Global and Local Kernel Methods for Dataset Shift, Scalable Inference and Optimization

by

Davit Gogolashvili

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy in the Doctoral School N.130: Computer Science, Telecommunications and Electronics of Paris of the Sorbonne University

Examining Committee:

Evgeny Burnaev	*Skolkovo Institute of Science and Technology*	Reviewer
Lorenzo Rosasco	*University of Genoa*	Reviewer
Maurizio Filippone	*EURECOM*	Advisor
Maria Zuluaga	*EURECOM*	Examiner
Marco Lorenzi	*INRIA*	Examiner

Presented on the 16th December 2022

“There are three kinds of lies: lies, damned lies, and statistics.”

The origin of this phrase is unclear, but Mark Twain attributed it to Benjamin Disraeli

SORBONNE UNIVERSITÉ

Abstract

Doctor of Philosophy

by **Davit Gogolashvili**

In many real world problems, the training data and test data have different distributions. The most common settings for dataset shift often considered in the literature are covariate shift and target shift. In this thesis, we investigate nonparametric models applied to the dataset shift scenario.

We develop a novel framework to accelerate Gaussian process regression (GPR). In particular, we consider localization kernels at each data point to down-weigh the contributions from other data points that are far away, and we derive the GPR model stemming from the application of such localization operation. Through a set of experiments, we demonstrate the competitive performance of the proposed approach compared to full GPR, other localized models, and deep Gaussian processes. Crucially, these performances are obtained with considerable speedups compared to standard global GPR due to the sparsification effect of the Gram matrix induced by the localization operation.

We propose a new method for estimating the minimizer x^* and the minimum value f^* of a smooth and strongly convex regression function f from the observations contaminated by random noise.

Acknowledgements

First, I would like to thank my supervisor Maurizio Filippone, for his guidance through the last three years.

I want to thank all of my colleagues, in particular those who I worked most closely with:

Motonobu Kanagawa, to whom I am grateful for the project that forms the basis of this thesis.

Matteo Zecchin, for the important conversations and extremely productive working relationship.

Bogdan Kozyrskiy, for making the chapter 4 of the thesis possible.

Arya Akhavan, for the endless and inspiring discussions and an amazing time in Paris and Genova.

Alexandre Tsybakov, without whom the chapter 5 of this thesis would not exist. I feel fortunate to have worked with Sasha.

It is important that I acknowledge CREST in Paris and MaLGa in Genova for hosting me and allowing me to be a part of their fantastic academic environment. I would like to highlight important conversations with Lorenzo Rosasco, Nicolas Schreuder and Vassilis Apidopoulos.

I would like to thank my fellow PhD students and ESRs from the Windmill project for the fantastic time spent together.

On a personal note, I would like to thank my family, particularly my parents who gave me every opportunity.

Finally, I would like to thank my amazing wife Tina and my daughter Mariam. This thesis exists only because of them.

Contents

List of Figures

List of Tables

Abbreviations

GP	**G**aussian **P**rocess
GPs	**G**aussian **P**rocesse**s**
IW	**I**mportance **W**eighting
KDE	**K**ernel **D**ensity **E**stimator
KNN	**K**-**N**earest **N**eighbors
KRR	**K**ernel **R**idge **R**egression
LSGPR	**L**ocally **S**moothed **G**aussian **P**rocess **R**egression
MLL	**M**arginal **L**og-**L**ikelihood
MSE	**M**ean-**S**quared **E**rror
ONB	**O**rtho**N**ormal **B**asis
RBF	**R**adial-**B**asis **F**unction
RKHS	**R**eproducing **K**ernel **H**ilbert **S**pace

Notation

(X, Y)	input-output pair
ρ^{te}	Testing measure
ρ^{tr}	Training measure
f_ν	The regression function of ν
K	Mercer kernel
$K_{\mathbf{xx}}$	The Gramm matrix, whose i, jth element is $K(x_i, x_j)$.
k	Smoothing kernel
T_ν	The covariance operator
L_ν	The integral operator
$\|\cdot\|$	The Euclidean norm
$\|\cdot\|_{\mathrm{op}}$	Operator norm from $\mathcal{H}$ to $\mathcal{H}$, i.e. for $U : \mathcal{H} \to \mathcal{H}$, $\|U\|_{\mathrm{op}} = \sup_{\|u\|_{\mathcal{H}} \leq 1} \|Uu\|_{\mathcal{H}}$.
$\mu_{\min}(U)$	The smallest eigenvalue of a linear operator U.
ℓ^2	The space consisting of all sequences $x = (x_n)_{n \in \mathbb{N}}$ satisfying $\sum_n \|x_n\|^2 < \infty$.
$L^2(X, \nu)$	Class of square integrable with respect to the measure ν functions on X.
$a \vee b$	$\max(a, b)$
$a \wedge b$	$\min(a, b)$
$\mathbb{I}(A)$	Indicator function of a set A
$\boldsymbol{m}$	A d-dimensional multi index $\boldsymbol{m} = (m_1, \ldots, m_d)$, where $m_j \geq 0$ are integers.
$\|\boldsymbol{m}\|$	The absolute value of the multi index defined as $\|\boldsymbol{m}\| = m_1 + \ldots + m_d$
$\boldsymbol{m}!$	The factorial of the multi index defined as $\boldsymbol{m}! = m_1! \ldots m_d!$
$u^{\boldsymbol{m}}$	The power of the multi index defined as $u^{\boldsymbol{m}} = u_1^{m_1} \ldots u_d^{m_d}$, for $u \in \mathbb{R}^d$.
$D^{\boldsymbol{m}}$	The differentiation operator $D^{\boldsymbol{m}} = \frac{\partial^{\|\boldsymbol{m}\|}}{\partial u_1^{m_1} \ldots \partial u_d^{m_d}}$.
$[n]$	The set that contains all positive integers j, such that $1 \leq j \leq n$.
$\lfloor \beta \rfloor$	Biggest integer smaller than or equal to β.

To my family:
Tina and Mariam

Chapter 1

Learning Theory

1.1 The Classical Learning Problem

In the classical learning problems one considers the random vector (x, y) sampled from the (unknown) Borel probability measure (distribution) ρ defined on $Z = X \times Y$. The ultimate goal is to understand how the response variable $y \in Y$ depends on the value of the observation vector $x \in X$. Obviously, the response y for every fixed x is generated according to the conditional measure $\rho(y|x)$, related to the initial measure ρ trough the equality

$$\rho(x, y) = \rho(y|x)\rho_X(x)$$

where ρ_X is a marginal probability measure on X, defined by $\rho_X(S) = \rho\left(\pi^{-1}(S)\right)$ where $\pi : X \times Y \to X$ is the projection. Asking for the estimation of the response distribution for a given input point x is too much. In many learning situations understanding the expected value of the conditional output is sufficient. To this order we define the function $f_\rho : X \to Y$ by

$$f_\rho(x) = \int_Y y d\rho(y|x).$$

The function f_ρ is called the *regression function* of ρ. For each $x \in X$, $f_\rho(x)$ is the average of the y coordinate of $\{x\} \times Y$. Figure 1.1 visualizes the regression function together with the densities associated with the conditional and marginal measures. The regression function can be viewed as a minimizer of

$$\mathcal{E}(f) := \mathcal{E}_\rho(f) := \int_Z (f(x) - y)^2 d\rho \tag{1.1}$$

among all square integrable (w.r.t. ρ) functions $f : X \to Y$.

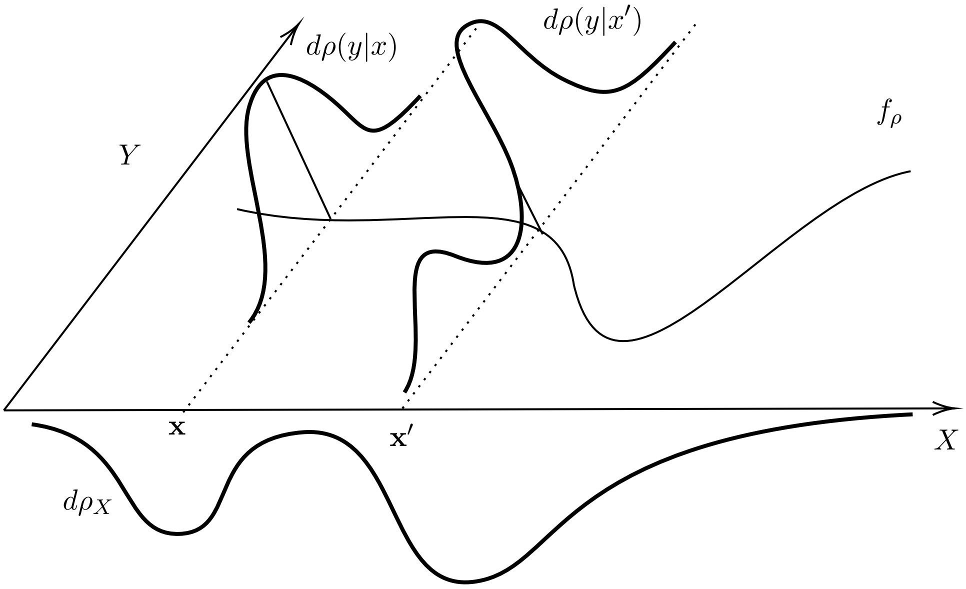


FIGURE 1.1: The regression function f_ρ is the average of the y coordinate of $\{x\} \times Y$.

In the applications, the measure ρ (and hence also the regression function) is usually unknown. Therefore, it is impossible to understand how the conditional expectation behaves. But it is often possible to observe data according to the distribution ρ and to estimate the regression function from these data.

To be more precise, denote by

$$\mathbf{z} \in Z^n, \quad \mathbf{z} = \left(\left(x_1, y_1\right), \ldots, \left(x_n, y_n\right)\right)$$

collection of independent and identically distributed (i.i.d.) random variables from ρ, called the *training sample* in Z^n. Our problem then is given the data $\mathbf{z}$, how to find a good approximation $f_{\mathbf{z}}(x)$ to $f_\rho(x)$ at the new point x, coming from the *same* distribution as the training data. The following property of the regression function

$$\mathcal{E}\left(f_\rho\right) = \inf_{f \in L^2(X, \rho_X)} \mathcal{E}(f) \tag{1.2}$$

suggests to consider the minimization of so called *empirical risk functional*

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{n} \sum_{i=1}^{n} \left(f\left(x_i\right) - y_i\right)^2 .$$

The choice of the function class where the minimization of the empirical risk is performed is of great importance. This class of functions $\mathcal{H}$ is called the *hypothesis space* in learning theory. A typical choice for $\mathcal{H}$ is the space of polynomials, or more generally the spaces produced by other (then monomials) basis functions like *wavelets*. The class of piecewise constant functions is another popular space. In this work, we focus primarily on reproducing kernel Hilbert spaces (RKHS) that correspond to positive definite functions.

1.2 Reproducing Kernel Hilbert Spaces

Definition 1.1. We say that the continuous function $K : X \times X \to \mathbb{R}$ is a *Mercer kernel* if it symmetric and for any $n \in \mathbb{N}, (c_1, \ldots, c_n) \subset \mathbb{R}$ and $(x_1, \ldots, x_n) \subset X$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K\left(x_i, x_j\right) \geq 0.$$

Remark 1.2. Definition 2.1 can be equivalently stated thus: A symmetric function K is positive definite if the matrix $K_{\mathbf{xx}} \in \mathbb{R}^{n \times n}$ with elements $\left[K_{\mathbf{xx}}\right]_{ij} = K\left(x_i, x_j\right)$ is positive semidefinite for any finite set $\mathbf{x} = (x_1, \ldots, x_n) \in X^n$ of any size $n \in \mathbb{N}$.

In the remainder, for simplicity, kernel always means positive definite kernel. For $\mathbf{x} = \{x_1, \ldots, x\} \in X^n$, the matrix $K_{\mathbf{xx}}$ is the kernel matrix or *Gram matrix*.

Example 1.1 (Polynomial kernels). *Let $X \subset \mathbb{R}^d$. For $m \in \mathbb{N}$, the polynomial kernel $K_m : X \times X \to \mathbb{R}$ is defined by*

$$K_m\left(x, x'\right) = \left(x^{\top} x' + c\right)^m, \quad x, x' \in X.$$

Example 1.2 (Gaussian RBF Kernels). *Let $X \subset \mathbb{R}^d$. For $\ell > 0$, a Gaussian RBF kernel $K_\ell : X \times X \to \mathbb{R}$ is defined by*

$$K_\ell\left(x, x'\right) = \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right), \quad x, x' \in X.$$

Example 1.3 (Matérn kernels). *Let $X \subset \mathbb{R}^d$. For positive constants α and ℓ, the Matérn kernel $K_{\alpha,\ell} : X \times X \to \mathbb{R}$ is defined by*

$$K_{\alpha,\ell}\left(x, x'\right) = \frac{2^{1-\alpha}}{\Gamma(\alpha)} \left(\frac{\sqrt{2\alpha}\,\|x - x'\|}{\ell}\right)^{\alpha} B_\alpha\left(\frac{\sqrt{2\alpha}\,\|x - x'\|}{\ell}\right), \quad x, x' \in X,$$

where Γ is the gamma function, and B_α is the modified Bessel function of the second kind of order α.

Remark 1.3. The scaling is chosen so that for $\alpha \to \infty$ we obtain the Gaussian RBF kernels in Example 1.2. That is, for a Matérn kernel $K_{\alpha,\ell}$ with $\ell > 0$ being fixed, we have

$$\lim_{\alpha\to\infty} K_{\alpha,\ell}\left(x,x'\right) = \exp\left(-\frac{\|x-x'\|^2}{2\ell^2}\right), \quad x,x' \in X.$$

Example 1.4 (**Wiener kernel**)**.** *Let $X \subset \mathbb{R}^d$. The Wiener kernel $K : X \times X \to \mathbb{R}$ is defined by*

$$K(x,x') = \prod_{i=1}^{d} x_i \wedge x_i', \quad x,x' \in X,$$

where x_i and x_i' are the ith coordinate of the vectors x and x'.

Given a kernel K we are going to associate with it a *reproducing kernel Hilbert space.* A RKHS is a Hilbert space of real-valued functions on X with the property that, for each $x \in X$, the evaluation functional Ev_x, which associates f with $f(x)$, $\mathrm{Ev}_x\, f \to f(x)$, is a bounded, linear functional. The boundedness means that there exists nonegative constant C_x such that

$$|\mathrm{Ev}_x\, f| = |f(x)| \leqslant C_x\|f\|, \quad \text{for all } f \text{ in the RKHS.}$$

If $\mathcal{H}$ is a RKHS, then by the Ritz representation theorem, for every evaluation functional Ev_x there exists a unique element K_x in $\mathcal{H}$ with the *reproducing property*

$$\mathrm{Ev}_x\, f = \langle K_x, f\rangle_{\mathcal{H}} = f(x), \quad \forall f \in \mathcal{H}. \tag{1.3}$$

It can be proved (see for example Cucker and Zhou (2007), Theorem 2.9) that to every RKHS $\mathcal{H}$ there corresponds a unique Mercer kernel $K(x,x')$ of two variables in X, called the *reproducing kernel* of $\mathcal{H}$ (hence the terminology RKHS), that has the reproducing property (1.3). Converse is also true: given a kernel K on $X \times X$ we can construct a unique RKHS of real-valued functions on X with K as its reproducing kernel.

