M, R*.*

We recall that it was established in [17] that the rate in (2.3.2) is minimax optimal for the class of data-generating distributions $\tilde{\mathcal{M}}$ obtained by replacing the conditions ν *satisfies Assumption B.2.1, 3. and 4.* by the slightly different assumption T_ν *has polynomial eigenvalue decay* $\mu_{\nu,i} \asymp i^{-b}$. Parametrizing $\mathcal{M}$ only by the asymptotics of the effective dimension, specified by b, tends to make $\mathcal{M}$ larger, while imposing in addition Assumption B.2.1 3. tends to make $\mathcal{M}$ smaller. Proving minimax optimality for our class $\mathcal{M}$ requires proving new lower bounds.

B.2.2 Improved Error Bound

Sometimes we can do even better: Assume that there is an exceptional set E_m of indices such that the smoothness of f^* is low on each set $\mathcal{X}_j$, $j \in E_m$ and higher on each $\mathcal{X}_j$, $j \in E_m^c$. For ease of reading we shall only analyze the most simple case given by:

Assumption B.2.4. *There are* $r_l, r_h \in (0, \frac{1}{2}]$*, with* $r_l < r_h$ *(corresponding to low smoothness and high smoothness) and there are* $R_l > 0$*,* $R_h > 0$ *such that*

$$||T_j^{-r_l} f_j^*||_{\mathcal{H}_j} \leq R_l\,, \quad \forall j \in E_m\,, \quad ||T_j^{-r_h} f_j^*||_{\mathcal{H}_j} \leq R_h\,, \quad \forall j \in E_m^c$$

and

$$\left(\sum_{j\in E_m} p_j\right) \leq \frac{R_h^2}{R_l^2}\left(\frac{1}{m}\right)^{1-\frac{r_l}{r_h}}\,.$$

Thus, by equation (B.1.4), global smoothness is given by the small number r_l, while local smoothness on the complement of the exceptional set is higher. We emphasize that this is an additional assumption on the sampling distribution which restricts the class of models $\mathcal{M}$ to a subclass $\mathcal{M}'$. Assumption B.2.4 then ensures that the probability of the exceptional set is so small that the error bound will actually be governed by the higher smoothness r_h, leading to a faster rate of convergence over the subclass $\mathcal{M}'$. More precisely,

Theorem B.2.5. *If the cardinality* $m = m_n$ *of partitions satisfies*

$$m_n = \left(\frac{R_h^2 n}{\sigma^2}\right)^{\alpha}\,, \quad \alpha < \frac{2br_h}{2br_h + b + 1} \tag{B.2.5}$$

and if $n_j \sim \frac{n}{m}$*, then there exists a choice* $(\lambda_n)_n$ *such that the sequence*

$$a_n := R_h\left(\frac{\sigma^2}{R_h^2 n}\right)^{\frac{b(r_h+\frac{1}{2})}{2br_h+b+1}}\,. \tag{B.2.6}$$

is an upper rate of convergence for the sequence of estimated solutions $(\hat{f}_{\mathcal{D}}^{\lambda_n})_n$ *over the subclass of models* $\mathcal{M}'$ *- defined by requiring Assumption B.2.4 for all* $m = m_n$ *for each* n *sufficiently large -, i.e.*

$$\limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}'} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|f^* - \hat{f}_{\mathcal{D}}^{\lambda_n}\right\|_{L^2(\nu)}^2\right]^{\frac{1}{2}}}{a_n} < C\,, \tag{B.2.7}$$

where C *does not depend on the model parameters* σ, M, R_h*.*

Note that a matching lower bound is proved in [17] only for the class $\tilde{\mathcal{M}}$.

B.3 KRR Nyström Subsampling

In this section, we recall the popularKRR Nyström subsampling method. For simplicity, we restrict ourselves to so called *Plain Nyström*, which works as follows: Given a training set $x_1, ..., x_n$ of random inputs, we sample uniformly at random without replacement $l \leq n$ points $\tilde{x}_1, ..., \tilde{x}_l$. Now the crucial idea is to seek for an estimator for the unknown f^* in a reduced space

$$\mathcal{H}_l = \{\ f\ :\ f = \sum_{j=1}^{l} \alpha_j K(\tilde{x}_j, \cdot)\ ,\ \alpha \in \mathbb{R}^l\ \}\,.$$

In [77] it is shown that the solution of the minimization problem

$$\min_{f\in\mathcal{H}_l}\ \frac{1}{n}\sum_{j=1}^{n}(f(x_j) - y_j)^2 + \lambda||f||_{\mathcal{H}_l}^2$$

is given by

$$\hat{f}_{n,l}^{\lambda} = \frac{1}{n}\sum_{j=1}^{l}\alpha_j K(\tilde{x}_j, \cdot)\ , \quad \alpha = (\frac{1}{n}\mathbb{K}_{nl}^*\mathbb{K}_{nl} + \lambda\mathbb{K}_{ll})^{\dagger}\ \mathbb{K}_{nl}^*\mathbf{y}\ , \tag{B.3.1}$$

where $(\mathbb{K}_{nl})_{ij} = K(x_i, \tilde{x}_j)$, $(\mathbb{K}_{ll})_{kj} = K(\tilde{x}_k, \tilde{x}_j)$, $i = 1, ..., n$, $k, j = 1, ..., l$ and $A^{\dagger}$ denotes the generalized inverse of a matrix A.

Clearly, one aims at minimizing the number l of subsamples needed for preserving (the expected) minimax optimality. We amplify the results in [77] by explicitly computing how l needs to grow when the total number of samples n tends to infinity.

We consider the setting of Section B.2 with $m = 1$. Granted Assumption B.2.1 and AssumptionB.2.2, one has:

Theorem B.3.1. *If the number* $l = l_n$ *of subsampled points satisfies*

$$l_n \geq n^{\beta}\ , \quad \beta > \frac{b+1}{2br+b+1}\ , \tag{B.3.2}$$

and if $r \in [0, \frac{1}{2}]$ *then there exists a choice* $(\lambda_n)_n$ *such that the sequence* $(a_n)_n$ *given in* (2.3.2) *is an upper rate of convergence for the sequence of estimated solutions* $(\hat{f}_{n,l_n}^{\lambda_n})_n$ *over the family of models* $\mathcal{M}$ *- defined*

by dropping the condition Assumption B.2.1,3. on the data generating distribution - i.e.

$$\limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|f^* - \hat{f}^{\lambda_n}_{n,l_n}\right\|^2_{L^2(\nu)}\right]^{\frac{1}{2}}}{a_n} < C\,,$$

for some $C < \infty$, not depending on the model parameters σ, M, R.

We remark that, as above, it was established in [17] that the rate in (2.3.2) is minimax optimal for the class of data-generating distributions $\tilde{\mathcal{M}}$.

B.4 LocalNysation

In this section we establish, that upper rates of convergence are preserved if the number of partitions is not too large and if locally the number of subsampled points is large enough. For simplicity we assume that the local sample size is roughly the same on each partition, i.e satisfies $n_j = \lfloor \frac{n}{m_n} \rfloor$ and that the number $l = l_n$ of subsample points also is equal on each subsample.

For $j = 1, ..., m$, and $1 \leq l \leq \frac{n}{m}$ let $\tilde{I}_{j,l} := \{i_{j,1}, ..., i_{j,l}\} \subseteq I_j$, with I_j as above ($\tilde{I}_{j,l}$ denotes the set of indices of subsampled inputs on each $\mathcal{X}_j$). For each subsample $\mathcal{D}_j$, with a regularization parameter $\lambda > 0$, we compute a local estimator

$$\hat{f}^{\lambda}_{\mathcal{D}_j} := \frac{m}{n} \sum_{i\in\tilde{I}_{j,l}} \alpha^{(i)}_j(\lambda)\hat{K}_j(x_i, \cdot) \in \hat{\mathcal{H}}_{j,l}\,,$$

where $\alpha_j \in \mathbb{R}^{\frac{n}{m}}$ is given in (B.3.1), with n replaced by $\frac{n}{m}$. The overall estimator is constructed as above and defined by

$$\hat{f}^{\lambda}_{\mathcal{D}} := \sum_{j=1}^{m} \hat{f}^{\lambda}_{\mathcal{D}_j}\,, \tag{B.4.1}$$

which by construction decomposes according to the direct sum $\mathcal{H} = \hat{\mathcal{H}}_1 \oplus ... \oplus \hat{\mathcal{H}}_m$.