RKHS and the Eigenvalues of the Integral Operator.A RKHS can be defined in terms of the eigenvalues and eigenfunctions of the *integral operator*. Let X be a compact space equipped with a strictly positive finite Borel measure ν and let $K(x,x) < \infty$. Define the integral operator $L_\nu : L^2(X,\nu) \to L^2(X,\nu)$ as

$$(L_\nu f)(x) = \int K(x',x)f(x')d\nu(x'), \tag{1.4}$$

where $L^2(X,\nu)$ is the space of square integrable functions with respect to ν.

Then there exists an orthonormal sequence of continuous eigenfunctions, $\phi_1, \phi_2, \ldots$ in $L^2(X,\nu)$ and eigenvalues $\mu_1 \geq \mu_2 \geq \ldots \geq 0$, with $L_\nu\phi_i = \mu_i\phi_i, \forall i \in \mathbb{N}$ and

$$K(x,x') = \sum_{i=1}^{\infty} \mu_i\phi_i(x)\phi_i(x'), \tag{1.5}$$

where the convergence is absolute and uniform. Moreover the sum $\sum_{i=1}^{\infty}\mu_i$ is convergent and

$$\sum_{i=1}^{\infty}\mu_i = \int_X K(x,x)d\nu.$$

From the expansion (1.5) it follows that $K(x,x')$ corresponds to a dot product in $\ell^2 :=$, since $K(x,x') = \langle \Phi(x), \Phi(x')\rangle_{\ell^2}$ with

$$\begin{aligned} \Phi : X &\to \ell^2 \\ x &\mapsto \left(\sqrt{\mu_i}\phi_i(x)\right)_{k\in\mathbb{N}}. \end{aligned}$$

The map Φ is well-defined and continuous. The space ℓ^2 is called the *feature space* and the function Φ *feature map.*

How let us describe the RKHS in terms of eigenfunctions and eigenvalues of L_ν. It can be shown (ref) that the functions $\{\sqrt{\mu_i}\phi_i\}$ forms an orthonormal basis (ONB) in $\mathcal{H}$. If the RKHS is infinite dimensional, L_ν has infinitely many positive eigenvalues μ_i, $i \geq 1$ and

$$\mathcal{H} = \left\{ f = \sum_{i=1}^{\infty} a_i\phi_i : \left\{\frac{a_i}{\sqrt{\mu_i}}\right\}_{i=1}^{\infty} \in \ell^2 \right\}, \tag{1.6}$$

where the inner product between $f = \sum a_i\phi_i$ and $g = \sum b_i\phi_i$ is given by

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty}\frac{a_ib_i}{\mu_i}.$$

Remark 1.4. When $\dim\mathcal{H} = N < \infty$, integral operator L_ν has finite number of positive eigenvalues and ℓ^2 in (1.6) is replaced by $\mathbb{R}^N$.

RKHS as a Linear Combination of the Kernels. Another way to construct a RKHS, closely related to the construction (1.6), is by taking a completion of the linear space of all functions

$$x \mapsto \sum_{i=1}^{n} a_iK\left(x_i, x\right), \quad a_1,\ldots,a_n \in \mathbb{R},\ x_1,\ldots,x_n \in X,\ n\in\mathbb{N}$$

relative to the norm induced by the inner product

$$\left\langle \sum_{i=1}^{n} a_iK\left(x_i,\cdot\right), \sum_{j=1}^{m} b_jK\left(x_j',\cdot\right)\right\rangle_{\mathcal{H}} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_ib_jK\left(x_i, x_j'\right).$$

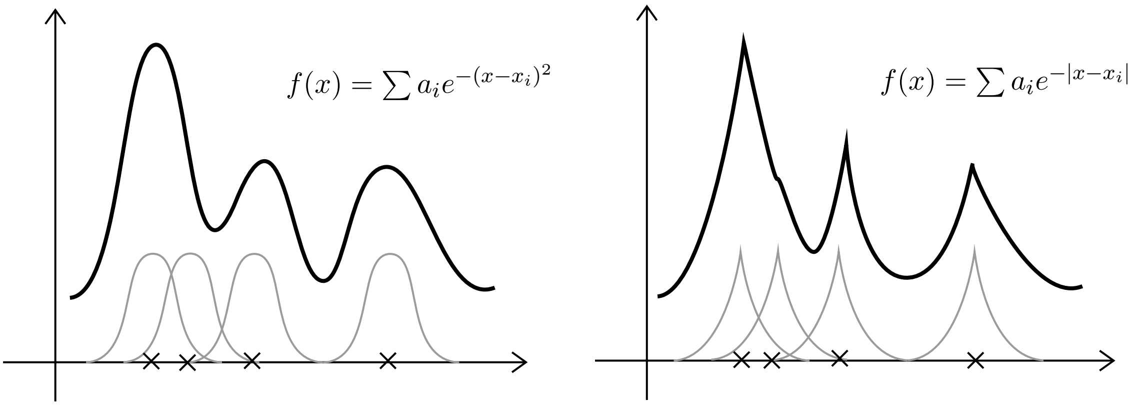


FIGURE 1.2: Smoothness of the function in RKHS induced by the smoothness of the reproducing kernel. The left panel shows the function from the RKHS with Gaussian RBF kernel. All members of this RKHS are infinitely differentiable. On the right panel we plot one member from RKHS reproduced by the Laplace kernel.

This leads us to the following representation by the feature maps $\{K(x_i, \cdot)\}$

$$\mathcal{H} = \left\{ f = \sum_{i=1}^{\infty} a_i K(x_i, \cdot) : \{a_i\}_{i=1}^{\infty} \subset \mathbb{R}, \{x_i\}_{i=1}^{\infty} \subset X, \|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{\infty} a_i a_j K(x_i, x_j) < \infty \right\}.$$

Form the construction above it is apparent that the function f in the RKHS inherit the smoothness properties of the underling kernel K. More precisely, if the kernel K is m-times continuously differentiable, then so are the functions in $\mathcal{H}$. The smoothness property of the RKHS norm can be seen in the following example on the RKHS of a Matern kernel.

Example 1.5 (Matérn kernel RKHS). *Let $K_{\alpha,\ell}$ be the Matérn kernel on $X \subset \mathbb{R}^d$ with sufficiently smooth boundary and let $s := \alpha + d/2$ be an integer. Then the RKHS of $K_{\alpha,\ell}$ is norm-equivalent to the Sobolev space $W_2^s(X)$ of order s defined by*

$$W_2^s(X) := \left\{ f \in L^2(X, \upsilon) : \|f\|_{W_2^s(X)}^2 := \sum_{|\boldsymbol{m}| \leq s} \|D^{\boldsymbol{m}} f\|_{\upsilon}^2 < \infty \right\}$$

where υ is a Lebesgue measure on $\mathbb{R}^d$.

1.3 Kernel Least Squares and Regularization

After we have decided on a set $\mathcal{H}$ which we shall use in approximating f_ρ, we can define the *empirical risk minimizer* by

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(f). \tag{1.7}$$

Clearly, the empirical risk minimizer can not perform better then the minimizer of the true risk $\mathcal{E}(f)$ over a set $\mathcal{H}$

$$f_{\mathcal{H}} := \arg\min_{f\in\mathcal{H}} \mathcal{E}(f). \tag{1.8}$$

The function $f_{\mathcal{H}}$ is called the *target function* and is known to be the best approximator of f_ρ from $\mathcal{H}$:

$$\|f_\rho - f_{\mathcal{H}}\|_{\rho_X} = \mathrm{dist}_{\rho_X}(f_\rho, \mathcal{H}) := \inf_{f\in\mathcal{H}} \|f_\rho - f\|_{\rho_X}. \tag{1.9}$$

The quantity above depends on the choice of $\mathcal{H}$ but is independent of sample $\mathbf{z}$. Let us note that if $f_\rho \in \mathcal{H}$, then $f_{\mathcal{H}} = f_\rho$. Understanding how well $\mathcal{H}$ does in approximating functions is critical to understanding the advantages and disadvantages of such a choice. Giving precise quantitative estimates to (1.9) is the subject of *approximation theory* and is outside the scope of the thesis. For the interested reader we refer to (DeVore et al., 2004; DeVore, 1998).

In the previous section we introduce the RKHS and show that the RKHS norm $\|f\|_{\mathcal{H}}$ can be interpreted as a measure of a complexity (smoothness) of a function $f \in \mathcal{H}$. Think of f with many oscillations having RKHS norm large. Suppose we want to perform the empirical risk minimization over $\mathcal{H}$ based on the finite data $\mathbf{z}$:

$$f_{\mathbf{z},\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_{\mathbf{z}}(f).$$

As the space $\mathcal{H}$ is potentially big, the solution of the above problem is not unique. In fact, it can be any function that interpolates the data $\mathbf{z}$. Each of the interpolants would have different performance properties [1] therefore, without further restriction the ERM over $\mathcal{H}$ is not well-posed. Instead, we can restrict the set $\mathcal{H}$ over the functions where the norm does not exceed the certain threshold level g and perform the ERM over the smaller class of functions. This leads to the following constrained optimization problem

$$\begin{aligned} \min_{f\in\mathcal{H}} \quad & \frac{1}{n}\sum_{i=1}^{n} \left(f\left(x_i\right) - y_i\right)^2 \\ \text{s.t.} \quad & \|f\|_{\mathcal{H}}^2 \le g \end{aligned}$$

Instead of restricting the set of functions over which one minimizes, one can rewrite the above optimization problem in a more traditional Lagrangian form, where the constraint on the norm enters as a penalty term added to the empirical risk functional. This leads

[1] Notice that the interpolation does not lead to an unreasonable estimate. It was shown by Belkin et al. (2019) that the learning methods interpolating the training data can achieve optimal rates. Achievability was shown for the Nadaraya-Watson smoother with a singular smoothing kernel.

us to the **regularized kernel least squares** or **kernel ridge regression**

$$f_{\mathbf{z},\lambda} := \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(f\left(x_i\right) - y_i\right)^2 + \lambda\|f\|_{\mathcal{H}}^2 \right\}. \tag{1.10}$$

Here λ is the so-called *regularization parameter*. The first term measures the closeness to the data, while the second term penalizes "roughness" in the function measured by the RKHS norm, and λ establishes the trade-off between the two.

The optimization (1.10) in many practical problems is defined on an infinite-dimensional function space. In the case of Matérn kernel RKHS corresponds to the Sobolev space of functions. Remarkably, (1.10) has an expicit, finite-dimensional, unique solution given in the following proposition.

Proposition 1.5. *Let $\lambda > 0$. The solution to (1.10) can be expressed as*

$$f_{\mathbf{z},\lambda}(x) = \sum_{i=1}^{n} a_i K\left(x, x_i\right), \tag{1.11}$$

where

$$\left(a_1, \ldots, a_n\right)^{\top} := \left(K_{\mathbf{xx}} + n\lambda I\right)^{-1} \mathbf{y} \in \mathbb{R}^n.$$

Here $\mathbf{y} = \left(y_1, \ldots, y_n\right)^{\top} \in Y^n$.

1.4 Effective Dimension

Kernel ridge regression (1.11) introduced in the previous section can be equivalently written as

$$f_{\mathbf{z},\lambda}(x) = K_{x\mathbf{x}}\left(K_{\mathbf{xx}} + n\lambda I\right)^{-1} \mathbf{y},$$

where $K_{x\mathbf{x}} = \left(K\left(x, x_1\right), \ldots, K\left(x, x_n\right)\right)$. The vector $w(x) := \left(K_{\mathbf{xx}} + n\lambda I\right)^{-1} K_{x\mathbf{x}}^{\top}$ gives the weights applied to the output vector $\mathbf{y}$. So the KRR is the weighted average of the output vector, where the weights are defined by $w(x)$. However, understanding the form of the weight function is made complicated by the matrix inversion of $K_{\mathbf{xx}} + n\lambda I$ and the fact that $K_{\mathbf{xx}}$ depends on the specific locations of the n datapoints.

Spectral analysis of the kernel matrix $K_{\mathbf{xx}}$ helps to gain some intuition behind the weighting procedure in KRR. To this order we define the n-dimensional vector of fitted values at the training points $\mathbf{f} = (f_{\mathbf{z},\lambda}(x_1), \ldots, f_{\mathbf{z},\lambda}(x_n))^{\top}$. Then

$$\mathbf{f} = K_{\mathbf{xx}}\left(K_{\mathbf{xx}} + n\lambda I\right)^{-1} \mathbf{y}.$$

The matrix $L_{\mathbf{x}}(\lambda) := K_{\mathbf{xx}}\left(K_{\mathbf{xx}} + n\lambda I\right)$, linearly acting on the output vector $\mathbf{y}$, is called the *smoother matrix.* Important property of smoother matrix is its independence from $\mathbf{y}$; $L_{\mathbf{x}}(\lambda)$ depends only on the input data $\mathbf{x}$ and λ.

We define the *empirical effective dimension* as

$$\mathcal{N}_{\mathbf{x}}(\lambda) = \operatorname{Tr} L_{\mathbf{x}}(\lambda),$$

the sum of diagonal entries of $L_{\mathbf{x}}(\lambda)$. For the special case when $X = \mathbb{R}^d$, $K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}$ and $\lambda = 0$, the smoother matrix $L_{\mathbf{x}}$ is a projection operator and the trace gives the dimension of the projection space and coincides with the number of parameters involved in the estimation.

Using the eigen-decomposition of the Gramm matrix $K_{\mathbf{xx}} = \sum_{i=1}^{n} \mu_i \phi_i \phi_i^\top$, where μ_i is the ith eigenvalue (real and non-negative) and $\phi_i \in \mathbb{R}^n$ is the corresponding eigenvector, we can write

$$\mathbf{f} = \sum_{i=1}^{n} \frac{\mu_i}{\mu_i + n\lambda} \phi_i \langle \phi_i, \mathbf{y} \rangle \quad \text{and} \quad \mathcal{N}_{\mathbf{x}}(\lambda) = \sum \frac{\mu_i}{\mu_i + n\lambda}.$$

Notice that the eigenvectors are not affected by the regularization parameter λ. Therefore, the KRR computes the coordinates of $\mathbf{y}$ w.r.t. the basis $\{\phi_i\}$ formed by the eigenvectors of the Gramm matrix. When $\mu_i / \left(\mu_i + \sigma_n^2\right)$ is small the component in $\mathbf{y}$ along the eigenvector ϕ_i is effectively eliminated.

1.5 KRR - Function Reconstruction Point of View

The learning problem can be formulated as a problem of a function reconstruction, where the function values are available only on the finite set in its domain. Adopting the terminology from the signal processing by *sampling* we call the process of converting continuous function into a sequence of values. In signal processing reconstructing the continuous-time signals from the samples (discrete-time signals) is of great importance. The classical *Whittaker-Shannon-Nyquist Sampling Theorem* gives conditions on a function on $\mathbb{R}$ so that it can be perfectly reconstructed from its sampling values at integer points:

Shannon's Theorem. *If a function $f \in L^2(\mathbb{R}, \upsilon)$ has its Fourier transform supported on $[-\pi, \pi]$, then*

$$f(x) = \sum_{t \in \mathbb{Z}} f(t) \phi(x - t), \tag{1.12}$$

where $\phi(x) = \frac{\sin \pi x}{\pi x}$.

One can immediately notice the analogy between (1.12) and the solution (1.11) of the KRR. The function $\phi(x-t)$ in the Shannon theorem corresponds to the feature map $K(t,x)$ and the solutions in both cases are represented by respective linear combinations.

Now let us introduce the regularized least squares algorithm from the Shannon sampling theory point of view. To this order we introduce the *sampling operator* $S_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}^n$ associated with a set $\mathbf{x} = \{x_1, \dots, x_n\} \subset X$ as

$$(S_{\mathbf{x}} f)_i = f\left(x_i\right) = \langle f, K_{x_i} \rangle_{\mathcal{H}} .$$

Its adjoint operator $S_{\mathbf{x}}^{\top} \colon \mathbb{R}^n \to \mathcal{H}$ is given by

$$S_{\mathbf{x}}^{\top}(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} y_i K_{x_i}.$$

The sampling problem we are interested in is: *reconstruct the function f_ρ based on the* $\mathbf{y} \in \mathbb{R}^n$. Note that in the Shannon case the values of the function are known exactly, while in the learning theory they are realizations from the distribution $\rho(y|x)$.

Intuitively, the function f provides a good reconstruction to f_ρ if its values on the set $\mathbf{x}$ are close to the actually observed ones, this is $S_{\mathbf{x}} f \approx \mathbf{y}$. From this we can deduce that for all g in $\mathcal{H}$

$$\langle S_{\mathbf{x}} f, S_{\mathbf{x}} g \rangle_{\mathbb{R}^n} \approx \langle \mathbf{y}, S_{\mathbf{x}} g \rangle_{\mathbb{R}^n} \implies \langle S_{\mathbf{x}}^{\top} S_{\mathbf{x}} f, g \rangle_{\mathcal{H}} \approx \langle S_{\mathbf{x}}^{\top} \mathbf{y}, g \rangle_{\mathcal{H}}.$$

As the approximate equality above holds for all $g \in \mathcal{H}$ we should expect that $S_{\mathbf{x}}^{\top} S_{\mathbf{x}} f \approx S_{\mathbf{x}}^{\top} \mathbf{y}$ or

$$f = \left(S_{\mathbf{x}}^{\top} S_{\mathbf{x}} + \lambda I \right)^{-1} S_{\mathbf{x}}^{\top} \mathbf{y}$$

where the regularization term λI is added to avoid the invertibility issues of $S_{\mathbf{x}}^{\top} S_{\mathbf{x}}$. As the following reconstruction theorem suggests, the heuristic just described is closely related to the minimization problem (1.10).

Theorem 1.6. *If $S_{\mathbf{x}}^{\top} S_{\mathbf{x}} + \lambda I$ is invertible, then $f_{\mathbf{z},\lambda}$ exists, is unique and*

$$f_{\mathbf{z},\lambda} = \left(S_{\mathbf{x}}^{\top} S_{\mathbf{x}} + \lambda I \right)^{-1} S_{\mathbf{x}}^{\top} \mathbf{y}.$$

Proof. Note that the regularizes loss in (1.10) can be rewritten in terms of the operators $S_{\mathbf{x}}$ and $S_{\mathbf{x}}^{\top}$ as follows

$$\mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_{\mathcal{H}}^2 = \left\langle \left(S_{\mathbf{x}}^{\top} S_{\mathbf{x}} + \lambda I \right) f, f \right\rangle_{\mathcal{H}} - 2 \left\langle S_{\mathbf{x}}^{\top} \mathbf{y}, f \right\rangle_{\mathcal{H}} + \|\mathbf{y}\|.$$

Taking the functional derivative of this objective function with respect to $f \in \mathcal{H}$, setting it equal to 0 and arranging the resulting equation yields

$$\left(S_{\mathbf{x}}^{\top} S_{\mathbf{x}} + \lambda I\right) f = S_{\mathbf{x}}^{\top} \mathbf{y}.$$

Solving the system of equations above w.r.t. f finishes the proof. □

Remark 1.7. For $\lambda \searrow 0$, $f_{\mathbf{z},\lambda}$ corresponds to the Moore–Penrose solution of the operator equation $S_{\mathbf{x}} f = \mathbf{y}$.

Let us fix $\lambda > 0$ and let the number of points n increase to infinity. By the low of large numbers the objective functional in (1.10) becomes

$$f_{\lambda} := \underset{f \in \mathcal{H}}{\arg\min} \left\{ \int_{X \times Y} \left(f\left(x\right) - y\right)^2 d\rho(x,y) + \lambda \|f\|_{\mathcal{H}}^2 \right\}. \tag{1.13}$$

Note that $\mathcal{H}$ is not compact, therefore the existence of f_{λ} is not immediate. Our next result proves that f_{λ} exists and is unique.

Proposition 1.8. *For all $\lambda > 0$ the function*

$$f_{\lambda} = \left(T + \lambda\right)^{-1} L f_{\mathcal{H}} \tag{1.14}$$

is the unique solution of (1.13).

Proof. Note that (1.13) is equivalent to

$$f_{\lambda} = \underset{f \in \mathcal{H}}{\arg\min} \left\{ \|f - f_{\rho}\|_{\rho_X}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\} = \underset{f \in \mathcal{H}}{\arg\min} \left\{ \|f - f_{\mathcal{H}}\|_{\rho_X}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

where the last equality comes from the equality $\|f - f_{\rho}\|_{\rho_X}^2 = \|f - f_{\mathcal{H}}\|_{\rho_X}^2 + \|f_{\rho} - f_{\mathcal{H}}\|_{\rho_X}^2$ and the fact that $\|f_{\rho} - f_{\mathcal{H}}\|_{\rho_X}^2$ is independent form f. Now, consider the functional

$$\varphi(f) = \|f - f_{\mathcal{H}}\|_{\rho_X}^2 + \lambda \left\| T^{-1/2} f \right\|_{\rho_X}^2.$$

If a point f_{λ} minimizes φ, then it must be a zero of the derivative $D\varphi$. That is, f_{λ} satisfies $\left(I + \lambda T^{-1}\right) f_{\lambda} = f_{\mathcal{H}}$, which implies $f_{\lambda} = \left(T + \lambda I\right)^{-1} T f_{\mathcal{H}}$. □

The basic question one can ask is: *for the fixed regularization parameter, how well f_{λ} approximates the target function $f_{\mathcal{H}}$.*

Theorem 1.9. *Define f_{λ} by* (1.14). *If $L^{-r} f_{\mathcal{H}} \in L^2(X, \rho_X)$ for some $0 < r \leqslant 1$, then*

$$\|f_{\lambda} - f_{\mathcal{H}}\|_{\rho} \leqslant \lambda^r \left\| L^{-r} f_{\mathcal{H}} \right\|_{\rho} \tag{1.15}$$

moreover

$$\|f_\lambda\|_{\mathcal{H}} \leq \left\|L^{-r} f_{\mathcal{H}}\right\|_\rho, \quad \textit{for} \quad r \in (0.5, 1]. \tag{1.16}$$

Proof. By the identity $A(A+\lambda I)^{-1} = I - \lambda(A+\lambda)^{-1}$ valid for $\lambda > 0$ and any bounded self-adjoint positive operator, we have

$$\left(I - (T+\lambda)^{-1}T\right) f_{\mathcal{H}} = \lambda(T+\lambda)^{-1} f_{\mathcal{H}} = \lambda^r \left(\lambda^{1-r}(T+\lambda)^{-(1-r)}\right)\left((T+\lambda)^{-r}L^r\right) L^{-r} f_{\mathcal{H}}.$$

From the equality above by taking the norm we have

$$\|f_\lambda - f_{\mathcal{H}}\|_{\rho_X} \leq \lambda^r \left\|\lambda^{1-r}(T+\lambda)^{-(1-r)}\right\|_{\text{op}} \left\|(T+\lambda)^{-r}L^r\right\|_{\text{op}} \left\|L^{-r} f_{\mathcal{H}}\right\|_{\rho_X}.$$

Note that $\left\|\lambda^{1-r}(T+\lambda)^{-(1-r)}\right\|_{\text{op}} \leq 1$ and $\|(T+\lambda)^{-r}L^r\|_{\text{op}} \leq 1$.

Regarding the second estimate, if $r > 1/2$, since $\|T\|_{\text{op}} \leq 1$, we obtain,

$$\begin{aligned}
\|f_\lambda\|_{\mathcal{H}} &= \left\|(T+\lambda)^{-1} L f_{\mathcal{H}}\right\|_{\mathcal{H}} \\
&= \left\|(T+\lambda)^{-1} L L^{r-\frac{1}{2}} L^{\frac{1}{2}-r} f_{\mathcal{H}}\right\|_{\mathcal{H}} \\
&\leq \|L\|_{\text{op}}^{r-\frac{1}{2}} \left\|L^{-r} f_{\mathcal{H}}\right\|_{\rho_X}.
\end{aligned}$$

□

1.6 Generalization performance of KRR - Kernel independent case

We apply the following Bernstein inequality (Caponnetto and De Vito, 2007, Proposition 2) for Hilbert space valued random variables to provide the upper rate on $\|f_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X}$.

Proposition 1.10. *Let (Z, ρ) be a probability space and let ξ be a random variable on Z taking value in a real separable Hilbert space H. Assume that there are two positive constants L and σ such that*

$$\mathbb{E}\left[\|\xi - \mathbb{E}[\xi]\|_H^m\right] \leq \frac{1}{2} m! \sigma^2 L^{m-2}, \quad \forall m \geq 2 \tag{1.17}$$

then, for any $\delta \in (0, 1]$

$$\left\|\frac{1}{n}\sum_{i=1}^n \xi\left(z_i\right) - \mathbb{E}[\xi]\right\|_H \leq \frac{2L\log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}}$$

with probability at least $1-\delta$*. In particular,* (1.17) *holds if*

$$\|\xi\|_H \leq \frac{L}{2}, \quad \mathbb{E}\left[\|\xi\|_H^2\right] \leq \sigma^2.$$

The following is our first result on the learning properties of kernel ridge regression.

Theorem 1.11. *Let* $L^{-r} f_\rho \in L^2(X, \rho_X)$*, for* $r \in (0.5, 1]$ *and assume that the output is bounded,* $|y| \leq M$*. Furthermore, let* n *and* λ *satisfy the following condition*

$$\lambda = \left(\frac{8\log(6/\delta)}{\sqrt{n}}\right)^{\frac{2}{2r+1}} \tag{1.18}$$

Then, with probability greater than $1-\delta$*, it holds*

$$\|f_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X} \leq 3\,(M+R)\left(\frac{8\log\left(6/\delta\right)}{\sqrt{n}}\right)^{\frac{2r}{2r+1}}. \tag{1.19}$$

Proof. We provide the proof here as it is quite instructive. Most of the proofs used in this thesis will follow the same line of arguments. In order to establish upper bounds, we split $\|f_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X}$ into two parts:

$$\|f_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X} \leq \|f_{\mathbf{z},\lambda} - f_\lambda\|_{\rho_X} + \|f_\lambda - f_{\mathcal{H}}\|_{\rho_X}, \tag{1.20}$$

the *estimation error* $\|f_{\mathbf{z},\lambda} - f_\lambda\|_{\rho_X}$ and the *approximation error* $\|f_\lambda - f_{\mathcal{H}}\|_{\rho_X}$. A bound on the approximation error has already been given in Theorem 1.9. To obtain an upper bound on estimation error we use the following decomposition

$$\begin{aligned} f_{\mathbf{z},\lambda} - f_\lambda &= (T_{\mathbf{x}} + \lambda)^{-1} S_{\mathbf{x}}^\top \mathbf{y} - (T+\lambda)^{-1} T f_{\mathcal{H}} \\ &= (T_{\mathbf{x}} + \lambda)^{-1} \left\{ \left(S_{\mathbf{x}}^\top \mathbf{y} - T f_{\mathcal{H}}\right) + (T - T_{\mathbf{x}})\, f_\lambda \right\} \\ &= (T+\lambda)^{-\frac{1}{2}} \left\{ I - (T+\lambda)^{-\frac{1}{2}} (T - T_{\mathbf{x}})\,(T+\lambda)^{-\frac{1}{2}} \right\}^{-1} (T+\lambda)^{-\frac{1}{2}} \\ &\quad \left\{ \left(S_{\mathbf{x}}^\top \mathbf{y} - T f_{\mathcal{H}}\right) + (T - T_{\mathbf{x}})\, f_\lambda \right\}. \end{aligned} \tag{1.21}$$

where $T_{\mathbf{x}} := S_{\mathbf{x}}^\top S_{\mathbf{x}}$ and $T := T_{\rho_X}$. Assuming that

$$S_1 := \left\| (T+\lambda)^{-\frac{1}{2}} (T - T_{\mathbf{x}})\,(T+\lambda)^{-\frac{1}{2}} \right\|_{\mathrm{HS}} < 1, \tag{1.22}$$

where $\|A\|_{\mathrm{HS}}^2 = \operatorname{Tr}\left(A^\top A\right)$, and using the Neumann series expansion we obtain

$$\begin{aligned}
\left\|\left\{I-(T+\lambda)^{-\frac{1}{2}}\left(T-T_{\mathbf{x}}\right)(T+\lambda)^{-\frac{1}{2}}\right\}^{-1}\right\|_{\mathrm{op}} &= \left\|\sum_{n=0}^{\infty}\left[(T+\lambda)^{-\frac{1}{2}}\left(T-T_{\mathbf{x}}\right)(T+\lambda)^{-\frac{1}{2}}\right]^n\right\|_{\mathrm{op}} \\
&\leq \sum_{n=0}^{\infty}\left\|(T+\lambda)^{-\frac{1}{2}}\left(T-T_{\mathbf{x},\mathbf{w}}\right)(T+\lambda)^{-\frac{1}{2}}\right\|_{\mathrm{op}}^n \\
&\leq \sum_{n=0}^{\infty}\left\|(T+\lambda)^{-\frac{1}{2}}\left(T-T_{\mathbf{x},\mathbf{w}}\right)(T+\lambda)^{-\frac{1}{2}}\right\|_{\mathrm{HS}}^n \\
&\leq \frac{1}{1-S_1}.
\end{aligned}$$

From (1.21) by taking the $L^2(X,\rho_X)$ norm from both sides and using the isometry property we get

$$\|f_{\mathbf{z},\lambda}-f_\lambda\|_{\rho_X} = \left\|T^{\frac{1}{2}}\left(f_{\mathbf{z},\lambda}-f_\lambda\right)\right\|_{\mathcal{H}} \leq \frac{S_2+S_3}{1-S_1}, \tag{1.23}$$

where

$$\begin{aligned}
S_2 &:= \left\|(T+\lambda)^{-\frac{1}{2}}\left(S_{\mathbf{x}}^\top \mathbf{y} - Lf_{\mathcal{H}}\right)\right\|_{\mathcal{H}}, \\
S_3 &:= \left\|(T+\lambda)^{-\frac{1}{2}}\left(T-T_{\mathbf{x}}\right)f_\lambda\right\|_{\mathcal{H}}.
\end{aligned}$$