Theorem B.4.1. *Let $r \in (0, \frac{1}{2}]$. If the number $m = m_n$ of partitions satisfies*

$$m_n \leq n^{\alpha}\,, \quad \alpha < \frac{2br}{2br + b + 1}\,, \tag{B.4.2}$$

and if the number $l = l_n$ of subsampled points on each local set satisfies

$$\frac{n^{\beta}}{m_n} \leq l_n \leq \frac{n}{m_n}\,, \quad \beta > \alpha + \frac{b+1}{2br + b + 1}\,, \tag{B.4.3}$$

then there exists a choice $(\lambda_n)_n$ such that the sequence $(a_n)_n$ given in (2.3.2) is an upper rate of convergence for the sequence of estimated solutions $(\hat{f}^{\lambda_n}_{\mathcal{D}})_n$ over the family of models $\mathcal{M}$ - defined by requiring the conditions in Assumption B.2.2 for each $m = m_n$ for n sufficiently large - , i.e.

$$\limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|f^* - \hat{f}^{\lambda_n}_{\mathcal{D}}\right\|^2_{\mathcal{H}}\right]^{\frac{1}{2}}}{a_n} < C\,, \tag{B.4.4}$$

Table B.1: Computational Cost

whole KRR	localized KRR	Nyström	localnysed
$\mathcal{O}(n^3)$	$\mathcal{O}\left((\frac{n}{m})^3\right)$	$\mathcal{O}(nl^2+l^3)$	$\mathcal{O}(\frac{n}{m}l^2+l^3)$
	$1 \leq m \leq n^\alpha$	$n^\beta \leq l \leq n$	$\frac{n^\beta}{m} \leq l \leq \frac{n}{m}$

where C does not depend on the model parameters σ, M, R.

Clearly, as in Theorem B.2.5, a version of the above result still holds if global smoothness is violated on an exceptional set E_m of small probability as amplified in Assumption B.2.4, thus passing from the model class $\mathcal{M}$ to a smaller model class $\mathcal{M}'$. We leave a precise formulation (and its proof) to the reader.

B.5 Conclusion

We have shown that the twofold effect of partitioning and subsampling may substantially reduce computational cost (see Table B.1), if the number of local sets grows sufficiently slowly and if the number of subsampled inputs grows sufficiently large with the sample size. In both cases we were able to improve or amplify the existing results. Furthermore, we derived a rigorous version of the principle *In partitioning, low smoothness on exceptional sets of small probability does not affect convergence.* This is not in contradiction to the known results on minimax optimality, since this phenomenon only occurs for data generating distributions varying over a restricted set $\mathcal{M}'$. We remark that, based on the simulations in the thesis [33], it already was observed there that simulations tend to be better than global smoothness predicts. The thesis [33] suggested that this effect could be due to violation of global smoothness on subsets of measure zero. While this certainly is possible, it seems much more probable that global smoothness could be violated on larger subsets of small probability (but possibly different from zero). Thus our theorems might be an additional explication of the properties of numerical simulations observed in [33].

Supplementary Material
LocalNysation: Combining Localized Kernel Regression and Nyström Subsampling

For ease of reading we make use of the following conventions:

- we are interested in a precise dependence of multiplicative constants on the parameters σ, M, R, m, n and p
- the dependence of multiplicative constants on various other parameters, including the kernel parameter κ, the parameters arising from the regularization method, $b > 1$, $r > 0$, etc. will (generally) be omitted
- the value of C might change from line to line
- the expression "for n sufficiently large" means that the statement holds for $n \geq n_0$, with n_0 potentially depending on all model parameters (including σ, M and R).

B.6 Operators and norms

We introduce all operators in more detail.

We let $\mathbf{Z} = \mathcal{X} \times \mathbb{R}$ denote the sample space, where the input space $\mathcal{X}$ is a standard Borel space endowed with a fixed unknown probability measure ν. The kernel space $\mathcal{H}$ is assumed to be separable, equipped with a measurable positive semi-definite kernel K, bounded by κ, implying continuity of the inclusion map $I_\nu : \mathcal{H} \longrightarrow L^2(\nu)$. Moreover, we consider the covariance operator $T_\nu = I_\nu^* I_\nu = \mathbb{E}[K_X \otimes K_X^*]$, which can be shown to be positive self-adjoint trace class (and hence is compact). Given a sample $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, we define the sampling operator $S_\mathbf{x} : \mathcal{H} \longrightarrow \mathbb{R}^n$ by $(S_\mathbf{x} f)_i = \langle f, K_{x_i} \rangle_\mathcal{H}$. The empirical covariance operator is given by $T_\mathbf{x} = S_\mathbf{x}^* S_\mathbf{x} : \mathcal{H} \longrightarrow \mathcal{H}$. We furthermore introduce $\bar{T} = \kappa^{-2} T$ and similarly $\bar{T}_\mathbf{x} = \kappa^{-2} T_\mathbf{x}$, $\bar{S}_\mathbf{x} = \kappa^{-1} S_\mathbf{x}$.

For a partition $\{\mathcal{X}_1, ..., \mathcal{X}_m\}$ of $\mathcal{X}$, we denote by $\hat{\mathcal{H}}_j$ the local RKHS with (extended) kernel $\hat{K}_j$, supported on $\mathcal{X}_j$, with associated covariance operator $\bar{T}_j = \kappa_j^{-2} T_j = \kappa_j^{-2} \mathbb{E}_{\nu_j}[\hat{K}_j(X, \cdot) \otimes \hat{K}_j(X, \cdot)^*]$. Given a sample $\mathbf{x}_j = (x_{j,1}, \ldots, x_{j,n_j}) \in \mathcal{X}_j^{n_j}$, we define the sampling operator $S_{\mathbf{x}_j} : \mathcal{H}_j \longrightarrow \mathbb{R}^{n_j}$ similarly by $(S_{\mathbf{x}_j} f)_i = \langle f, \hat{K}_j(x_i, \cdot) \rangle_{\mathcal{H}_j}$.

Lemma B.6.1. *Given $j \in [m]$ let $p_j = \nu(\mathcal{X}_j)$ and $\nu_j(A) = \nu(A|\mathcal{X}_j)$, for a measurable $A \subset \mathcal{X}$. One has*

$$L^2(\mathcal{X}, \nu) = \bigoplus_{j=1}^{m} p_j L^2(\mathcal{X}_j, \nu_j)$$

with

$$||f||^2_{L^2(\nu)} = \sum_{j=1}^m p_j ||f_j||^2_{L^2(\nu_j)} \ ,$$

where $f = \oplus_{j=1}^m f_j$.

The next Lemma states that the global effective dimension can be expressed as the sum of the local ones.

Lemma B.6.2 (Effective Dimension)**.** *For any* $\lambda \in [0,1]$

$$\sum_{j=1}^m \mathcal{N}_{\nu_j}(\bar{T}_j, \lambda) = \mathcal{N}_\nu(\bar{T}, \lambda) \ .$$

B.7 Proofs of Section B.2

We let $f^* \in \mathcal{H}$, i.e. $f^* = \oplus_{j=1}^m \hat{f}_j^*$, with $\hat{f}_j^* \in \hat{\mathcal{H}}_j$. Note that $\hat{f}_j$ is defined on all of $\mathcal{X}$. We shall use the following error decomposition:

$$f^* - \hat{f}^\lambda_{\mathcal{D}} = \sum_{j=1}^m \hat{f}_j^* - \hat{f}^\lambda_{\mathcal{D}_j} = \sum_{j=1}^m r_\lambda(\bar{T}_{\mathbf{x}_j}) \hat{f}_j^* + \sum_{j=1}^m g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j} \hat{f}_j^* - \bar{S}^*_{\mathbf{x}_j} \mathbf{y}_j) \ ,$$

where $\hat{f}^\lambda_{\mathcal{D}}$ is given in (B.2.1), with $r_\lambda(t) = 1 - g_\lambda(t)t$ and with $g_\lambda(t) = (t+\lambda)^{-1}$.

Proposition B.7.1 (Approximation Error $L^2(\nu)-$ norm)**.** *Let* ρ *be a source distribution belonging to* $\mathcal{M}$*, defined in Assumption B.2.2. For any* $\lambda \in (0,1]$*, one has*

$$\mathbb{E}_{\rho^{\otimes n}} \Big[\| \sum_{j=1}^m r_\lambda(\bar{T}_{\mathbf{x}_j}) \hat{f}_j^* \|^2_{L^2(\nu)} \Big] \leq C R^2 \sum_{j=1}^m p_j \mathcal{B}^2_{n_j}(\bar{T}_j, \lambda) \lambda^{2(r+\frac{1}{2})} \ ,$$

where $\mathcal{B}^2_{n_j}(\bar{T}_j, \lambda)$ *is defined in Proposition B.10.3 .*

Proof of Proposition B.7.1. According to Lemma B.6.1, by assumption $\mathbf{SC}(r, R)$ we have

$$\begin{aligned}
\mathbb{E}_{\rho^{\otimes n}} \Big[\| \sum_{j=1}^m r_\lambda(\bar{T}_{\mathbf{x}_j}) \hat{f}_j^* \|^2_{L^2(\nu)} \Big] &= \sum_{j=1}^m p_j \mathbb{E}_{\rho^{\otimes n}} \Big[\| r_\lambda(\bar{T}_{\mathbf{x}_j}) \hat{f}_j^* \|^2_{L^2(\nu_j)} \Big] \\
&= \sum_{j=1}^m p_j \mathbb{E}_{\rho^{\otimes n}} \Big[\| \sqrt{\bar{T}_j} r_\lambda(\bar{T}_{\mathbf{x}_j}) \hat{f}_j^* \|^2_{\hat{\mathcal{H}}_j} \Big] \\
&\leq C R^2 \sum_{j=1}^m p_j \mathbb{E}_{\rho^{\otimes n}} \Big[\| \sqrt{\bar{T}_j} r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}_j^r \|^2 \Big] \ . \qquad \text{(B.7.1)}
\end{aligned}$$

We bound for any $j \in [m]$ the expectation by first deriving a probabilistic estimate. For any $\eta \in (0,1]$, with probability at least $1 - \eta$

$$\begin{aligned}
||\sqrt{\bar{T}_j} r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}_j^r|| &\leq C \log^2(2\eta^{-1}) \mathcal{B}_{n_j}(\bar{T}_j, \lambda) \ ||\bar{T}_j^{\frac{1}{2}} (\bar{T}_j + \lambda)^{\frac{1}{2}}|| \ ||(\bar{T}_{\mathbf{x}_j} + \lambda)^{\frac{1}{2}} r_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j} + \lambda)^r|| \ ||(\bar{T}_j + \lambda)^r \bar{T}_j^r|| \\
&\leq C \log^2(2\eta^{-1}) \mathcal{B}_{n_j}(\bar{T}_j, \lambda) \lambda^{r+\frac{1}{2}} \ .
\end{aligned}$$

Here we have used that

$$||(\bar{T}_{\mathbf{x}_j}+\lambda)^{\frac{1}{2}} r_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}+\lambda)^r|| \leq C\lambda^{r+\frac{1}{2}}$$

and that for $s \in [0, \frac{1}{2}]$

$$||(\bar{T}_j+\lambda)^s \bar{T}_j^s|| \leq ||(\bar{T}_j+\lambda)\bar{T}_j||^s \leq 1$$

by Proposition B.11.1 and the spectral theorem. Also, from Proposition B.11.1 and Proposition B.10.3

$$||(\bar{T}_{\mathbf{x}_j}+\lambda)^{-\frac{1}{2}}(\bar{T}_j+\lambda)^{\frac{1}{2}}|| \leq ||(\bar{T}_{\mathbf{x}_j}+\lambda)^{-1}(\bar{T}_j+\lambda)||^{\frac{1}{2}} \leq \sqrt{8}\log(2\eta^{-1})\mathcal{B}_{n_j}^{\frac{1}{2}}(\bar{T}_j,\lambda)\ .$$

From Lemma B.11.2, by integration

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\sqrt{\bar{T}_j}\, r_\lambda(\bar{T}_{\mathbf{x}_j})\bar{T}_j^r\right\|^2\right] \leq C\mathcal{B}_{n_j}^2(\bar{T}_j,\lambda)\lambda^{2(r+\frac{1}{2})}\ .$$

Combining this with (B.7.1) gives

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\sum_{j=1}^m r_\lambda(\bar{T}_{\mathbf{x}_j})\hat{f}_j^*\right\|_{L^2(\nu)}^2\right] \leq CR^2\sum_{j=1}^m p_j\mathcal{B}_{n_j}^2(\bar{T}_j,\lambda)\lambda^{2(r+\frac{1}{2})}\ .$$

□

Proposition B.7.2 (Sample Error $L^2(\nu)-$ norm)**.** *Let ρ be a source distribution belonging to $\mathcal{M}$, defined in Assumption B.2.2. For any $\lambda \in (0,1]$, one has*

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\sum_{j=1}^m g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\right\|_{L^2(\nu)}^2\right] \leq C\sum_{j=1}^m p_j\ \mathcal{B}_{n_j}^2(\bar{T}_j,\lambda)\lambda\ \left(\frac{M}{n_j\lambda}+\sigma\sqrt{\frac{\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda)}{n_j\lambda}}\right)^2\ ,$$

where C does not depend on $(\sigma, M, R) \in \mathbb{R}_+^3$.

Proof of Proposition B.7.2. Recall that $||\sqrt{\bar{T}_j}\hat{f}||_{\hat{\mathcal{H}}_j} = ||\hat{f}||_{L^2(\nu_j)}$. According to Lemma B.6.1 we have

$$\begin{aligned}\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\sum_{j=1}^m g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\right\|_{L^2(\nu)}^2\right] &= \sum_{j=1}^m p_j\mathbb{E}_{\rho^{\otimes n}}\left[\left\|g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\right\|_{L^2(\nu_j)}^2\right] \\ &= \sum_{j=1}^m p_j\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\sqrt{\bar{T}_j}g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\right\|_{\hat{\mathcal{H}}_j}^2\right]. \quad \text{(B.7.2)}\end{aligned}$$

We bound the expectation for each separate subsample of size n_j by first deriving a probabilistic estimate and then by integration. For this reason, we use (B.10.2) and Proposition B.11.1 and write for any $\hat{f}_j \in \hat{\mathcal{H}}_j$, $j \in [m]$

$$\begin{aligned}||\sqrt{\bar{T}_j}\hat{f}_j||_{\hat{\mathcal{H}}_j} &\leq ||\sqrt{\bar{T}_j}(T_j+\lambda)^{-1/2}||\ ||(T_j+\lambda)^{1/2}(T_{\mathbf{x}_j}+\lambda)^{-1/2}||\ ||(\bar{T}_{\mathbf{x}_j}+\lambda)^{1/2}\hat{f}_j||_{\hat{\mathcal{H}}_j} \\ &\leq ||\bar{T}_j(T_j+\lambda)^{-1}||^{1/2}\ ||(T_j+\lambda)(T_{\mathbf{x}_j}+\lambda)^{-1}||^{1/2}\ ||(\bar{T}_{\mathbf{x}_j}+\lambda)^{1/2}\hat{f}_j||_{\hat{\mathcal{H}}_j} \\ &\leq C\log(4\eta^{-1})\mathcal{B}_{n_j}^{1/2}(\bar{T}_j,\lambda)\ ||(\bar{T}_{\mathbf{x}_j}+\lambda)^{1/2}\hat{f}_j||_{\hat{\mathcal{H}}_j}\ , \quad \text{(B.7.3)}\end{aligned}$$

holding with probability at least $1-\frac{\eta}{2}$. We proceed by splitting

$$(\bar{T}_{\mathbf{x}_j}+\lambda)^s g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}f_\rho - \bar{S}_{\mathbf{x}_j}^\star\mathbf{y}_j) = H_{\mathbf{x}_j}^{(1)}\cdot H_{\mathbf{x}_j}^{(2)}\cdot h_{\mathbf{z}_j}^\lambda\ , \quad \text{(B.7.4)}$$

with

$$\begin{aligned} H^{(1)}_{\mathbf{x}_j} &:= (\bar{T}_{\mathbf{x}_j}+\lambda)^{\frac{1}{2}} g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}+\lambda)^{\frac{1}{2}}, \\ H^{(2)}_{\mathbf{x}_j} &:= (\bar{T}_{\mathbf{x}_j}+\lambda)^{-\frac{1}{2}} (\bar{T}+\lambda)^{\frac{1}{2}}, \\ h^{\lambda}_{\mathbf{z}_j} &:= (\bar{T}+\lambda)^{-\frac{1}{2}} (\bar{T}_{\mathbf{x}_j} f_\rho - \bar{S}^{\star}_{\mathbf{x}_j}\mathbf{y}_j) \, . \end{aligned}$$

The first term is clearly bounded. The second term is now estimated using (B.10.2) once more. One has with probability at least $1-\frac{\eta}{4}$

$$H^{(2)}_{\mathbf{x}_j} \leq \sqrt{8}\log(8\eta^{-1})\mathcal{B}_{\frac{n}{m}}(\bar{T}_j,\lambda)^{\frac{1}{2}} \, .$$

Finally, $h^{\lambda}_{\mathbf{z}_j}$ is estimated using Proposition B.10.2:

$$h^{\lambda}_{\mathbf{z}_j} \leq 2\log(8\eta^{-1})\left(\frac{M}{n_j\sqrt{\lambda}} + \sigma\sqrt{\frac{\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda)}{n_j}}\right) \, ,$$

holding with probability at least $1-\frac{\eta}{4}$. Thus, combining the estimates following (B.7.4) with (B.7.3) gives for any $j=1,...,m$

$$||\sqrt{\bar{T}_j}g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}f_\rho - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j)||_{\hat{\mathcal{H}}_j} \leq C_{\blacktriangle}\log^3(8\eta^{-1})\mathcal{B}_{n_j}(\bar{T}_j,\lambda)\sqrt{\lambda}\,\left(\frac{M}{n_j\lambda}+\sigma\sqrt{\frac{\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda)}{n_j\lambda}}\right) \, ,$$

with probability at least $1-\eta$. By integration using Lemma B.11.2 one obtains