We have to find an upper bound for each of S_i. To do so, notice that

$$S_i = \left\|\frac{1}{n}\sum_{k=1}^{n}\xi_k - \mathbb{E}[\xi_k]\right\|_F, \quad i=1,2,3,$$

with appropriate choice of the random variable ξ and the norm $\|\cdot\|_F$. Indeed, in order to let the equality above hold, we define the operator valued random variable $\xi_1 : X \to \mathrm{HS}\left(\mathcal{H}\right)$, where $\mathrm{HS}\left(\mathcal{H}\right)$ is the space of Hilbert-Schmidt operators on $\mathcal{H}$, as follows

$$\xi_1(x)[\cdot] = (T+\lambda)^{-\frac{1}{2}}K_x\left\langle K_x,\cdot\right\rangle_{\mathcal{H}}(T+\lambda)^{-\frac{1}{2}}.$$

Moreover, $\xi_2 : Z \to \mathcal{H}$ is defined by

$$\xi_2(x,y) = (T+\lambda)^{-\frac{1}{2}}K_x y.$$

Finally, $\xi_3 : X \to \mathcal{H}$ is defined by

$$\xi_3(x) = (T+\lambda)^{-\frac{1}{2}}K_x f_\lambda(x).$$

Application of Proposition 1.10 to each of S_i yields to the following bounds with probability at least $1 - \delta/3$

$$S_i \leq \frac{2L_i \log(6/\delta)}{n} + \sigma_i \sqrt{\frac{2\log(6/\delta)}{n}} \tag{1.24}$$

where, as it can be straightforwardly verified, the constants L_i and σ_i are given by the expressions

$$L_1 = \frac{2}{\lambda}, \quad \sigma_1^2 = \frac{1}{\lambda^2}, \tag{1.25}$$

$$L_2 = \frac{2M}{\sqrt{\lambda}}, \quad \sigma_2^2 = \frac{M^2}{\lambda}, \tag{1.26}$$

$$L_3 = \frac{2\|f_\lambda\|_{\mathcal{H}}}{\sqrt{\lambda}}, \quad \sigma_3^2 = \frac{\|f_\lambda\|_{\mathcal{H}}}{\lambda}. \tag{1.27}$$

Let us verify (1.22). From (1.25) and Proposition 1.10, with probability greater than $1 - \delta/3$, we have

$$S_1 \leq 2\log\left(\frac{6}{\delta}\right)\left(\frac{2}{n\lambda} + \frac{1}{\sqrt{n}\lambda}\right) \tag{1.28}$$

Choosing $\lambda\sqrt{n} \geq 8\log\left(\frac{6}{\delta}\right)$ we get $S_1 \leq 3/4 \implies 1/(1 - S_1) \leq 4$. From this and Proposition 1.10 applied to the terms S_1 and S_2 we can write,

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_{\rho_X} \leq \frac{S_2 + S_3}{1 - S_1} \leq 8\log\left(\frac{6}{\delta}\right)(M + \|f_\lambda\|_{\mathcal{H}})\left(\frac{2}{n\sqrt{\lambda}} + \frac{1}{\sqrt{n\lambda}}\right) \tag{1.29}$$

with probability at least $1 - \delta$.

Combining the decomposition in (1.20) with the bounds (1.29), (1.15) and choosing λ as in (2.21) we finally get

$$\begin{aligned}
\|f_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X} &\leq 8\log\left(\frac{6}{\delta}\right)(M + R)\left(\frac{2}{n\sqrt{\lambda}} + \frac{1}{\sqrt{n\lambda}}\right) + \lambda^r R \\
&\leq 8\log\left(\frac{6}{\delta}\right)(M + R)\,\lambda^r\left(\frac{2}{n\lambda^{r+1/2}} + \frac{1}{\sqrt{n}\lambda^{r+1/2}}\right) \\
&\leq 3\,(M + R)\,\lambda^r,
\end{aligned}$$

with probability at least $1 - \delta$.

□

1.7 Generalization performance of KRR - Kernel dependent case

The generalization bound of the previous section is kernel independent, except the requirement $\|L^{-r}f_{\mathcal{H}}\|_{\rho_X} < \infty$. However, most of the kernels used in practice usually have some extra regularity properties. For instance a Matérn kernel 1.5 with $\alpha \to \infty$ is infinitely differentiable as well as the functions in the corresponding reproducing kernel Hilbert space. For the regression function belonging to this space faster rates can be obtained.

The smoothness of the kernel is better described by the decay rate of the eigenvalues in the expansion (1.5). For $0 < s \leq 1$ we assume that the *eigenvalue decay* satisfies a polynomial upper bound of order $1/s$:

$$\mu_i \sim O\left(i^{-\frac{1}{s}}\right), \quad \forall i \in \mathbb{N}. \tag{1.30}$$

The above eigenvalue decay measures the smoothness (capacity) of the corresponding reproducing space. The smaller the value of s, the smoother the reproduction space is,while for $s = 1$ the condition (1.30) is satisfied for any bounded kernels. A classical example is the Matérn kernel of smoothness $\beta - d/2$, in which case $s = d/(2\beta)$ and condition (1.30) is equivalent to assuming $\mathcal{H}$ to be a Sobolev space.

Central concept to obtain tighter bounds is the *effective dimension* $\mathcal{N}_\rho : (0, \infty) \to [0, \infty)$ defined by

$$\mathcal{N}_\rho(\lambda) := \mathrm{Tr}\left((T_\rho + \lambda)^{-1} T_\rho\right) = \sum_{i\geq 1} \frac{\mu_i}{\mu_i + \lambda}.$$

Proposition 1.12. *Let $\mathcal{H}$ be a separable RKHS on X and ρ is the probability measure on X. Then the condition (1.30) is equivalent to the following upper bound of the effective dimension*

$$\mathcal{N}_\rho(\lambda) \leq C\lambda^{-s}. \tag{1.31}$$

Proof. The proof can be found in (Fischer and Steinwart, 2020, Lemma 11). □

Theorem 1.13. *Let K be a bounded measurable kernel on X with $\|K\|_\infty = 1$ and separable RKHS $\mathcal{H}$. Moreover, let ρ be a distribution on $X \times [-M, M]$, where $M > 0$ is some constant. Assume that the extended sequence of eigenvalues of the integral operator L satisfies the assumption (1.30). Let $L^{-r}f_{\mathcal{H}} \in L^2(X, \rho_X)$, for $r \in (0.5, 1]$. Furthermore, let n and λ satisfy the following condition*

$$\lambda = \left(\frac{8\log(6/\delta)}{\sqrt{n}}\right)^{\frac{2}{2r+s}} \tag{1.32}$$

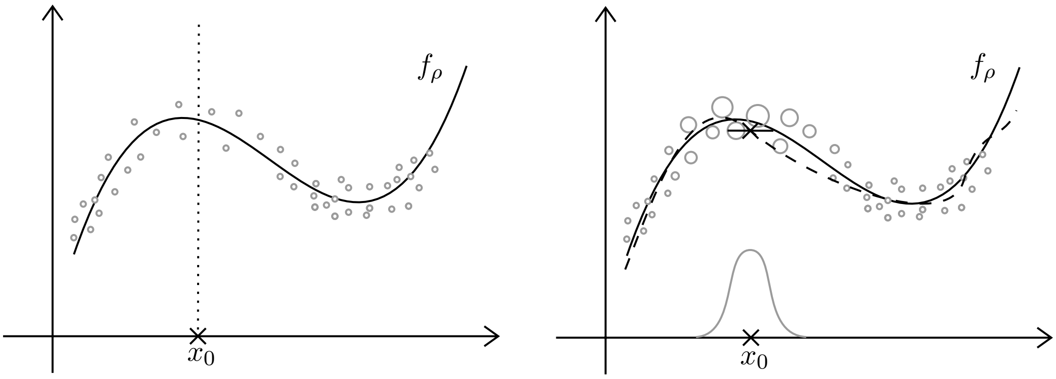


FIGURE 1.3: The definition of the regression function as a conditional average can not be directly applied for the learning problems with a finite data (left panel). Nadaraya-Watson regression takes a weighted average of the training output points with the weights defined by the kernel function k (right panel).

Then, with probability greater than $1-\delta$, *it holds*

$$\|f_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X} \le 3\,(M+R)\left(\frac{8\log(6/\delta)}{\sqrt{n}}\right)^{\frac{2r}{2r+s}}. \tag{1.33}$$

Proof. The proof follows the same line of arguments as for Theorem 1.11 except the variance part σ_i^2 in (1.25), (1.26),(1.27). Loosely speaking the term $1/\lambda$ in all of the σ_i^2 should be replaced by $1/\lambda^s$, which is the upper bound on the effective dimension. It results in a better concentration for each of S_i. □

1.8 Local Methods of Regression - Kernel Smoothing

In the previous sections we describe the kernel ridge regression and gave the function reconstruction point of view. The algorithmic idea was inspired by the variational formulation (1.2) of the regression function. In this section we describe a class of regression function estimation techniques that achieve the flexibility in estimation the regression function f_ρ over the domain $X = \mathbb{R}^d$ by estimation a different but simple models separately at each query point x_0. Recall that the regression function is the conditional expectation of the output for a given input point x_0 :

$$f_\rho(x_0) = \int_Y y d\rho(y|x_0). \tag{1.34}$$

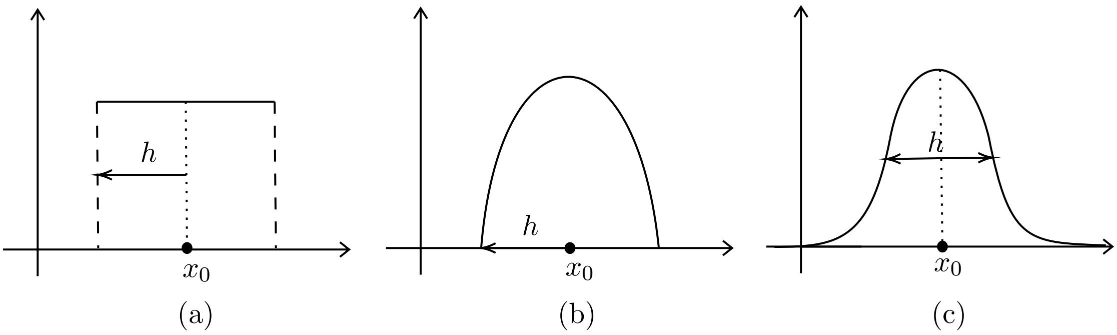


FIGURE 1.4: Examples of local kernels: (a) Rectangular kernel $k(x) = \mathbb{I}(\|x\| \leq 1)$. (b) Epanechnikov kernel $k(x) = \frac{d+2}{2V_d}(1 - \|x\|^2)\mathbb{I}(\|x\| \leq 1)$. (c) Gaussian kernel $k(x) = \frac{1}{2\pi}e^{-\|x\|^2}$.

This definition is our starting point to define a *local kernel smoother.* Note that the definition (1.34) can not be applied directly as for a given target point x_0 there could be no corresponding responses (see the left panel of Figure 1.3). The idea is to relax the definition of the conditional expectation, as illustrated in the right panel of Figure 1.3, and compute the weighted average in the neighbourhood of the target point. Weights are assigned to the data points according to the smoothing kernel [2] k, leading to so called Nadaraya-Watson regression defined as follows

$$f^{\mathrm{NW}}_{\mathbf{x},h}(x_0) = \sum_{i=1}^{n} \frac{y_i k\left(\frac{x_0 - x_i}{h}\right)}{\sum_{j=1}^{n} k\left(\frac{x_0 - x_j}{h}\right)}, \tag{1.35}$$

where $k : \mathbb{R}^d \to \mathbb{R}$ is an integrable function satisfying $\int k(x)dx = 1$. Some classical examples of smoothing kernels are illustrated in Figure 1.4.

Note that unlike KRR, where the regression function is reconstructed over the entire domain X, local kernel smoother evaluates the regression function at a fixed target point x_0. Heuristic described in section 1.5 to reconstruct the function on RKHS $\mathcal{H}$, based on the operators $S_{\mathbf{x}}$ and $S_{\mathbf{x}}^\top$, can also be adapted for the Nadaraya-Watson estimator. To this order we define the corresponding operators as follows

$$S_{\mathbf{x}_0} f = (f(x_0), \ldots, f(x_0))^\top \in \mathbb{R}^n, \quad \text{and} \quad S_{\mathbf{x}}^\top a = \frac{1}{n}\sum_{i=1}^{n} a_i k_h(x_0 - x_i)$$

where $\mathbf{x}_0 = (x_0, \ldots, x_0)^\top \in \mathbb{R}^n$ and $k_h(x) := \frac{1}{h}k(x/h)$. Then, Nadaraya-Watson estimator follows from the equality $S_{\mathbf{x}}^\top S_{\mathbf{x}_0} f = S_{\mathbf{x}}^\top \mathbf{y}$ derived in section 1.5. Indeed, we

[2] Should not be confused with the reproducing kernels. Smoothing kernels are not positive definite in general.

have

$$S_{\mathbf{x}}^T S_{\mathbf{x}_0} f = f(x_0)\frac{1}{n}\sum_{i=1}^{n} k_h(x_0 - x_i) = \frac{1}{n}\sum_{i=1}^{n} y_i k_h(x_0 - x_i) \implies f(x_0) = \frac{\sum_{i=1}^n y_i k_h\left(x_0 - x_i\right)}{\sum_{i=1}^n k_h\left(x_0 - x_i\right)}.$$

1.9 Local polynomial estimators

Nadaraya-Watson estimator (1.35) can be obtained by a local constant least squares approximation of the outputs y_i :

$$f_{\mathbf{x},h}^{\mathrm{NW}}(x_0) = \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^{n} \left(y_i - \theta\right)^2 k_h\left(x_i - x_0\right).$$

The kernel k assigns small weights to the points x_i far away from x_0, making the contribution of these points in the least squares small. The constant θ is the parameter to be fitted. Instead of locally fitting a constant to the data, we can locally fit a more general function, which depends on several parameters. The most popular example is the *local polynomial kernel estimate.*

Suppose that the regression function belongs to the Hölder class defined below.

Definition 1.14. The Hölder class $\mathcal{F}_\beta(L)$ on $\mathbb{R}^d$ is defined as the set of $\ell = \lfloor\beta\rfloor$ times differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ whose derivative $D^s f(x)$ atisfying the the following inequality

$$\left| f(x) - \sum_{|\boldsymbol{m}|\le l} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(x)(x - x')^{\boldsymbol{m}} \right| \le L|x - x'|^\beta, \quad \forall x, x' \in \mathbb{R}^d,$$

where $\boldsymbol{m} = (m_1, \ldots, m_d)$ is a d-dimensional multi-index and $D^{\boldsymbol{m}}$ is a differentiation operator $D^{\boldsymbol{m}} = \frac{\partial^{|\boldsymbol{m}|}}{\partial x_1^{m_1} \ldots \partial x_d^{m_d}}$.