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\sqrt{\bar{T}_j}g_\lambda(\bar{B}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}f_\rho - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j)\right\|^2_{\hat{\mathcal{H}}_j}\right]^{\frac{1}{2}} \leq C\mathcal{B}_{n_j}(\bar{T}_j,\lambda)\sqrt{\lambda}\,\left(\frac{M}{n_j\lambda}+\sigma\sqrt{\frac{\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda)}{n_j\lambda}}\right) \, .$$

Combining this with (B.7.2) implies

$$\mathbb{E}_{\rho^{\otimes n}}\left[\|\sum_{j=1}^m g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}^*_j - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j)\|^2_{L^2(\nu)}\right] \leq C\sum_{j=1}^m p_j\,\mathcal{B}^2_{n_j}(\bar{T}_j,\lambda)\lambda\,\left(\frac{M}{n_j\lambda}+\sigma\sqrt{\frac{\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda)}{n_j\lambda}}\right)^2 \, ,$$

where C does not depend on $(\sigma,M,R)\in\mathbb{R}^3_+$. □

We are now ready to prove Theorem B.2.3.

Proof of Theorem B.2.3. Let the regularization parameter λ_n be chosen as

$$\lambda_n = R\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{b}{2br+b+1}} \tag{B.7.5}$$

and assume that $n_j = \lfloor \frac{n}{m_n} \rfloor$. Lemma B.10.5 yields $\mathcal{B}_{\frac{n}{m_n}}(\bar{T}_j,\lambda_n)\leq 2$ provided n is sufficiently large and

$$m_n \leq n^\alpha \, , \qquad \alpha < \frac{2br}{2br+b+1} \, .$$

From Proposition B.7.1 and Proposition B.7.2 we immediately obtain (recalling the definition of a_n in

B.2.3)

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|f^* - \hat{f}_{\mathcal{D}}^{\lambda_n}\right\|_{L^2(\nu)}^2\right] \leq C\sum_{j=1}^{m_n} p_j a_n^2 + p_j\lambda_n \left(\frac{Mm_n}{n\lambda_n} + \sigma\sqrt{\frac{m_n \mathcal{N}_{\nu_j}(\bar{T}_j,\lambda_n)}{n\lambda_n}}\right)^2$$
$$\leq C\left(a_n^2 + \lambda_n\left(\frac{Mm_n}{n\lambda_n}\right)^2 + \sigma^2\frac{m_n}{n}\sum_{j=1}^{m_n} p_j\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda_n)\right) .$$

Under Assumption B.2.2 we obtain

$$\sigma^2\frac{m_n}{n}\sum_{j=1}^{m_n} p_j\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda_n) \leq C\frac{\sigma^2}{n}\mathcal{N}_\nu(\bar{T},\lambda_n) \leq C\frac{\sigma^2}{n}\lambda_n^{-\frac{1}{b}} = CR^2\lambda_n^{2(r+\frac{1}{2})} = Ca_n^2 .$$

Moreover,

$$\lambda_n\left(\frac{Mm_n}{n\lambda_n}\right)^2 \leq \lambda_n^{2r+1} ,$$

provided that

$$m_n \leq n^\alpha , \qquad \alpha < \frac{b(r+1)}{2br+b+1} .$$

As a result,

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|f^* - \hat{f}_{\mathcal{D}}^{\lambda_n}\right\|_{L^2(\nu)}^2\right] \leq Ca_n^2 ,$$

where C does not depend on the model parameter $(\sigma, M, R) \in \mathbb{R}_+$. □

Proof of Theorem B.2.5. Assume that $n_j = \lfloor\frac{n}{m_n}\rfloor$. Let the regularization parameter λ_n be given by

$$\lambda_n = R_h\left(\frac{\sigma^2}{R_h^2 n}\right)^{\frac{b}{2br_h+b+1}} \tag{B.7.6}$$

and let

$$m_n = \left(\frac{R_h^2 n}{\sigma^2}\right)^\alpha , \quad \alpha < \frac{2br_h}{2br_h+b+1} . \tag{B.7.7}$$

As above, Lemma B.10.5 yields $\mathcal{B}_{\frac{n}{m_n}}(\bar{T}_j,\lambda_n) \leq 2$ provided n is sufficiently large.

From Proposition B.7.1 we immediately obtain for the approximation error, recalling the definition of a_n in B.2.6

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\sum_{j=1}^{m_n} r_{\lambda_n}(\bar{T}_{\mathbf{x}_j})\hat{f}_j^*\right\|_{L^2(\nu)}^2\right] \leq C\left(R_l^2\left(\sum_{j\in E_{m_n}} p_j\right)\lambda_n^{2(r_l+\frac{1}{2})} + R_h^2\left(\sum_{j\in E_{m_n}^c} p_j\right)\lambda_n^{2(r_h+\frac{1}{2})}\right)$$
$$\leq C\left(R_l^2\frac{R_h^2}{R_l^2}\left(\frac{1}{m_n}\right)^{1-\frac{r_l}{r_h}}\lambda_n^{2(r_l+\frac{1}{2})} + R_h^2\lambda_n^{2(r_h+\frac{1}{2})}\right)$$
$$\leq CR_h^2\left(\left(\frac{1}{m_n}\right)^{1-\frac{r_l}{r_h}}\lambda_n^{2(r_l+\frac{1}{2})} + \lambda_n^{2(r_h+\frac{1}{2})}\right) .$$

Here we have used that by Assumption B.2.4

$$\left(\sum_{j\in E_{m_n}} p_j\right) \leq \frac{R_h^2}{R_l^2}\left(\frac{1}{m_n}\right)^{1-\frac{r_l}{r_h}} \quad \text{and} \quad \left(\sum_{j\in E_{m_n}^c} p_j\right) \leq 1 .$$

Finally, the choice (B.7.7) ensures that

$$\left(\frac{1}{m_n}\right)^{1-\frac{r_l}{r_h}} \lambda_n^{2(r_l+\frac{1}{2})} = \left(\frac{\sigma^2}{R_h^2 n}\right)^{\alpha(1-\frac{r_l}{r_h})} \lambda_n^{2(r_l+\frac{1}{2})} \leq \left(\frac{\sigma^2}{R_h^2 n}\right)^{\frac{2br_h}{2br_h+b+1}(1-\frac{r_l}{r_h})} \lambda_n^{2(r_l+\frac{1}{2})} = \lambda_n^{2(r_h+\frac{1}{2})} .$$

As a result, the approximation error satisfies

$$\mathbb{E}_{\rho^{\otimes n}}\left[\Big\| \sum_{j=1}^{m_n} r_{\lambda_n}(\bar{T}_{\mathbf{x}_j}) \hat{f}_j^* \Big\|_{L^2(\nu)}^2\right] \leq C R_h^2 \lambda_n^{2(r_h+\frac{1}{2})} = C a_n^2 .$$

The bound for the sample error follows exactly as in the proof of Theorem B.2.3. □

B.8 Proofs of Section B.3

For proving Theorem B.3.1 we use the non-asymptotic error decomposition given in Theorem 2 of [77], somewhat reformulated and streamlined using our estimate B.10.2. We adopt the notation and idea of [77] and write $\hat{f}_{n,l}^{\lambda} = g_{\lambda,l}(\bar{T}_{\mathbf{x}}) S_{\mathbf{x}}^* \mathbf{y}$, with $g_{\lambda,l}(\bar{T}_{\mathbf{x}}) = V(V^* \bar{T}_{\mathbf{x}} V + \lambda)^{-1} V^*$ and $VV^* = P_l$, the projection operator onto $\mathcal{H}_l$, $l \leq n$. Consider

$$||\sqrt{\bar{T}}(\hat{f}_{n,l}^{\lambda} - f_\rho)||_{\mathcal{H}} \leq T_1 + T_2$$

with

$$T_1 = ||g_{\lambda,l}(T_{\mathbf{x}})(S_{\mathbf{x}}^* \mathbf{y} - T_{\mathbf{x}} f_\rho)||_{L^2(\nu)} = ||\sqrt{\bar{T}} g_{\lambda,l}(T_{\mathbf{x}})(S_{\mathbf{x}}^* \mathbf{y} - T_{\mathbf{x}} f_\rho)||_{\mathcal{H}}$$

and

$$T_2 = ||\sqrt{\bar{T}} g_{\lambda,l}(T_{\mathbf{x}})(T_{\mathbf{x}} f_\rho - f_\rho)||_{\mathcal{H}} .$$

Proposition B.8.1 (Expectation Sample Error KRR-Nyström)**.**

$$\mathbb{E}_{\rho^{\otimes n}}\left[\big\| g_{\lambda,l}(T_{\mathbf{x}})(S_{\mathbf{x}}^* \mathbf{y} - T_{\mathbf{x}} f_\rho)\big\|_{L^2(\nu)}^2\right]^{\frac{1}{2}} \leq C\, \sqrt{\lambda} \mathcal{B}_n(\bar{T}, \lambda) \left(\frac{M}{n\lambda} + \sigma\sqrt{\frac{\mathcal{N}_\nu(\bar{T}, \lambda)}{n\lambda}}\right)$$

where C does not depend on $(\sigma, M, R) \in \mathbb{R}_+^3$.

Proof of Proposition B.8.1. For estimating T_1 we use Proposition B.10.3 and obtain for any $\lambda \in (0, 1]$ with probability at least $1 - \eta$

$$\begin{aligned} T_1 &\leq C \log(2\eta^{-1}) \mathcal{B}_n(\bar{T}, \lambda)\, ||(\bar{T}_{\mathbf{x}} + \lambda)^{1/2} g_{\lambda,l}(T_{\mathbf{x}})(S_{\mathbf{x}}^* \mathbf{y} - T_{\mathbf{x}} f_\rho)||_{\mathcal{H}} \\ &\leq C \log^2(4\eta^{-1}) \mathcal{B}_n^2(\bar{T}, \lambda)\, ||(\bar{T}_{\mathbf{x}} + \lambda)^{1/2} g_{\lambda,l}(T_{\mathbf{x}})(\bar{T}_{\mathbf{x}} + \lambda)^{1/2}|| \\ &\qquad ||(\bar{T} + \lambda)^{-1/2}(S_{\mathbf{x}}^* \mathbf{y} - T_{\mathbf{x}} f_\rho)||_{\mathcal{H}} . \end{aligned}$$

From Proposition 6 in [77] and from the spectral Theorem we obtain

$$||(\bar{T}_{\mathbf{x}} + \lambda)^{1/2} g_{\lambda,l}(T_{\mathbf{x}})(\bar{T}_{\mathbf{x}} + \lambda)^{1/2}|| \leq 1 .$$

Thus, applying Proposition $B.10.1$ one has with probability at least $1-\eta$

$$T_1 \leq C\log^3(8\eta^{-1})\,\sqrt{\lambda}\mathcal{B}_n^2(\bar{T},\lambda)\,\left(\frac{M}{n\lambda}+\sigma\sqrt{\frac{\mathcal{N}_\nu(\bar{T},\lambda)}{n\lambda}}\right)\,,$$

where C does not depend on $(\sigma,M,R)\in\mathbb{R}^3_+$. Integration using Lemma B.11.2 gives the result. □

Before we proceed we introduce the *computational error*: For $u\in[0,\frac{1}{2}]$, $\lambda\in(0,1]$ define

$$\mathcal{C}_u(l,\lambda) := ||(Id-VV^*)(\bar{T}+\lambda)^u||\,.$$

The proof of the following Lemma can be found in [77], proof of Theorem 2.

Lemma B.8.2. *For any* $u\in[0,\frac{1}{2}]$

$$\mathcal{C}_u(l,\lambda)\leq\mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2u}\,.$$

Lemma B.8.3. *If* λ_n *is defined by* (B.7.5) *and if*

$$l_n\geq n^\beta\qquad\beta>\frac{b+1}{2br+b+1}$$

one has with probability at least $1-\eta$

$$\mathcal{C}_{\frac{1}{2}}(l_n,\lambda_n)\leq C\log(2\eta^{-1})\sqrt{\lambda_n}\,,$$

provided n *is sufficiently large.*

Proof of Lemma B.8.3. Using Proposition 3 in [77] one has with probability at least $1-\eta$

$$\begin{aligned}\mathcal{C}_{\frac{1}{2}}(l,\lambda_n)&\leq\sqrt{\lambda_n}\,||(T_{\mathbf{x}_l}+\lambda_n)^{-1}(T+\lambda_n)||^{\frac{1}{2}}\\&\leq C\log(2\eta^{-1})\sqrt{\lambda_n}\mathcal{B}_l^{\frac{1}{2}}(\bar{T},\lambda_n)\,.\end{aligned}$$

Recall that $\mathcal{N}_\nu(\bar{T},\lambda)\leq C_b\lambda^{-\frac{1}{b}}$, implying

$$\mathcal{B}_l(\bar{T},\lambda_n)\leq C\left(1+\left(\frac{2}{l\lambda_n}+\sqrt{\frac{\lambda_n^{-\frac{1}{b}}}{l\lambda_n}}\right)^2\right)\,.$$

Straightforward calculation shows that

$$\frac{2}{l_n\lambda_n}=o(1)\,,\quad\text{if }\;l_n\geq n^\beta\,,\;\beta>\frac{b}{2br+b+1}$$

and

$$\sqrt{\frac{\lambda_n^{-\frac{1}{b}}}{l_n\lambda_n}}=o(1)\,,\quad\text{if }\;l_n\geq n^\beta\,,\;\beta>\frac{b+1}{2br+b+1}\,.$$

Thus, $\mathcal{C}_{\frac{1}{2}}(l_n,\lambda_n)\leq C\log(2\eta^{-1})\sqrt{\lambda_n}$, with probability at least $1-\eta$. □

Proposition B.8.4 (Expectation Approximation- and Computational Error KRR-Nyström)**.** *Assume that*

$$l_n \geq n^{\beta} \ , \qquad \beta > \frac{b+1}{2br+b+1}$$

and $(\lambda_n)_n$ *is chosen according to* (B.7.5)*. If* n *is sufficiently large*

$$\mathbb{E}_{\rho^{\otimes n}}\left[\left\|\sqrt{\bar{T}}g_{\lambda_n,l_n}(T_{\mathbf{x}})(T_{\mathbf{x}}f_\rho - f_\rho)\right\|^2_{L^2(\nu)}\right]^{\frac{1}{2}} \leq C\, a_n \ ,$$

where C *does not depend on* $(\sigma, M, R) \in \mathbb{R}^3_+$*.*

Proof of Proposition B.8.4. Using that $f_\rho \in \Omega_\nu(r,R)$ one has for any $\lambda \in (0,1]$

$$T_2 \leq CR\ (\ (a) + (b) + (c)\ )\ , \tag{B.8.1}$$

with

$$(a) = ||\sqrt{\bar{T}}(Id - VV^*)\bar{T}^r|| \ , \qquad (b) = \lambda||\sqrt{\bar{T}}g_{\lambda,l}(T_{\mathbf{x}})\bar{T}^r||$$

and

$$(c) = ||\sqrt{\bar{T}}g_{\lambda,l}(T_{\mathbf{x}})(\bar{T}_{\mathbf{x}} + \lambda)(Id - VV^*)\bar{T}^r|| \ .$$

Since $(Id - VV^*)^2 = (Id - VV^*)$ we obtain by Lemma B.8.2

$$(a) \leq \mathcal{C}_{\frac{1}{2}}(l,\lambda)\ \mathcal{C}_r(l,\lambda) \leq \mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2r+1} \ .$$

Furthermore, using (B.10.2) , with probability at least $1 - \frac{\eta}{2}$

$$\begin{aligned}(b) &\leq C\log^2(8\eta^{-1})\lambda\mathcal{B}_n^{\frac{1}{2}+r}(\bar{T},\lambda)\ ||(\bar{T}_{\mathbf{x}} + \lambda)^{1/2}g_{\lambda,l}(T_{\mathbf{x}})(\bar{T}_{\mathbf{x}} + \lambda)^r|| \\ &\leq C\log^2(8\eta^{-1})\lambda^{\frac{1}{2}+r}\mathcal{B}_n^{\frac{1}{2}+r}(\bar{T},\lambda) \ ,\end{aligned}$$

by again using Proposition 6 in [77].