Functions from the Hölder class can be well approximated by the Taylor polynomial of order l in the neighbourhood of the target point z as follows

$$f(x) \approx \sum_{|\boldsymbol{m}|\le l} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(z)(z - x)^{\boldsymbol{m}} = \theta^\top(z) U\left(\frac{x - z}{h}\right) \quad ,$$

where x is sufficiently close to z and

$$U(u) = \left(\frac{u^{\boldsymbol{m}^{(1)}}}{\boldsymbol{m}^{(1)}}, \ldots, \frac{u^{\boldsymbol{m}^{(S)}}}{\boldsymbol{m}^{(S)}}\right)^\top, \quad \theta(z) \quad = \left(h^{|\boldsymbol{m}^{(1)}|} D^{\boldsymbol{m}^{(1)}} f(z), \ldots, h^{|\boldsymbol{m}^{(S)}|} D^{\boldsymbol{m}^{(S)}} f(z)\right)^\top \quad ,$$

with $S = \text{card}\{\boldsymbol{m} : |\boldsymbol{m}| \leq l\}$, where the numeration is such that $\boldsymbol{m}^{(1)} = (0, \ldots, 0), \boldsymbol{m}^{(2)} = (1, 0, \ldots, 0), \ldots, \boldsymbol{m}^{(d+1)} = (0, \ldots, 0, 1)$. After approximating the objective function by its Taylor expansion, we define the *local polynomial estimator* of $\theta(z)$ (see e.g. (Tsybakov, 2009)) as follows

$$\hat{\theta}_n(z) = \arg\min_{\theta \in \mathbf{R}^S} \sum_{i=1}^{n} \left[y_i - \theta^\top U\left(\frac{x_i - z}{h}\right)\right]^2 k_h\left(x_i - z\right) \ .$$

Being the weighted least squares estimator, $\hat{\theta}_n(z)$ can be expressed analytically as

$$\hat{\theta}_n(z) = B_n(z)^{-1} D_n(z) \ ,$$

where the matrix $B_n(z)$ and the vector $D_n(z)$ are defined as

$$B_n(z) = \frac{1}{nh^d} \sum_{i=1}^{n} U\left(\frac{x_i - z}{h}\right) U^\top\left(\frac{x_i - z}{h}\right) k\left(\frac{x_i - z}{h}\right) \ ,$$
$$D_n(z) = \frac{1}{nh^d} \sum_{i=1}^{n} y_i U\left(\frac{x_i - z}{h}\right) k\left(\frac{x_i - z}{h}\right) \ .$$

The *local polynomial estimator* of $f_\rho(z)$ is simply the first coordinate of the vector $\hat{\theta}_n(z)$

$$f_n(z) = U(0)\hat{\theta}_n(z).$$

Furthermore, properly normalized coordinates of $\hat{\theta}_n(z)$ provide estimators of the derivatives $D^{\boldsymbol{m}^{(2)}} f(z), \ldots, D^{\boldsymbol{m}^{(S)}} f(z)$. The gradient method described in chapter 5 is based on this simple observation.

1.10 Bayesian Methods

Kernel ridge regression is closely related to the *Gaussian process regression*, where Gaussian processes are random functions serving as priors on function spaces.

Definition 1.15. A Gaussian process is a set of random variables $f(x)$ indexed by the input set X such that for each finite subset $\{x_1, \ldots, x_m\} \subset X$ the random vector $(f(x_1), \ldots, f(x_m))^\top$ is a multivariate normal.

The connection between KRR and GPR is a consequence of the duality between the Hilbert space spanned by GP and its associated RKHS. Let us describe it in more detail.

The finite-dimensional distributions of a Gaussian process are determined by the *mean function* and *covariance kernel*, defined by

$$m(x) = \mathbf{E}f(x), \quad K(x,x') = \mathbf{E}[f(x)f(x')]. \tag{1.36}$$

When $m(x) = 0$ for all $x \in X$, we call the process $f(x)$ zero-mean GP.

Let $f(x),\ x \in X$ be a zero-mean GP with covariance kernel $\mathbf{E}f(x)f(x') = K(x,x')$. Let $\mathbb{H}$ be the completion of the linear space of all random variables

$$\xi = \sum_{i=1}^{n} a_i f(x_i), \quad a_1, \dots, a_n \in \mathbb{R},\ x_1, \dots, x_n \in X,\ n \in \mathbb{N}$$

relative to the norm induced by the inner product $\langle \xi_1, \xi_2 \rangle_{\mathbb{H}} = \mathbf{E}(\xi_1 \xi_2)$. If $\mathcal{H}_K$ is RKHS with reproducing kernel K then the map $f(x) \mapsto K_x$ from $\mathbb{H}$ to $\mathcal{H}$ is one to one and preserves the scalar product:

$$\langle f(x), f(x') \rangle_{\mathbb{H}} = \mathbf{E}f(x)f(x') = K(x,x') = \langle K_x, K'_x \rangle_{\mathcal{H}}.$$

So the spaces $\mathbb{H}$ and $\mathcal{H}$ are isometrically isomorphic.

Consider zero-mean GP $f(x),\ x \in X$ with a covariance kernel K. Define another Gaussian process $y(x) = f(x)+\varepsilon$, which is the corrupted version of the process $f(x)$, corrupted by Gaussian noise ε with variance σ^2. Let us fix $x \in X$ and compute $\mathbf{E}[f(x)|y(x_i) = y_1, \dots, y(x_n) = y_n]$. The joint distribution of $f(x), y_1, \dots, y_n$ is zero-mean Gaussian with covariance matrix given by

$$\begin{pmatrix} K(x,x) & K\left(x,x_1\right), \dots, & K\left(x,x_n\right) \\ K\left(x,x_1\right) & & \\ \vdots & K_{\mathbf{xx}} + \sigma^2 I & \\ K\left(x,x_n\right) & & \end{pmatrix}$$

where $K_{\mathbf{xx}}$ is a Gram matrix of $\mathbf{x} \in X^n$. Using properties of the multivariate normal distribution, as given, e.g., in (see e.g. Rasmussen and Williams 2006, Appendix A.2), we have

$$\mathbf{E}[f(x)|y(x_i) = y_1, \dots, y(x_n) = y_n] = K_{x\mathbf{x}}\left(K_{\mathbf{xx}} + \sigma^2 I\right)^{-1}\mathbf{y}$$

where $K_{x\mathbf{x}} = \left(K\left(x,x_1\right), \dots, K\left(x,x_n\right)\right)$. This is exactly (1.11) with $\sigma^2 = \lambda n$.

1.11 Notes

For almost a century, the theory of reproducing kernels has been applied in various fields of pure mathematics. The theory of RKHSs was introduced by Aronszajn (1950). In the statistical literature, the reproducing kernels were first introduced in the context of time series analysis Parzen (1961, 1962b, 1963). The *smoothing spline* as an optimization problem in an RKHS were introduced by Kimeldorf and Wahba (1971, 1970). The latter article provides a closed form expression for the solution to the regulirized least squares problem, where the regularization is a square norm or seminorm in an RKHS - *the representer theorem.* In Kimeldorf and Wahba (1970) connection between Gaussian processes and spline methods was established. Kanagawa et al. (2018) provides a more recent overview of the connection between Bayesian and frequentist approaches.

The Nadaraya-Watson estimator is proposed by Nadaraya (1964) and Watson (1964). An overview of the literature on this estimator as well as local polynomial estimators can be found, for example, in the books Fan and Gijbels (2018); Györfi et al. (2002); Hastie et al. (2009). For the minimax analysis we refer to the book Tsybakov (2009).

The generalization bounds of KRR presented in this chapter are based on operator and spectral methods developed for kernel-based algorithms (Blanchard and Mücke, 2018; Caponnetto and De Vito, 2007; Smale and Zhou, 2005, 2007; Steinwart et al., 2009; Zhang, 2002, 2005). The main differences between these methods and the more traditional covering number techniques used in (Cucker and Smale, 2002a; DeVore et al., 2004; Vapnik, 1998) consist in developing nonasymptotic upper rates estimates of integral operators via concentration inequalities. However, in the early investigation (De Vito et al., 2005a,b; Smale and Zhou, 2005, 2007) operator and spectral methods of analysis of kernel-based algorithms failed to compare with similar results obtained using entropy methods (DeVore et al., 2004). This gap was filled by (Caponnetto and De Vito, 2007) utilizing the notion of *effective dimension*, which encodes the crucial properties of the marginal distribution via integral operators.

The connection between learning and sampling theory was investigated in Smale and Zhou (2004). The inverse problem point of view to the regulirized least-squares was given in De Vito et al. (2005b). The main idea is to associate certain linear operator equation to the kernel least squares problem. It turns out that the regulirized least squares solution can be obtained as a Tikhonov solution of the linear operator equation. Similar arguments can also be found in (Vapnik, 1998, Appendix to chapter 1).

Chapter 2

Covariate shift

2.1 Introduction

In the previous chapter we describe the classical learning scenario when the training and testing distributions are the same. However, in many real-world applications of supervised learning, training and testing distributions are different. The most common setting in the literature is the one in which the conditional distributions of labels given inputs are the same but the marginal distributions over the inputs differ across training and testing instances. This situation is referred to as *covariate shift* (Shimodaira, 2000), which is a special case of sample selection bias (Heckman, 1979). Covariate shift naturally arises in many common learning scenarios. In active learning problems, the training data points are sampled by the learner at will, while the test data points are bounded to be sampled from the environment distribution (Cortes et al., 2008; MacKay, 1992; Pukelsheim, 2006). In domain adaptation the training data is drawn from a source domain that differs from the target domain, to which the learner is required to transfer its knowledge (Ben-David et al., 2007; Cortes and Mohri, 2014; Jiang and Zhai, 2007; Mansour et al., 2009b; Zhang et al., 2012). Covariance shift also occurs in off-policy reinforcement learning, when a learner is required to evaluate a policy using data generated by interacting with the environment using a different policy (Precup et al., 2000; Thomas et al., 2015).

A common approach to address covariate shift is to consider so-called *importance weighted* risk minimization. The idea is to correct the notion of risk in a way that matches the risk associated with the target distribution. While making the risk estimate unbiased seems natural, in some learning scenarios, it does not significantly improve over unweighted risk minimization, and it sometimes even negatively affects its performance (Cortes et al., 2010).

Cortes et al. (2010) provide an empirical and theoretical analysis of importance weighting (IW). As pointed out in this work, the weighting function $w(x)$ is unbounded, or extremely large, in many practical cases and IW leads to poor performance. Large values of $w(x)$ are unfortunately unavoidable whenever regions of the input space with high testing probability are not properly covered by the training measure. To measure the degree of a singularity of the testing measure with respect to the training one, the notion of *transfer-exponent* was introduced in Kpotufe and Martinet (2021). It was shown that under severe covariate shifts, learning becomes hard (in a minimax sense) irrespectively of the learning approach. With this result in mind, a natural question is whether IW adaptation (whenever IW is well defined) maintains optimal learning rates and simply affects the constants in the generalization bounds. It is not clear if the poor performance of IW correction observed by Cortes et al. (2010) is related to the hardness of the problem and not to the importance weighting approach itself.

Parametric models under covariate shift were studied in Shimodaira (2000) and it was shown that IW adaptation is the asymptotically optimal strategy (the importance weighted maximum likelihood estimator is consistent) when the target function does not belong to the hypothesis class, i.e., the model is *misspecified*. Although consistency is guaranteed by this modification, the weighted maximum likelihood estimator is no longer asymptotically efficient. For *well-specified* models the situation is different. In this case, the asymptotically optimal strategy is uniform weighting $w(x) = 1$. The case of model misspecification was further investigated by Wen et al. (2014).

Less is known for high-capacity models. The robustness to covariate shift of over-parameterized models was studied in Tripuraneni et al. (2021), in which kernel methods with random feature approximation were considered. In the context of over-parameterized deep neural networks optimized by stochastic gradient descent, it was empirically observed by Byrd and Lipton (2019) that IW impacts only the early stage of training and its impact diminishes after the model separates (in the classification settings) the training data. Later, Xu et al. (2021) provided theoretical insights into this phenomenon. Minimax results under covariate shift in nonparametric classification were given by Kpotufe and Martinet (2021). Nonparametric regression over the Hölder class was considered by Pathak et al. (2022) where a refined version of the singularity measure was given. In particular, the Nadaraya-Watson estimator was shown to be minimax optimal over the introduced class of training and testing measure families.

In this chapter, we study the theoretical properties of IW adaptation for kernel ridge regression (KRR) under covariate shift. Our technical tools are based on operator and spectral methods developed for kernel-based algorithms (Blanchard and Mücke, 2018;

Caponnetto and De Vito, 2007; Smale and Zhou, 2005, 2007; Steinwart et al., 2009; Zhang, 2002, 2005).

Covariate shift is a phenomenon affecting the marginal distributions over the inputs, therefore it is not surprising that the notion of effective dimension plays an important role in our investigation. For kernel methods, although the feature space is very large, only a few of the features with sufficiently large eigenvalues play a role in an actual fit. The number of these eigenvectors gives the effective dimension and is controlled through the regularization parameter. We show that in the case of bounded importance weights, importance weighted kernel ridge regression attains the optimal rate of convergence (in the minimax sense) known for the learning problems without covariate shift (Caponnetto and De Vito, 2007). We also show that the optimal regularization parameter for importance weighted KRR (IW-KRR) under covariate shift is large and scales according to the supremum norm of the importance weights. Relaxing the boundedness condition on the importance weights, the rates of convergence become capacity independent, matching those found in Smale and Zhou (2005, 2007). Furthermore, by extending the generalization bounds for IW-KRR to arbitrarily weighted KRR, we are able to highlight several factors that should be considered in order to obtain successful re-weighting procedures. As we will see, the approximation properties of the model class $\mathcal{H}$ play a crucial role, providing insights into popular re-weighting functions.

In the independent work of Ma et al. (2022) similar problems are discussed. For the bounded IW scenario minimax, optimal rates are provided. These are achieved by unweighted (uniformly weighted) KRR. Under weaker assumptions, namely boundedness of the second moment of the importance weights, the sub-optimality of the unweighted KRR was experimentally demonstrated. For this scenario, a clipped version of IW correction, with a carefully chosen truncation level and regularization parameter, was shown to be optimal (up to a logarithmic factor). Let us elaborate more on the similarities and differences of our work.

1. Our analysis emphasizes the importance of the boundedness condition of the importance weights in the case of IW-KRR. In Theorem 2.5 we show that even a slight relaxation of the boundedness condition leads to sub-optimal rates for IW-KRR. In the case of bounded importance weights, the rates of convergence of IW-KRR are the same as those found by Ma et al. (2022) for the uniformly weighted KRR.

2. We show that IW correction using the *clipped* weighting function achieves the optimal rates (without logarithmic factor), whenever the truncation level and regularization parameter are properly tuned. Similarly to Ma et al. (2022), uniform boundedness of eigenvalues is also required in our analysis. However, Theorem

2.14 relaxes this assumption by imposing a stronger moment condition on the importance weights.