The last term gives with probability at least $1 - \frac{\eta}{2}$

$$\begin{aligned}(c) &\leq C\log(8\eta^{-1})||(\bar{T}_{\mathbf{x}} + \lambda)^{1/2}g_{\lambda,l}(T_{\mathbf{x}})(\bar{T}_{\mathbf{x}} + \lambda)||\ \mathcal{C}_r(l,\lambda) \\ &\leq C\log(8\eta^{-1})\sqrt{\lambda}\ \mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2r} \ .\end{aligned}$$

Combining the estimates for (a), (b) and (c) gives

$$T_2 \leq CR\log^2(8\eta^{-1})\left(\mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2r+1} + \lambda^{\frac{1}{2}+r}\mathcal{B}_n^{\frac{1}{2}+r}(\bar{T},\lambda) + \sqrt{\lambda}\ \mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2r}\right) \ .$$

We now choose λ_n according to (B.7.5) . Notice that by Lemma B.10.4 one has $\mathcal{B}_n(\bar{T},\lambda_n) \leq C$ for any n sufficiently large. Applying Lemma B.8.3 we obtain, with probability at least $1 - \eta$

$$T_2 \leq C\log^2(8\eta^{-1})R\lambda_n^{r+\frac{1}{2}} \ ,$$

provided n is sufficiently large and

$$l_n \geq n^{\beta} \ , \qquad \beta > \frac{b+1}{2br+b+1} \ .$$

The result follows from integration by applying Lemma A.3.3 and recalling that $a_n = R\lambda_n^{r+\frac{1}{2}}$. □

Proof of Theorem B.3.1. The proof easily follows by combining Proposition B.8.1 and Proposition B.8.4 . In particular, the estimate for the sample error by choosing $\lambda = \lambda_n$ follows by recalling that $\mathcal{N}_\nu(\bar{T}, \lambda_n) \leq C_b \lambda_n^{-\frac{1}{b}}$, by definition of $(a_n)_n$ in (B.2.3) by Lemma B.10.4 and by

$$\frac{M}{n\lambda_n} = o\left(\sigma\sqrt{\frac{\lambda_n^{-\frac{1}{b}}}{n\lambda_n}}\right) .$$

□

B.9 Proofs of Section B.4

In this section we give a sketch of proof of the main result.

Proposition B.9.1 (Sample Error KRR Localnysed)**.** *Let $m_n \leq n^\alpha$ with $\alpha < \frac{2br}{2br+b+1}$ and let λ_n be defined as in* (B.7.5)*. If n is sufficiently large, one has*

$$\mathbb{E}_{\rho^{\otimes n}}\left[\Big\| \sum_{j=1}^{m} g_{\lambda_n,l}(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\Big\|_{L^2(\nu)}^2\right]^{\frac{1}{2}} \leq Ca_n ,$$

where C does not depend on the model parameter σ, M, R.

Proof of Proposition B.9.1. Applying Proposition B.8.1 we obtain

$$\begin{aligned}\mathbb{E}_{\rho^{\otimes n}}\left[\Big\| \sum_{j=1}^{m} g_{\lambda,l}(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\Big\|_{L^2(\nu)}^2\right] &= \sum_{j=1}^{m} p_j \mathbb{E}_{\rho^{\otimes n}}\left[\big\| g_{\lambda,l}(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\big\|_{L^2(\nu_j)}^2\right] \\ &\leq C\sum_{j=1}^{m} p_j \ \mathcal{B}_{\frac{n}{m}}^2(\bar{T}_j, \lambda)\lambda \ \left(\frac{Mm}{n\lambda} + \sigma\sqrt{\frac{m\mathcal{N}_{\nu_j}(\bar{T}_j, \lambda)}{n\lambda}}\right)^2 .\end{aligned}$$

Arguing as in the proof of Theorem B.2.3, using Lemma B.10.5, implies the result. □

Proposition B.9.2 (Approximation and Computational Error KRR Localnysed)**.** *Let λ_n be defined by* (B.7.5)*. Assume $l_n \geq n^\beta$, $m_n \leq n^\alpha$ with*

$$\alpha < \frac{2br}{2br+b+1} , \qquad \beta > \frac{b+1}{2br+b+1} .$$

Then, if n is sufficiently large

$$\mathbb{E}_{\rho^{\otimes n}}\left[\Big\| \sum_{j=1}^{m_n} g_{\lambda,l_n}(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\Big\|_{L^2(\nu)}^2\right]^{\frac{1}{2}} \leq Ca_n ,$$

where C does not depend on the model parameter σ, M, R.

Proof of Proposition B.9.2. Let $\lambda \in (0, 1]$. For proving this Proposition we combine techniques from both the partitioning and subsampling approach. More precisely:

$$\mathbb{E}_{\rho^{\otimes n}}\Big[\|\sum_{j=1}^{m} g_{\lambda,l}(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\|_{L^2(\nu)}^2\Big] = \sum_{j=1}^{m} p_j \mathbb{E}_{\rho^{\otimes n}}\Big[\|g_{\lambda,l}(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\|_{L^2(\nu_j)}^2\Big]$$
$$= \sum_{j=1}^{m} p_j \mathbb{E}_{\rho^{\otimes n}}\Big[\|\sqrt{\bar{T}_j} g_{\lambda,l}(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)\|_{\hat{\mathcal{H}}_j}^2\Big] .$$

We shall decompose as in B.8.1, with $\bar{T}$ replaced by $\bar{T}_j$ and $\bar{T}_{\mathbf{x}}$ replaced by $\bar{T}_{\mathbf{x}_j}$,

$$||\sqrt{\bar{T}_j} g_{\lambda,l}(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}_{\mathbf{x}_j}^*\mathbf{y}_j)||_{\hat{\mathcal{H}}_j} \leq CR\,(\,(a)+(b)+(c)\,) = (*)\,.$$

Following the lines of the proof of Proposition B.8.4 leads to an upper bound (with probability at least $1-\eta$) for the rhs of the last inequality, which is

$$(*) \leq CR\log^2(8\eta^{-1})\left(\mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2r+1} + \lambda^{\frac{1}{2}+r}\mathcal{B}_n^{\frac{1}{2}+r}(\bar{T},\lambda) + \sqrt{\lambda}\,\mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2r}\right)$$
$$\leq CR\log^2(8\eta^{-1})\lambda^{r+\frac{1}{2}}\left(\mathcal{B}_l^{2r+1}(\bar{T}_j,\lambda) + \mathcal{B}_{\frac{n}{m}}^{r+\frac{1}{2}}(\bar{T}_j,\lambda) + \mathcal{B}_l^{2r}(\bar{T}_j,\lambda)\right)\,.$$

We proceed by estimating each term separately by choosing $\lambda = \lambda_n$, $l = l_n$ and $m = m_n$. By Lemma B.6.2, each local effective dimension is bounded by the global one. Thus (recall the Definition of $\mathcal{B}_l(\bar{T}_j,\lambda)$ in Proposition B.10.3)

$$\mathcal{B}_{l_n}(\bar{T}_j,\lambda_n) \leq 1 + \left(\frac{2}{l_n\lambda_n} + \sigma\sqrt{\frac{\lambda_n^{-\frac{1}{b}}}{l_n\lambda_n}}\right)^2\,.$$

Straightforward calculation shows that

$$\frac{2}{l_n\lambda_n} = o(1)\,, \quad \text{if } l_n \geq n^{\beta'}\,,\ \beta' > \frac{b}{2br+b+1} \tag{B.9.1}$$

and

$$\sqrt{\frac{\lambda_n^{-\frac{1}{b}}}{l_n\lambda_n}} = o(1)\,, \quad \text{if } l_n \geq n^{\beta'}\,,\ \beta' > \frac{b+1}{2br+b+1}\,. \tag{B.9.2}$$

Furthermore

$$\mathcal{B}_{\frac{n}{m_n}}^{r+\frac{1}{2}}(\bar{T}_j,\lambda_n) = \left[1 + \left(\frac{2m_n}{n\lambda_n} + \sqrt{\frac{m_n\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda_n)}{n\lambda}}\right)^2\right]^{r+\frac{1}{2}}$$
$$\leq \left[1 + \left(\frac{2m_n}{n\lambda_n} + \sqrt{\frac{m_n\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda_n)}{n\lambda}}\right)^2\right]$$
$$\leq 2\left[1 + \left(\frac{2m_n}{n\lambda_n}\right)^2 + \left(\frac{m_n\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda_n)}{n\lambda}\right)\right]\,.$$

since $r+\frac{1}{2} \leq 1$ and $(1+A)^{r+\frac{1}{2}} \leq 1+A$ for any $A \geq 0$. Combining the last steps results in (by integration

using Lemma B.11.2) and Assumption B.2.1, 3.

$$\begin{aligned}\mathbb{E}_{\rho^{\otimes n}}\left[\|\sum_{j=1}^{m_n} g_{\lambda_n,l_n}(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j}\hat{f}_j^* - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j)\|^2_{L^2(\nu)}\right] &\leq CR^2\lambda_n^{2(r+\frac{1}{2})}\sum_{j=1}^{m_n} p_j\left(1+\left(\frac{2m_n}{n\lambda_n}\right)^2+\left(\frac{m_n\mathcal{N}_{\nu_j}(\bar{T}_j,\lambda_n)}{n\lambda}\right)\right)\\ &\leq CR^2\lambda_n^{2(r+\frac{1}{2})}\sum_{j=1}^{m_n} p_j\left(1+\left(\frac{2m_n}{n\lambda_n}\right)^2+\left(\frac{m_n\mathcal{N}_{\nu}(\bar{T},\lambda_n)}{n\lambda}\right)\right)\\ &\leq CR^2\lambda_n^{2(r+\frac{1}{2})}\sum_{j=1}^{m_n} p_j\left(1+\left(\frac{2m_n}{n\lambda_n}\right)^2+\left(\frac{m_n\lambda_n^{-\frac{1}{b}}}{n\lambda}\right)\right)\\ &\leq CR^2\lambda_n^{2(r+\frac{1}{2})}\sum_{j=1}^{m_n} p_j\left(1+\left(\frac{2m_n}{n\lambda_n}\right)^2+m_n\lambda_n^{2r}\right)\,.\end{aligned}$$

Furthermore,

$$\frac{2m_n}{n\lambda_n} = o\left(\sqrt{m}_n\lambda_n^r\right)\,,$$

provided

$$m_n \leq n^\alpha\,, \qquad \alpha < \frac{2(br+1)}{2br+b+1}\,.$$

Finally, $\sqrt{m}_n\lambda_n^r = o(1)$ if

$$m_n \leq n^\alpha\,, \qquad \alpha < \frac{2br}{2br+b+1}\,.$$

Finally, conditions (B.9.1) and (B.9.2) are satisfied by choosing $m_n \leq n^\alpha$ and $l_n \geq \frac{n^\beta}{m_n}$ with $\beta' = \beta+\alpha$. □

B.10 Probabilistic Inequalities

Proposition B.10.1 ([17]). *For $n \in \mathbb{N}$, $\lambda \in (0,1]$ and $\eta \in (0,1]$, one has with probability at least $1-\eta$:*

$$\left\|(\bar{T}+\lambda)^{-\frac{1}{2}}\ \left(\bar{T}_{\mathbf{x}}f_\rho - \bar{S}^\star_{\mathbf{x}}\mathbf{y}\right)\right\|_{\mathcal{H}_1} \ \leq\ 2\log(2\eta^{-1})\left(\frac{M}{n\sqrt{\lambda}}+\sqrt{\frac{\sigma^2\mathcal{N}_\nu(\bar{T},\lambda)}{n}}\right)\,.$$

Proposition B.10.2 ([17], Proposition 5.3). *Let $x_1,...,x_n$ be an iid sample, drawn according to ν on $\mathcal{X}$. For any $\lambda \in (0,1]$ and $\eta \in (0,1)$ one has with probability at least $1-\eta$:*

$$\left\|(\bar{T}+\lambda)^{-1}(\bar{T}-\bar{T}_{\mathbf{x}})\right\|_{\mathrm{HS}} \ \leq 2\log(2\eta^{-1})\left(\frac{2}{n\lambda}+\sqrt{\frac{\mathcal{N}_\nu(\bar{T},\lambda)}{n\lambda}}\right)\,.$$

Proposition B.10.3 ([62]). *Let $x_1,...,x_n$ be an iid sample, drawn according to ν on $\mathcal{X}$. Define*

$$\mathcal{B}_n(\bar{T},\lambda) := \left[1+\left(\frac{2}{n\lambda}+\sqrt{\frac{\mathcal{N}_\nu(\bar{T},\lambda)}{n\lambda}}\right)^2\right] \tag{B.10.1}$$

For any $\lambda > 0$, $\eta \in (0,1]$, with probability at least $1-\eta$ one has

$$\left\|(\bar{T}_{\mathbf{x}}+\lambda)^{-1}(\bar{T}+\lambda)\right\| \leq 8\log^2(2\eta^{-1})\mathcal{B}_n(\bar{T},\lambda)\,. \tag{B.10.2}$$

Lemma B.10.4. *If λ_n is defined by* (B.7.5)

$$\mathcal{B}_n(\bar{T}, \lambda_n) \leq 2 \,,$$

provided n is sufficiently large.

Proof of Lemma B.10.4. The proof is a straightforward calculation using Definition (B.7.5) and recalling that $\mathcal{N}_\nu(\bar{T}, \lambda) \leq C_b \lambda^{-\frac{1}{b}}$. □

Lemma B.10.5. *If λ_n is defined by* (B.7.5) *and if*

$$m_n \leq n^\alpha \,, \qquad \alpha < \frac{2br}{2br + b + 1} \,,$$

one has for any $j \in [m]$

$$\mathcal{B}_{\frac{n}{m}}(\bar{T}_j, \lambda_n) \leq \mathcal{B}_{\frac{n}{m}}(\bar{T}, \lambda_n) \leq 2 \,,$$

provided n is sufficiently large.

Proof of Lemma B.10.5. The first inequality follows from Lemma B.6.2 , since $\mathcal{N}_{\nu_j}(\bar{T}_j, \lambda) \leq \mathcal{N}_\nu(\bar{T}, \lambda)$, $j \in [m]$. For proving the second inequality, recall that $\mathcal{N}_\nu(\bar{T}, \lambda_n) \leq C_b \lambda_n^{-\frac{1}{b}}$ and $\sigma\sqrt{\frac{\lambda_n^{-\frac{1}{b}}}{n\lambda_n}} = R\lambda_n^r$. Using the definition of λ_n in (B.7.5) yields

$$\frac{2m}{n\lambda_n} = o\left(\sqrt{m}\lambda_n^r\right) \,,$$

provided

$$m \leq n^\alpha \,, \qquad \alpha < \frac{2(br+1)}{2br + b + 1} \,.$$

Finally, $\sqrt{m}\lambda_n^r = o(1)$ if

$$m \leq n^\alpha \,, \qquad \alpha < \frac{2br}{2br + b + 1} \,.$$

□

B.11 Miscellanea

Proposition B.11.1 (Cordes Inequality,[5], Theorem IX.2.1-2)**.** *Let A, B be to self-adjoint, positive operators on a Hilbert space. Then for any $s \in [0, 1]$:*

$$\|A^s B^s\| \leq \|AB\|^s \,. \tag{B.11.1}$$

Lemma B.11.2. *Let X be a non-negative random variable with $\mathbb{P}[X > C \log^u(k\eta^{-1})] < \eta$ for any $\eta \in (0, 1]$. Then $\mathbb{E}[X] \leq \frac{C}{k} u\Gamma(u)$.*

Proof. Apply $\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > t]dt$. □

Bibliography

[1] H. Avron, K. L. Clarkson, and D. P. Woodruff. Faster kernel ridge regression using sketching and preconditioning. arxiv preprint (1611.03220), 2017.

[2] F. Bach. Sharp analysis of low-rank kernel matrix approximations. *JMLR Workshop and Conference Proceedings*, 30, 2013.

[3] F. R. Bach and M. I. Jordan. Predictive low-rank decomposition for kernel methods. *ICML '05 Proceedings of the 22nd international conference on Machine learning*, 2005.

[4] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23(1):52–72, 2007.

[5] R. Bhatia. *Matrix Analysis*. Springer, 1997.

[6] R. Bhatia and J. Holbrook. Fréchet derivatives of the power function. *Indiana University Mathematics Journal*, 49 (3):1155–1173, 2000.

[7] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1987.

[8] L. Birgé. An alternative point of view on Lepski's method. *Institute of Mathematical Statistics, Beachwood*, 36:113–133, 2001. http://projecteuclid.org/euclid.lnms/1215090065.

[9] N. Bissantz, T. Hohage, and A. Munk. Consistency and rates of convergence of nonlinear tikhonov regularization with random noise. *Inverse Problems*, 20:1773, 2004. 6.

[10] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Analysis*, 45(6):2610–2636, 2007.

[11] G. Blanchard, M. Hoffmann, and M. Reiß. Optimal adaptation for early stopping in statistical inverse problems. arxiv preprint (1606.07702), 2016.

[12] G. Blanchard and N. Krämer. Optimal learning rates for kernel conjugate gradient regression. *Advances in Neural Information Processing Systems 23*, 2010.

[13] G. Blanchard and N. Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14(6):763–794, 2016.

[14] G. Blanchard and P. Massart. Discussion of "2004 IMS medallion lecture: Local Rademacher complexities and oracle inequalities in risk minimization", by V. Koltchinskii. *Annals of Statistics*, 34(6):2664–2671, 2006.

[15] G. Blanchard, P. Mathé, and N. Mücke. Lepski principle in supervised learning. work in preparation, 2017.

[16] G. Blanchard and N. Mücke. Parallelizing spectral algorithms for kernel learning. arXiv Preprint (1610.07487), 2016.

[17] G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 2017. doi:10.1007/s10208-017-9359-7.

[18] P. Bühlmann and B. Yu. Boosting with the l_2-loss: Regression and classification. *Journal of American Statistical Association*, 98(462):324–339, 2003.

[19] R. Camoriano, T. Angles, A. Rudi, and L. Rosasco. When subsampling meets early stopping. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[20] A. Caponnetto. Optimal rates for regularization operators in learning theory. Technical report, MIT, 2006.

[21] A. Caponnetto and Y. Yao. Cross-validion based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(2):161–183, 2010.

[22] N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[23] F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428, 2002.