3. Finally, our analysis differs from theirs. As discussed above, our technical tools are based on operator and spectral methods. The main ingredient to derive fast convergence rates consists in controlling the speed of *eigenvalue decay* in the variance part of estimation error bound through the effective dimension. In local methods, used in Ma et al. (2022), the estimates giving the optimal rates are based on a *local version of the Rademacher averages* where these are computed on a subset of functions with a small empirical error. The local averages applied to the kernel classes can be accurately described in terms of the kernel eigenvalues. Therefore, the spectral properties of kernels play a crucial role in both methods.

This chapter is structured as follows. In Section 2 we briefly recall the learning problem under covariate shift, we introduce some auxiliary notations and the importance weighted algorithms. Section 3 introduces the assumptions and the main result on the generalization of IW-KRR, followed by some remarks. In Section 4 we study the generalization properties of alternative reweighting algorithms, which allow us to provide insights into practically relevant weighting schemes. In Section 5, we consider the implication of our results in the context of classification. The final section provides computer simulations supporting our theoretical conclusions.

2.2 The Learning Problem under Covariate Shift

We consider the *covariate shift* setting of Shimodaira (2000), where $\rho^{\mathrm{tr}}(x,y)$ and $\rho^{\mathrm{te}}(x,y)$ share the same conditional distribution of the output y given the input x, but they differ in their marginal distributions on the input space X. More precisely, let $\rho(y|x)$ be the shared conditional distribution, and let $\rho_X^{\mathrm{tr}}(x)$ and $\rho_X^{\mathrm{te}}(x)$ be the marginal distributions of $\rho^{\mathrm{tr}}(x,y)$ and $\rho^{\mathrm{te}}(x,y)$ on X, respectively

$$\rho^{\mathrm{tr}}(x,y) = \rho(y|x)\rho_X^{\mathrm{tr}}(x), \quad \rho^{\mathrm{te}}(x,y) = \rho(y|x)\rho_X^{\mathrm{te}}(x). \tag{2.1}$$

Covariate shift refers to the setting in which $\rho_X^{\mathrm{tr}}(x)$ and $\rho_X^{\mathrm{te}}(x)$ differ.

In the following we consider the *regression* problem; while in Section 2.5 we extend the results to binary classification problems. The task is to estimate the regression function $f_\rho : X \to Y$ defined from the conditional distribution $\rho(y|x)$ as

$$f_\rho(x) := \int_Y y \; d\rho(y|x), \quad x \in X. \tag{2.2}$$

For this problem, we assume that we are given $n \in \mathbb{N}$ i.i.d. samples from the training distribution,

$$(x_1, y_1), \ldots, (x_n, y_n) \overset{i.i.d.}{\sim} \rho^{\mathrm{tr}}(x, y). \tag{2.3}$$

To compact the notation we denote the training set by $\mathbf{z} := \{z_1, \ldots, z_n\} \in Z^n$ where $z_i := (x_i, y_i)$.

The goal is to construct an estimator $f_{\mathbf{z}} : X \to Y$ of the regression function f_ρ based on the training data $\mathbf{z}$. In particular, we aim at making $f_{\mathbf{z}}$ a good approximation of f_ρ with respect to the testing input distribution ρ^{te}, *not* with respect to the training input distribution ρ^{tr}.

More precisely, for any given function $f : X \to Y$, we consider the following *risk* function

$$\mathcal{E}_{\rho^{\mathrm{te}}}(f) = \int_Z (f(x) - y)^2 d\rho^{\mathrm{te}}(x, y) = \int_X \int_Y (f(x) - y)^2 d\rho(y|x) \rho_X^{\mathrm{te}}(x), \tag{2.4}$$

and we wish to find an estimator $f_{\mathbf{z}}$ such that the risk $\mathcal{E}_{\rho^{\mathrm{te}}}(f_{\mathrm{z}})$ is small. It is well known that the minimizer of (2.4) over the space of square integrable functions is the regression function (2.2). However, the regression function (2.2) is unknown, and the goal of learning theory is to construct the function $f_{\mathbf{z}}$ from the finite-size sample set $\mathbf{z}$.

In the absence of covariate shift, the conventional approach consists in replacing the risk functional (2.4) with its finite sample approximation based on $\mathbf{z}$ (in this case sampled from $\rho_X^{\mathrm{te}} = \rho_X^{\mathrm{tr}}$)

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 , \tag{2.5}$$

and minimize (2.5) over some functional class $\mathcal{H}$ called the *model class* (Vapnik, 1998).

However, in the covariate shift scenario, the learning algorithm is bounded to approximate (2.4) based on a training set $\mathbf{z}$ sampled from ρ^{tr}. To overcome this limitation it is either possible to modify the training set $\mathbf{z}$ to make it resemble to samples coming from ρ^{te}, or to change the notion of risk (2.4). The latter approach leads to the definition of importance weighted risk. Under the absolute continuity assumption $d\rho^{\mathrm{te}} \ll d\rho^{\mathrm{tr}}$, we define the *importance weighting function*

$$w(x) = \frac{d\rho_X^{\mathrm{te}}(x)}{d\rho_X^{\mathrm{tr}}(x)}, \tag{2.6}$$

and the corresponding *importance weighted risk*

$$\mathcal{E}_{\rho^{\mathrm{te}}}(f) = \int_Z w(x)(f(x) - y)^2 d\rho^{\mathrm{tr}}(x, y). \tag{2.7}$$

Based on $\mathbf{z}$, we also define it empirical estimate

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{n}\sum_{i=1}^{n} w(x_i)(f(x_i) - y_i)^2. \tag{2.8}$$

In the following, we will study the properties of estimators that belong to the model class $\mathcal{H}$ associated with a reproducing kernel Hilbert space. We also assume that K is bounded, meaning that

$$\sup_{x\in X} K(x,x) \leq \kappa. \tag{2.9}$$

To avoid superfluous notations, we further assume $\kappa \leq 1$. This condition can always be achieved by properly scaling the kernel function.

2.2.1 Notations and Auxiliary Operators

Let ν be a measure under consideration (in the consequent it will be training, testing or other). By $L^2(X,\nu)$ we denote the Lebesgue spaces of square-integrable functions with respect to measure ν, with the norm given by

$$\|f\|_\nu = \left(\int_X f^2(x)d\nu(x))\right)^{\frac{1}{2}}.$$

For any function f for which the integral is finite, we set $\|f\|_{k,\nu} = \left(\int_X f^k(x)d\nu(x))\right)^{1/k}$.

Our analysis relies upon operators related to the RKHS that we introduce below. Crucial in our analysis are the *covariance* operator $T_\nu : \mathcal{H} \to \mathcal{H}$,

$$\left(T_\nu f\right)(x) = \int K\left(x', x\right) f(x')d\nu(x'), \tag{2.10}$$

and the *integral* operator $L_\nu : L^2(X,\nu) \to \mathcal{H}$,

$$\left(L_\nu f\right)(x) = \int K\left(x', x\right) f(x')d\nu(x'). \tag{2.11}$$

In both definitions the measure ν depends on the context. These operators have a similar definition, however they differ in their domains. Under the boundedness assumption (2.9), the covariance operator T_ν can be proved to be a positive trace class operator for any measure ν, namely,

$$\|T_\nu\|_{\text{op}} \leq \text{Tr}(T_\nu) = \int_X \text{Tr}(T_x)d\nu(x) \leq 1, \tag{2.12}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm from $\mathcal{H}$ to $\mathcal{H}$ and $T_x = K_x\langle K_x, \cdot\rangle_{\mathcal{H}}$. Positive trace class operators are known to have at most countably many non-zero eigenvalues, all being non-negative. We denote by $(\mu_i(T_\nu))_{i\geq 1}$ the ordered sequence of eigenvalues (with geometric multiplicities), possibly extended appending zeros in case of finitely many non-zero eigenvalues. From (2.12) we can conclude that the resulting sequence $(\mu_i(T_\nu))_{i\geq 1}$ is summable, since $\sum_{i=1}^{\infty}\mu_i(T_\nu) = \mathrm{Tr}(T_\nu) \leq 1$. Moreover, the spectral theorem gives

$$T_\nu = \sum_{i\geq 1}\mu_i\left\langle \cdot, \mu_i^{1/2}e_i\right\rangle_{\mathcal{H}}\mu_i^{1/2}e_i \quad \text{and} \quad L_\nu = \sum_{i\geq 1}\mu_i\langle \cdot, e_i\rangle_\nu e_i,$$

where $\{\mu_i^{1/2}e_i\}_{i=1}^{\infty}$ is an orthonormal basis (ONB) of $\mathrm{Ker}\, T_\nu^{\perp}$ and $\{e_i\}_{i=1}^{\infty}$ is an ONB of $(\ker L_\nu)^{\perp}$.

Finally, we define the *empirical covariance operator* $T_{\mathbf{x}} : \mathcal{H} \to \mathcal{H}$ such that $T_{\mathbf{x}} = S_{\mathbf{x}}^{\top}S_{\mathbf{x}}$. It can be shown that $T_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}K_{x_i}\langle K_{x_i}, \cdot\rangle_{\mathcal{H}}$ and, similar to (2.12), we then have

$$\|T_{\mathbf{x}}\|_{\mathrm{op}} \leq \mathrm{Tr}(T_{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n}K_{x_i}\langle K_{x_i}, \cdot\rangle_{\mathcal{H}} \leq 1. \tag{2.13}$$

2.2.2 Importance Weighted Risk Minimization Algorithm

We now introduce the importance weighted regularized least-squares algorithm (IW-RLS). The IW-RLS solution associated with the kernel K is the minimizer of the following weighted least-square optimization problem defined over the training set of samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{n}$ independently drawn according to ρ^{tr}

$$f_{\mathbf{z},\lambda}^{\mathrm{IW}} := \arg\min_{f\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}w(x_i)\left(f(x_i) - y_i\right)^2 + \lambda\|f\|_{\mathcal{H}}^2\right\} \tag{2.14}$$

where $\lambda = \lambda_n$ is any positive function of the number of examples n known as a *regularization parameter.*

In this section, we assume that the importance weighting function $w = d\rho_X^{\mathrm{te}}/d\rho_X^{\mathrm{tr}}$ is known and are mainly concerned to study the effects of the weights on generalization properties of IW-RLS. When the importance weights are not known, they can be estimated from the training and testing input data. An analysis of the effect of an error in the estimation of the reweighting function on the accuracy of the learning algorithm is given in Cortes et al. (2008).

We begin by describing the the data-free limit of (2.14). As we increase the number of training examples, $n \to \infty$, the functional in the minimization problem (2.14) becomes

$$\mathcal{E}_{\rho^{\mathrm{te}}}(f) + \lambda \|f\|_{\mathcal{H}}^2.$$

By the standard decomposition of $\mathcal{E}_\rho(f)$ we have

$$\mathcal{E}_{\rho^{\mathrm{te}}}(f) = \|f - f_\rho\|^2_{\rho_X^{\mathrm{te}}} + \mathcal{E}_{\rho^{\mathrm{te}}}(f_\rho).$$

As the last term is independent from f, in the data-free limit we have that (2.14) becomes

$$f_\lambda := \arg\min_{f \in \mathcal{H}} \left\{ \|f - f_\rho\|^2_{\rho_X^{\mathrm{te}}} + \lambda \|f\|^2_{\mathcal{H}} \right\}. \tag{2.15}$$

The following lemma describes the solution of the minimization problems (2.14) and (2.15).

Lemma 2.1. *For any $\lambda > 0$, the solutions $f_{\mathbf{z},\lambda}$ and f_λ exist and are unique. Moreover,*

$$f_{\mathbf{z},\lambda}^{\mathrm{IW}} = \left(S_{\mathbf{x}}^\top M_{\mathbf{w}} S_{\mathbf{x}} + \lambda \right)^{-1} S_{\mathbf{x}}^\top M_{\mathbf{w}} \mathbf{y} \tag{2.16}$$

where $\mathbf{y} = (y_1, \dots, y_n)$, $\mathbf{w} = (w(x_1), \dots, w(x_n))$ with $w(x) = d\rho_X^{\mathrm{tr}}/d\rho_X^{\mathrm{te}}(x)$ and $M_{\mathbf{w}}$ being the diagonal matrix with main diagonal entries $w(x_i), i = 1, \dots, n$.

For the infinite data case, the the solution of the minimization problem (2.15) is

$$f_\lambda = (T + \lambda)^{-1} L f_{\mathcal{H}}. \tag{2.17}$$

where $T = T_{\rho_X^{\mathrm{te}}}$ and $L = L_{\rho_X^{\mathrm{te}}}$.

Proof. The proof of (2.16) can be found in (Smale and Zhou, 2004, Theorem. 2). For (2.17) see (Cucker and Smale, 2002b, Proposition. 7). □

Remark 2.2. Assuming that the matrix $M_{\mathbf{w}}$ has full rank the solution (2.16) can be equivalently written as

$$f_{\mathbf{z},\lambda}^{\mathrm{IW}} = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i), \quad \alpha = \left(K_{\mathbf{xx}} + n\lambda M_{1/\mathbf{w}} \right)^{-1} \mathbf{y}, \tag{2.18}$$

where $K_{\mathbf{xx}}$ is the covariance matrix whose entries are given by $K_{ij} = K(x_i, x_j)$ and $M_{\mathbf{w}}$ is the diagonal matrix with main diagonal entries $1/w(x_i), i = 1, \dots, n$. Depending on the observation weight we rescale the regularizer accordingly: the higher the weight of an observation, the less we regularize.

2.3 Convergence Results

Let $f_{\mathcal{H}}$ be the projection of the regression function f_ρ onto the closure of $\mathcal{H}$ in $L^2(X, \rho_X^{\text{te}})$:

$$\|f_\rho - f_{\mathcal{H}}\|_{\rho_X^{\text{te}}} = \operatorname{dist}(f_\rho, \mathcal{H}) := \inf_{f \in \mathcal{H}} \|f_\rho - f\|_{\rho_X^{\text{te}}}.$$

Clearly non of the learning procedures over $\mathcal{H}$ can achieve better performance than $f_{\mathcal{H}}$. The goal of this section is to understand: (i) how well the weighted empirical risk minimizer (ERM) $f_{\mathbf{z},\lambda}$ approximates $f_{\mathcal{H}}$, (ii) what is the role of the importance weights in this approximation, (iii) how the decay of the regularization parameter λ affects the convergence rates.

There are various ways to measure the approximation error of f_ρ with respect to $f_{\mathbf{z},\lambda}$. First of all, let us notice that measuring the approximation quality using $L^2(X, \rho_X^{\text{tr}})$ norm does not lead to anything interesting as our goal is not to minimize the risk with respect to ρ^{tr}. In this chapter, we shall measure the performance of the model with respect to the $L^2(X, \rho_X^{\text{te}})$ norm.

2.3.1 Assumptions

We first introduce some basic assumptions and we then present the convergence results for importance weighted algorithms.

Assumption 1. There exist $r \geq 1/2$ and $R > 0$ such that $\|L^{-r} f_{\mathcal{H}}\|_{\rho_X^{\text{te}}} \leq R$.

Remark 2.3. The condition above can be equivalently stated as follows. There exists $r \geq 1/2$, $R > 0$ and $g \in L^2\left(X, \rho_X^{\text{te}}\right)$ such that $\|g\|_{\rho_X^{\text{te}}} \leq R$ and $f_{\mathcal{H}}(x) = \left(L^r g\right)(x)$, where L^r is defined by

$$L^r \left(\sum a_i e_i\right) = \sum \mu_i^r a_i e_i.$$

Assuming the finiteness of $\|L^{-r} f_{\mathcal{H}}\|_{\rho_X^{\text{te}}}$ is a common *source condition* in the inverse problem literature (Caponnetto, 2006; De Vito et al., 2005a; Smale and Zhou, 2004, 2007) and it characterizes the regularity of the target function $f_{\mathcal{H}}$. A bigger r corresponds to higher regularity and it can lead to faster convergence rates. In particular, the case $r = 0$ is equivalent to making no assumption, while when $r = 1/2$, we are requiring $f_{\mathcal{H}} \in \mathcal{H}$, since $\|L^{1/2} f\|_{\mathcal{H}} = \|f\|_{\rho_X^{\text{te}}}$. For $r \geq 1/2$ the image of the integral operator $L^r\left(L^2(X, \rho_X^{\text{te}})\right)$ becomes a subset of $\mathcal{H}$, which implies that the minimization of risk functional (2.4) over $\mathcal{H}$ has at least one solution in $\mathcal{H}$. This is referred to as the attainable case.

Assumption 2. Let $w = d\rho_X^{\mathrm{te}}/d\rho_X^{\mathrm{tr}}$. For some $q \in [0,1]$ there exist positive constants W and σ depending on q such that for all integer $m \geq 2$

$$\left(\int_X w(x)^{\frac{m-1}{q}} d\rho_X^{\mathrm{te}}(x)\right)^q \leq \frac{1}{2} m! W^{m-2} \sigma^2. \tag{2.19}$$

Remark 2.4. When $q = 0$ it corresponds to the case when $w(x)$ is uniformly bounded over X. In this case $W = \sup_{x \in X} w(x)$ and $\sigma^2 = \int w(x) d\rho_X^{\mathrm{te}} = \int w^2(x) d\rho_X^{\mathrm{tr}}$.

Obviously, if the training measure does not properly cover the support of the testing one, learning is impossible. It is not hard to check that the Assumption 2 is satisfied when $2\rho_X^{\mathrm{tr}}\{x : w(x) \geq t\} \leq W\sigma^2 \exp(-\frac{t^{1/q}}{W})$, restricting the behavior of large values of the Radon–Nikodym derivative.

The condition (2.19) can be written equivalently as a condition on the Rényi Divergence (Cortes et al., 2010; Mansour et al., 2009a) as follows

$$H_{(m-1)/q}(\rho_X^{\mathrm{te}} \| \rho_X^{\mathrm{tr}}) \leq \frac{1}{m-1}\left(\log m! + \log\left(\frac{W^{m-2}\sigma^2}{2}\right)\right)$$

where

$$H_\alpha(\rho_X^{\mathrm{te}} \| \rho_X^{\mathrm{tr}}) = \frac{1}{\alpha} \log \int_X w(x)^\alpha d\rho_X^{\mathrm{te}}(x)$$

is the Rényi Divergence with parameter α. Notice that for each fixed $q > 0$, we are imposing the growth condition on the Rényi Divergence w.r.t. the parameter m.

Assumption 3. For some $s \in (0,1]$ we assume that

$$E_s := 1 \vee \sup_{\lambda \in (0,1]} \sqrt{\mathcal{N}(\lambda)\lambda^s} < \infty \tag{2.20}$$

where $\mathcal{N}(\lambda) = \mathrm{Tr}\left[T(T+\lambda)^{-1}\right]$.

The constant E_s characterizes the marginal testing distribution ρ_X^{te} through $\mathcal{N}(\lambda)$, also termed as *degrees of freedom* (Zhang, 2005) or *effective dimension* (Caponnetto and De Vito, 2007). The boundedness of E_s was implicitly assumed in (Caponnetto and De Vito, 2007, Definition 1, (iii)) and it is satisfied, for instance, when the eigenvalues of T, $\mu_i(T)$, have an asymptotic order $\mathcal{O}\left(i^{-1/s}\right)$. In general, the eigenvalue assumption is a tighter measure for the complexity of the RKHS than more classical covering or entropy number assumptions (Steinwart et al., 2009). For the case $s = 1$, referred to as the *capacity independent* setting, E_1 is always bounded as $\mathcal{N}(\lambda)\lambda^s = \mathcal{N}(\lambda)\lambda \leq \kappa = 1$.

2.3.2 Rates of Convergence for IW-KRR

Now we are ready to state our main results for the importance weighted kernel ridge regression.

Theorem 2.5. *Let ρ^{te} and ρ^{tr} be the distributions on $X \times [-M, M]$, where $M > 0$ is some constant, satisfying Assumptions 1-3. Let $q \in [0,1]$ and $s \in (0,1]$. Furthermore, let n and λ satisfy the constraints $\lambda \leq \|T\|_{\text{op}}$ and*

$$\lambda = \left(\frac{8E_s^{1-q}(\sqrt{W}+\sigma)\log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{\frac{2}{2r+s+q(1-s)}} \tag{2.21}$$

for $\delta \in (0,1)$ and $r \geq 0.5$. Then, with probability greater than $1-\delta$, it holds

$$\|f_{\mathbf{z},\lambda}^{\text{IW}} - f_{\mathcal{H}}\|_{\rho_X^{\text{te}}} \leq 3\,(M+R) \left(\frac{8E_s^{1-q}(\sqrt{W}+\sigma)\log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{\frac{2r}{2r+s+q(1-s)}}. \tag{2.22}$$

The above theorem provides a convergence result in high probability for the importance weighted kernel ridge regression for the attainable case.

We first notice that the optimal choice of the regularization parameter (Caponnetto, 2006; Caponnetto and De Vito, 2007) depends on the characteristics of importance weights. Compared to the standard learning scenario, the optimal regularizer should be bigger under covariate shift and importance weighting correction is applied, and it depends on W and σ.

Second, we notice that equation (2.21) together with the condition $\lambda \leq \|T\|_{\text{op}}$ can be equivalently given as a condition on the number of observations as follows

$$\sqrt{n} \geq 8E_s^{1-q}(\sqrt{W}+\sigma)\|T\|_{\text{op}}^{-r-\frac{s+q(1-s)}{2}} \log\left(\frac{6}{\delta}\right).$$

Third, the slow tail decay of the importance weighting function does not go in favor of importance weighting adaptation. To see this let us consider two extreme cases when $q = 0$ and $q = 1$. When $q = 0$, w is bounded and $\|f_{\mathbf{z},\lambda}^{\text{IW}} - f_{\mathcal{H}}\|_{\rho_X^{\text{te}}}$ has an asymptotic rate of convergence $\mathcal{O}\left(n^{-\frac{r}{2r+s}}\right)$ for $s \in (0,1]$, which is optimal in the minimax sense of Caponnetto and De Vito (2007). The rate of convergence of importance weighted KRR achieves the same order as those attained by the KRR without covariate shift. For $q = 1$, the probability of observing large values of w decays exponentially, and the best learning rate that can be achieved is $\mathcal{O}\left(n^{-\frac{r}{2r+1}}\right)$, for all $s \in (0,1]$.

This last remark agrees with the earlier observation of Cortes et al. (2010) that importance weighting correction can succeed when the weights are bounded and it gives slower rates under weak assumption on the moment of the weight. As pointed out by Kpotufe and Martinet (2021), the slow rates are not only the consequence of importance weighting correction. In a minimax sense, such situations are hard irrespective of the learning approach. Later we will see that the optimal rates can be achieved under a much weaker assumption on the weights than the one put forward in Assumption 2.

A corollary of Theorem 2.5, encompassing all the values of $q \in [0,1]$, can be obtained for the special case corresponding to $s = 1$.

Corollary 2.6. *Let ρ^{te} and ρ^{tr} be as in Theorem 2.5. If λ satisfies the constraints $\lambda \leq \|T\|_{\mathrm{op}}$ and*

$$\lambda = \left(\frac{8(\sqrt{W} + \sigma) \log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{\frac{2}{2r+1}}$$

for $\delta \in (0,1)$ and $r \geq 0.5$. Then, with probability greater than $1-\delta$, it holds

$$\|f_{\mathbf{z},\lambda}^{\mathrm{IW}} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \leq 3\,(M+R) \left(\frac{8(\sqrt{W} + \sigma) \log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{\frac{2r}{2r+1}}.$$

The above bound is kernel independent (except the quantity R) and it matches the one given in Smale and Zhou (2007). The corollary states that using importance weighting adaptation when choosing the right regularization parameter λ gives us the same rate of convergence as those attained by the KRR in the absence of covariate shift.

Finite dimensional RKHS. We now turn to deriving some explicit consequences of our main theorems for specific classes of reproducing kernel Hilbert spaces. Our first corollary applies to problems for which the kernel has finite rank N, meaning that its eigenvalues satisfy $\mu_j(T) = 0$ for all $j > N$. Examples of such finite rank kernels include the linear kernel $K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}$, and the kernel $K(x,x') = (1 + \langle x, x' \rangle_{\mathbb{R}^d})^k$ generating polynomials of degree k.

Corollary 2.7. *Let ρ^{te} and ρ^{tr} be the distributions on $X \times [-M, M]$, satisfying Assumption 2 with $q = 1$ and let $\mathcal{N}(\lambda) \leq N$. If λ satisfies the constraints $\lambda \leq \|T\|_{\mathrm{op}}$ and*

$$\lambda = \left(\frac{8\sqrt{N}(\sqrt{W} + \sigma) \log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{2}, \quad \delta \in (0,1),$$

then, with probability greater than $1-\delta$, it holds

$$\|f_{\mathbf{z},\lambda}^{\mathrm{IW}} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \leq 3\,(M+R)\, \frac{8\sqrt{N}(\sqrt{W} + \sigma) \log\left(\frac{6}{\delta}\right)}{\sqrt{n}}.$$

The rate $\mathcal{O}\left(\sqrt{N/n}\right)$ is known to be optimal for the ridge regression without covariate shift.

Smoothness spaces. The Assumption 3 holds for the covariance operator T with the eigenvalues of the following order

$$\mu_i(T) \sim \mathcal{O}\left(i^{-\frac{1}{s}}\right). \tag{2.23}$$

If X is a Euclidean ball in $\mathbb{R}^d$, $\beta > d/2$ is some integer, and ρ_X^{te} is the uniform distribution on X, then the Sobolev space $\mathcal{H} := W^\beta(X)$ is an RKHS that satisfies (2.23) for $s := \frac{d}{2\beta}$.

For this particular choice of the eigenvalue decay we have the following

Corollary 2.8. *Let ρ^{te} and ρ^{tr} be as in Theorem 2.5. If λ satisfies the constraints $\lambda \leq \|T\|_{\text{op}}$ and*

$$\lambda = \left(\frac{8(\sqrt{W}+\sigma)\log\left(\frac{6}{\delta}\right)}{\sqrt{n}}\right)^{\frac{4\beta}{2\beta(2r+q)+d(1-q)}}$$

for $\delta \in (0,1)$ and $r \geq 0.5$. Then, with probability greater than $1-\delta$, it holds

$$\|f_{\mathbf{z},\lambda}^{\text{IW}} - f_{\mathcal{H}}\|_{\rho_X^{\text{te}}} \leq 3\,(M+R)\left(\frac{8(\sqrt{W}+\sigma)\log\left(\frac{6}{\delta}\right)}{\sqrt{n}}\right)^{\frac{4r\beta}{2\beta(2r+q)+d(1-q)}}. \tag{2.24}$$

For the special case when $q = 0$ and $r = 1/2$, we have the optimal rates for Sobolev spaces $\mathcal{O}((W/n)^{-\frac{\beta}{2\beta+d}})$, under covariate shift with bounded importance weights (Ma et al., 2022). Note that the rates reduce to the known optimal rates (Caponnetto and De Vito, 2007) in the case of no covariate shift.

The Sobolev space is not the only example when the condition (2.23) is satisfied. Another example is the RKHS with the Gaussian radial basis function (RBF) reproducing kernel and the distribution ρ_X^{te} satisfying the following condition

$$\rho_X^{\text{te}}\left(\mathbb{R}^d \backslash r_1 B\right) \leq r_1^{-\tau}, \quad r_1 > 0,$$

where B is the unit ball in $\mathbb{R}^d$. A bound of the form (2.23) can be established (Steinwart and Christmann, 2008, Theorem 7.34).

2.4 Effect of Using Incorrect Weights

In the Section 2.3 we have analyzed the generalization properties of the importance weighted KRR and concluded that importance weighting adaptation is an effective strategy whenever the regularizer is properly tuned (Theorem 2.5). It is important to notice these results have been derived under the assumption that the importance weights $w(x)$ can be perfectly estimated, see (2.6). In the following we relax this assumption and we study the performance of the weighted KRR in the case of a weighing function $v(x)$ that does not match the ratio between test and train marginal distributions.

In the case of "imperfect" weights $v(x)$, the optimization problem (2.14) becomes

$$f'_{\mathbf{z},\lambda} := \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} v(x_i)\left(f\left(x_i\right)-y_i\right)^2 + \lambda\|f\|_{\mathcal{H}}^2 \right\} \tag{2.25}$$

where now $v(x) = \frac{d\rho'_X(x)}{d\rho^{\mathrm{tr}}_X(x)}$ for some measure $\rho'_X \ll \rho^{\mathrm{tr}}_X$. In the data free scenario, the optimization problem (2.25) is equivalent to

$$f'_\lambda := \arg\min_{f\in\mathcal{H}} \left\{ \|f - f_\rho\|^2_{\rho'_X} + \lambda\|f\|^2_{\mathcal{H}} \right\},$$

whose solution can be expressed as

$$f'_\lambda = \left(T' + \lambda I\right)^{-1} L' f'_{\mathcal{H}}, \tag{2.26}$$

where $T' = T_{\rho'_X}$, $L' = L_{\rho'_X}$ and $f'_{\mathcal{H}}$ being the projection of the regression function f_ρ onto the closure of $\mathcal{H}$ in $L^2(X, \rho'_X)$.

In order to provide guarantees for the finite data scenario we need a set of assumptions on v and ρ'_X that are similar to Assumptions 2 and 3.

Assumption 4. For some $q' \in [0,1]$ there exist positive constants V and γ depending on q such that for all $m \geq 2$

$$\left(\int_X v(x)^{\frac{m-1}{q'}}\, d\rho'_X(x)\right)^{q'} \leq \frac{1}{2} m! V^{m-2}\gamma^2. \tag{2.27}$$

Assumption 5. For some $s' \in (0,1]$ we assume that

$$E'_{s'} := 1 \vee \sup_{\lambda\in(0,1]} \sqrt{\mathcal{N}'(\lambda)\lambda^{s'}} < \infty \tag{2.28}$$

where $\mathcal{N}'(\lambda) = \mathrm{Tr}\left[T'(T'+\lambda)^{-1}\right]$.