[24] E. De Vito and A. Caponnetto. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006.

[25] E. De Vito, S. Pereverzyev, and L. Rosasco. Adaptive kernel methods using the balancing principle. *Foundations of Computational Mathematics*, 10 (4):455–479, 2010.

[26] E. De Vito, L. Rosasco, and A. Caponnetto. Discretization error analysis for Tikhonov regularization. *Analysis and Applications*, 4(1):81–99, 2006.

[27] E. De Vito, L. Rosasco, A. Caponnetto, and U. De Giovannini. Learning from examples as an inverse problem. *J. of Machine Learning Research*, 6:883–904, 2005.

[28] R. DeVore, G. Kerkyacharian, D. Picard, and V.Temlyakov. Mathematical methods for supervised learning. *Foundations of Computational Mathematics*, 6(1):3–58, 2006.

[29] L. Dicker, D. Foster, and D. Hsu. Kernel methods and regularization techniques for nonparametric regression: Minimax optimality and adaptation. Technical report, Rutgers University, 2015.

[30] M. Dimassi and J. Sjöstrand. Trace asymptotics via almost analytic extensions. *Partial Differential equations and Mathematical physics.Prog. Nonl. Diff. Equ. Appl.*, 21:126 –142, 1996.

[31] M. Dimassi and J. Sjöstrand. *Spectral Asymptotics in the Semi-Classical Limit*. Cambridge University Press, 1999.

[32] S. Dirksen. *Noncommutative and vector-valued Rosenthal inequalities*. PhD thesis, Delft Univ. Technology, 2011.

[33] M. Eberts. *Adaptive rates for Support Vector Machines*. PhD thesis, Universität Stuttgart, 2014.

[34] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 2000.

[35] H. W. Engl, K. Kunisch, and A. Neubauer. Convergence rates for Tikhonov regularisation of nonlinear ill-posed problems. *Inverse Problems*, 5(4):523, 1989.

[36] J. C. Ferreira and V. A. Menegatto. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral equations and Operator Theory*, 64, 2009.

[37] K. Fukumizu, F. R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

[38] L. L. Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

[39] M. Gevrey. Sur la nature analytique des solutions des equations aux dérivées partielles (premier mémoire). *Annales scientifique de l' É.N.S.*, 35:129 – 190, 1918.

[40] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural network architectures. *Neural Computation*, 7(2):219–269, 1993.

[41] A. Goldenshluger and S. Pereverzev. On adaptive inverse estimation of linear functionals in hilbert scales. *Bernoulli*, 9(5):783–807, 2003.

[42] Z. Guo, S. Lin, and D. Zhou. Learning theory of distributed spectral algorithms. *Preprint*, 2016.

[43] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer, 2002.

[44] P. Halmos and V. Sunder. *Bounded Integral Operators on L^2-Spaces*. Springer, 1978.

[45] L. Hörmander. The spectral function of an elliptic operator. *Acta Math.*, 124:173–218, 1968.

[46] L. Hörmander. On the asymptotic distribution of the eigenvalues of pseudodifferential operators in R^n. *Ark. Mat.*, 17(2):297–313, 1979.

[47] L. Hörmander. *The analysis of linear partial differential operators 1*. Springer, 1983.

[48] L. Hörmander. *The analysis of linear partial differential operators 2*. Springer, 1983.

[49] L. Hörmander. *The analysis of linear partial differential operators 3*. Springer, 1985.

[50] V. Ivrii. *Microlocal analysis and precise spectral asymptotics*. Springer, 1998.

[51] Q. Jin and U. Tautenhahn. On the discrepancy principle for some newton type methods for solving nonlinear inverse problems. *Numer. Math.*, 111(4):509–558, 2009.

[52] M. Klein and J. Rama. Almost exponential decay of quantum resonance states and paley-wiener type estimates in gevrey spaces. *Ann. Henri Poinc.*, 11:499 – 537, 2010.

[53] H. Komatsu. Ultradistributions. I. structure theorems and a characterization. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.*, 20:25–105, 1973.

[54] O. Lepski. Some new ideas in nonparametric estimation. arxiv preprint (1603.03934), 2016.

[55] O. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory Probab. Appl.*, 35(3):454–466, 1990.

[56] O. Lepskii. Asymptotically minimax adaptive estimation I: Upper bounds. optimally adaptive estimates. *Theory Probab. Appl.*, 36(4), 1992.

[57] O. Lepskii. Asymptotically minimax adaptive estimation II: Schemes without optimal adaptation: Adaptive estimators. *Theory Probab. Appl.*, 37(3), 1993.

[58] S. Lin, X. Guo, and D.-X. Zhou. Distributed learning with regularized least squares. arXiv Preprint (1608.03339), 2016.

[59] J. Loubes and C. Ludena. Penalized estimators for non linear inverse problems. *ESAIM: PS*, 14:173–191, 2010.

[60] S. Loustau. Inverse statistical learning. *Electron. J. Statist.*, 7:2065–2097, 2013.

[61] S. Loustau and C. Marteau. Minimax fast rates for discriminant analysis with errors in variables. *Bernoulli*, 21(1):176–208, 2015.

[62] S. Lu, P. Mathé, and S. V. Pereverzev. Balancing principle in supervised learning for a general regularization scheme. Technical report, RICAM, 2016.

[63] P. Mathé. The Lepskii principle revisited. *Inverse Problems*, 22(3):L11–L15, 2006.

[64] P. Mathé and S. Pereverzev. Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(3):789, 2003.

[65] M. Meister and I. Steinwart. Optimal learning rates for localized svms. *Journal of Machine Learning Research*, 17(194):1–44, 2016.

[66] S. Mendelson and J. Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

[67] N. Mücke. Functional calculus for elliptic selfadjoint semiclassical pseudodifferential operators. *Diploma Thesis, Universität Potsdam*, 2010.

[68] A. Neubauer. Tikhonov regularisation for non-linear ill-posed problems: optimal convergence rates and finite-dimensional approximation. *Inverse Problems*, 5(4):541, 1989.

[69] F. O'Sullivan. Convergence characteristics of methods of regularization estimators for nonlinear operator equations. *SIAM J. Numer. Anal.*, 27(6):1635–1649, 1990.

[70] I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

[71] I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 30(1):143–148, 1985.

[72] G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and non-parametric regression. *JMLR*, 15:335–366, 2014.

[73] A. Rastogi and S. Sampath. Optimal rates for the regularized learning algorithms under general source condition. *Frontiers in Applied Mathematics and Statistics*, 3:3, 2017.

[74] M. Reed and B. Simon. *Functional Analysis I.* Academic Press, 1980.

[75] L. Rodino. *Linear Partial Differential Operators in Gevrey Spaces.* World Scientific, 1993.

[76] H. P. Rosenthal. On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970.

[77] A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems 28*, 2015.

[78] S. Smale and D. Zhou. Shannon sampling II: Connections to learning theory. *Appl. Comput. Harmon. Analysis*, 19(3):285–302, 2005.

[79] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximation. *Constructive Approximation*, 26(2):153–172, 2007.

[80] A. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. *In 17th International Conference on Machine Learning, Stanford*, pages 911–918, 2000.

[81] I. Steinwart and A. Christman. *Support Vector Machines.* Springer, 2008.

[82] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.

[83] C. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

[84] R. Tandon, S. Si, and P. Ravikumar. Kernel ridge regression via partitioning. arXiv Preprint (1608.01976), 2016.

[85] U. Tautenhahn. Error estimates for regularized solutions of non-linear ill-posed problems. *Inverse Problems*, 10(2):485, 1994.

[86] V. Temlyakov. Approximation in learning theory. *Constructive Approximation*, 27(1):33–74, 2008.

[87] A. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2008.

[88] G. Wahba. *Spline Models for Observational Data*, volume 59. SIAM CBMS-NSF Series in Applied Mathematics, 1990.

[89] C. Wang and D.-X. Zhou. Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, 27(1):55–67, 2011.

[90] C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems 13*, 2000.

[91] C. Xu, Y. Zhang, and R. Li. On the feasibility of distributed kernel regression for big data. arXiv Preprint (1505.00869), 2015.

[92] Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. arxiv preprint (1501.06195), 2017.

[93] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[94] T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.*, 17(9):2077–2098, 2005.

[95] Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression. *JMLR: Workshop and Conference Proceedings*, 30, 2013.

[96] D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18 (3):739–767, 2002.

[97] D.-X. Zhou. Distributed learning algorithms. Technical report, Mathematisches Forschungsinstitut Oberwolfach Report No. 33, 2016.

Eigenständigkeitserklärung:

Ich versichere, die Arbeit selbstständig verfasst zu haben sowie keine anderen Quellen und Hilfsmittel als die angegebenen verwendet zu haben.

Diese Arbeit ist bisher an keiner anderen Hochschule eingereicht worden.

Nicole Mücke