Equipped with the necessary assumptions, we are now ready to state the main result for an arbitrarily weighted KRR.

Theorem 2.9. *Let ρ^{te} and ρ' be the distributions on $X \times [-M, M]$, where $M > 0$ is some constant, satisfying Assumptions 1,4 and 5. Let $q' \in [0,1]$ and $s' \in (0,1]$. Furthermore,* *let n and λ satisfy the constraints $\lambda \leq \|T'\|_{\mathrm{op}}$ and*

$$\lambda = \left(\frac{8E'_{s'}(\sqrt{V} + \gamma) \log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{\frac{2}{2r+s'+q'(1-s')}} \tag{2.29}$$

for $\delta \in (0,1)$. Then, for $r \geq 0.5$, with probability greater than $1-\delta$, it holds

$$\|f'_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \leq 3\frac{(M+R)}{1-\mu} \left(\frac{8E'_{s'}(\sqrt{V} + \gamma) \log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{\frac{2r}{2r+s'+q'(1-s')}} + 4\|f'_{\mathcal{H}} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \tag{2.30}$$

with $\mu = \mu_{\max}\left((T+\lambda I)^{-1/2}(T-T')(T+\lambda I)^{-1/2}\right)$.

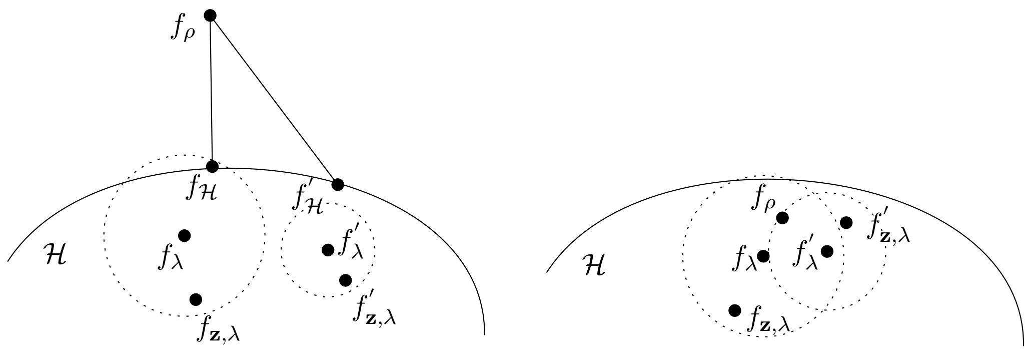
Remark 2.10. The scaling factor $1/(1-\mu)$ is well defined as $\mu < 1$. For the special case when $\rho'_X = \rho_X^{\mathrm{te}}$, both μ and $\|f'_{\mathcal{H}} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}}$ vanish, recovering the learning rates of Theorem 2.5.

Theorem 2.9 highlights some important aspects associated to "imperfectly" re-weighting the risk and its effects on the learning rates.

First we notice that a good choice of the weighting function heavily depends on the approximation properties of $\mathcal{H}$. Consider the misspecified scenario depicted in Figure 2.1a in which $f_\rho \notin \mathcal{H}$. For a correctly specified weighting function, the solution $f_{\mathbf{z},\lambda}$ concentrates around its data free limit f_λ, and the latter is a good approximation of $f_{\mathcal{H}}$, the projection of regression function f_ρ on $\mathcal{H}$ under the measure induced by ρ^{te}. In the case of "imperfect" weights, $f'_{\mathbf{z},\lambda}$ concentrates around f'_λ that approximates $f'_{\mathcal{H}}$, the projection of f_ρ under the measure induced by ρ'. The latter minimizes the projection error under the induced measure ρ', not the testing one ρ^{te}. Therefore, whenever the model class is misspecified and the weights are "imperfect", the learner approximates a "wrong" projection.

The situation is less dramatic when the model class is well-specified, $f_\rho \in \mathcal{H}$, or when f_ρ is well approximated by the elements of $\mathcal{H}$ (Figure 2.1b). Under a suitable choice of the regularization parameter, both f_λ and f'_λ are close to the regression function f_ρ. Therefore, becomes preferable to choose weighting functions with a small variance and not necessarily matching the ratio between train and test measures. This explains an uniform weighting function, which has zero variance, is often the best choice for universal kernels.

Another important quantity to be considered is the scaling factor $1/(1-\mu)$, which measures the distortion between candidate and testing measures. However, choosing the testing measure for ρ'_X that give rise the smallest scaling constant does not necessarily improve the generalization bound due to the large constant V.



(A) Misspecified

(B) well-specified

FIGURE 2.1: Difference between misspecified and well-specified scenarios. (a) IW adaptation could significantly reduce the approximation error with a price of high variance. (b) On the contrary, the performance of the model can be significantly improved by using the weighting function with better control of large values.

Remarkably, the bound and the optimal regularization parameter depend on the geometry of candidate measure ρ'_X. Indeed, in the well-specified scenario, the testing measure plays a role only through the scaling factor, defined by μ.

Below we consider some weighting procedures which are commonly used in practice.

Uniform weights. The uniform weights scenario corresponds to the case where $\rho'_X = \rho^{\text{tr}}_X$. This choice corresponds to the solution of the following optimization problem

$$\rho^{\text{tr}}_X = \underset{\rho' : \rho'_X \ll \rho^{\text{tr}}_X}{\arg\min} \left\| \frac{d\rho'_X}{d\rho^{\text{tr}}_X} \right\|_\infty ,$$

and it yields the smallest constant V in the generalization bound (2.30). In other words, the uniform weights leads to the smallest generalization bound whenever the difference between the projections of the regression function f_ρ on $\mathcal{H}$ w.r.t the training and testing input measures is small. This is the case in the well-specified scenario, when $f_\rho \in \mathcal{H}$, or when the kernel K is universal, meaning that the corresponding Hilbert space $\mathcal{H}$ is dense in $L^2(X, \rho^{\text{te}}_X)$. In both of these cases $f_\rho = f_\mathcal{H} = f'_\mathcal{H}$, and the term $\|f'_\mathcal{H} - f_\mathcal{H}\|_{\rho^{\text{te}}_X}$ in the bound (2.30) vanishes. The above discussion is formalized in the following corollary.

Corollary 2.11. *Let us assume that the conditions of Theorem 2.9 are satisfied with $\rho'_X = \rho^{\text{tr}}_X$. Furthermore, assume that either $f_\rho \in \mathcal{H}$, or $\mathcal{H}$ is dense in $L^2(X, \rho^{\text{te}}_X)$. Let n and λ satisfy the constraints $\lambda \leq \|T'\|_{\text{op}}$ and*

$$\lambda = \left(\frac{8E'_{s'} \log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{\frac{2}{2r+s'}}$$

for $\delta \in (0,1)$ *and* $s' > 0$. *Then, for* $r \geq 0.5$, *with probability greater than* $1-\delta$, *it holds*

$$\|f'_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \leq 3\frac{(M+R)}{1-\mu}\left(\frac{8E'_{s'}\log\left(\frac{6}{\delta}\right)}{\sqrt{n}}\right)^{\frac{2r}{2r+s'}}. \tag{2.31}$$

The generalization bound (2.31) depends only on the geometry of the training input measure. To achieve this bounds no assumption on the IW is needed. For the covariate shift with bounded importance wights, Ma et al. (2022) showed that the rates for unweighted KRR can be rewritten in terms of the testing measure geometry with the rates similar to those attained by IW-KRR (2.24) with $r = 1/2$ and $q = 0$.

Clipping of importance weights. Another popular weighting function is the one obtained after clipping the importance weights that exceed a maximum value D. Namely,

$$w_D(x) = \min\{w(x), D\}. \tag{2.32}$$

We denote the solution of the weighted KRR based on clipped importance weights by $f^D_{\mathbf{z},\lambda}$ and provide the following learning guarantee.

Theorem 2.12. *Assume that* $\int w(x)d\rho_X^{\mathrm{te}}(x) \leq \Sigma$ *and the eigenfunctions* $\{e_i\}_{i\geq 1}$ *of the covariance operator* T *is uniformly bounded*

$$\sup_{i\geq 1}\|e_i\|_\infty \leq 1. \tag{2.33}$$

Furthermore, let n *and* λ *satisfy the constraints*

$$\lambda = \left(\frac{16\Sigma E_s \log\left(\frac{6}{\delta}\right)}{\sqrt{n}}\right)^{\frac{2}{2r+s}} \quad \text{and} \quad D = 2\Sigma\mathcal{N}(\lambda), \tag{2.34}$$

for $\delta \in (0,1)$. *Then, for* $r \geq 0.5$, *with probability greater than* $1-\delta$, *it holds*

$$\|f^D_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \leq 2\,(M+R)\left(\frac{16\Sigma E_s\log\left(\frac{6}{\delta}\right)}{\sqrt{n}}\right)^{\frac{2r}{2r+s}}. \tag{2.35}$$

Remark 2.13. To achieve the optimal rates (2.35) the model class does not need to be well specified as the bias term is eliminated by the truncation of the weights to a maximum value D.

The detailed proof can be found in Appendix 2.8.3. A key step in the proof consists in showing that $\|T(T_D+\lambda)^{-1}\|$ is uniformly bounded over λ, where $T_D = \int T_x w_D(x)d\rho_X^{\mathrm{tr}}(x)$. Having establish that, the rest of the proof follows along the same lines of the one for Theorem 2.5.

Unfortunately, the condition (2.33) is not always satisfied and it is known that even C^∞-kernels on $[0,1]$, where $[0,1]$ is equipped with the Lebesgue measure, may not have uniformly bounded orthonormal basis (Zhou, 2002, Example 1). In general, the uniform boundedness property is hard to check. Even for the Gaussian RBF kernel on $[-1,1]$ it is unknown whether the condition (2.33) holds. Uniformly bounded eigenfunctions have been considered, e.g., by Mendelson and Neeman (2010) and Steinwart et al. (2009).

The condition (2.33) can be relaxed imposing a stronger moment condition on the weighting function. Indeed, hard covariate shifts, where the hardness is encoded in the moment condition, should be met by extra integrability condition for the eigenvalues e_i. This is more precisely stated in the following theorem.

Theorem 2.14. *Assume that the following conditions hold*

$$\|w\|_{k,\rho_X^{\mathrm{te}}} \le \Sigma \quad \textit{and} \quad \sup_{i\ge 1} \|e_i^2\|_{l,\rho_X^{\mathrm{te}}} \le 1,$$

where $1/k + 1/l = 1$. *Let* n, λ *and* D *satisfy the condition* (2.34). *Then*

$$\|f_{\mathbf{z},\lambda}^D - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \le 2\,(M+R) \left(\frac{16 E_s \Sigma \log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{\frac{2r}{2r+s}},$$

with probability greater than $1-\delta$.

Kpotufe Importance Weights. Another weighting function, motivated by the work of Kpotufe (2017), is

$$w_D(x) = \frac{\rho_X^{\mathrm{te}}(B_D(x))}{\rho_X^{\mathrm{tr}}(B_D(x))},$$

where $B_D(x)$ is the ball of radius D centered at the point x. Obviously, if $\rho_X^{\mathrm{te}} \ll \rho_X^{\mathrm{tr}}$, $w_D(x)$ approaches the importance weights as $D \to 0$; however, $w_D(x)$ can be defined even for the measures when IW is not well defined (Kpotufe and Martinet, 2021). For $D = \mathrm{diam}(X)$, we are in the case of the uniform weight $w_D(x) = 1$.

2.5 Binary Classification

In the following we show that the above results above can be applied to binary classification algorithms, i.e., when $Y = \{-1,1\}$. The problem of statistical learning in classification consists of predicting the value $y \in \{-1,1\}$ for a given $x \in X$. As in the regression setting here we also distinguish training and testing input distributions, while the conditional distribution is the same and supported on $\{-1,1\}$. We consider *binary* *classifiers*, namely functions $f : X \to \{-1, 1\}$ that assign a label to each point $x \in X$. We denote the *classification error* of a classifier as follows

$$\mathcal{R}(f) = \rho \{(x, y) \in Z = X \times \{-1, 1\} \mid f(x) \neq y\}.$$

It is well known that

$$\min_f \mathcal{R}(f) = \mathcal{R}(f_\rho)$$

where $f_\rho(x) = \int_{\mathbb{R}} y d\rho(y \,|\, x) = P(y = 1 \,|\, x) - P(y = -1 \,|\, x)$ is a regression function. The classifier $\operatorname{sgn}(f_\rho(x))$ is called the *Bayes rule.*

It can be shown (Bartlett et al., 2006; Bauer et al., 2007) that the excess misclassification error $\mathcal{R}(f) - \mathcal{R}(f_\rho)$ can be upper bounded as follows

$$\mathcal{R}(f) - \mathcal{R}(f_\rho) \leq \|f - f_\rho\|_{\rho_X^{\mathrm{te}}},$$

meaning that the Theorem 2.5 can be directly applied to achieve finite sample guarantees replacing f by $f_{\mathbf{z},\lambda}$. Mammen and Tsybakov (1999) first showed that one can attain fast rates under mild assumptions on the behavior of $f_\rho(x)$ in a neighborhood of the boundary $\{x : f_\rho(x) = 0\}$. Namely if the Tsybakov noise condition holds (Tsybakov, 2004) for $l \geq 0$,

$$\rho_X^{\mathrm{te}} \left(\{x \in X : f_\rho(x) \in [-\Delta, \Delta]\}\right) \leqslant B_l \Delta^l, \quad \forall \Delta \in [0, 1], \tag{2.36}$$

then

$$\mathcal{R}\left(f_{\mathbf{z},\lambda}\right) - \mathcal{R}\left(f_\rho\right) \leqslant 4c_\alpha \left\|f_z - f_\rho\right\|_{\rho_X^{\mathrm{te}}}^{\frac{2}{2-\alpha}},$$

with $\alpha = l/(l + 1)$ and $c_l = B_l + 1$(Bauer et al. (2007),Yao et al. (2007)). A direct application of Theorem 2.5 gives us

Corollary 2.15. *Assume that the assumptions of Theorem 2.5 is satisfied together with the margin condition (2.36) and let $f_\rho \in \mathcal{H}$. If λ satisfies the constraints $\lambda \leq \|T\|$ and*

$$\lambda = \left(\frac{8E_s^{1-q}(\sqrt{W} + \sigma) \log\left(\frac{6}{\delta}\right)}{\sqrt{n}}\right)^{\frac{2}{2r+s+q(1-s)}},$$

for $\delta \in (0, 1)$, $q \in [0, 1]$ and $r \geq 0.5$. Then with probability greater than $1 - \delta$, it holds

$$\mathcal{R}\left(f_{\mathbf{z},\lambda}\right) - \mathcal{R}\left(f_\rho\right) \leq 12c_\alpha \left(M + R\right) \left(\frac{8E_s^{1-q}(\sqrt{W} + \sigma) \log\left(\frac{6}{\delta}\right)}{\sqrt{n}}\right)^{\frac{4r}{(2r+s+q(1-s))(2-\alpha)}}.$$

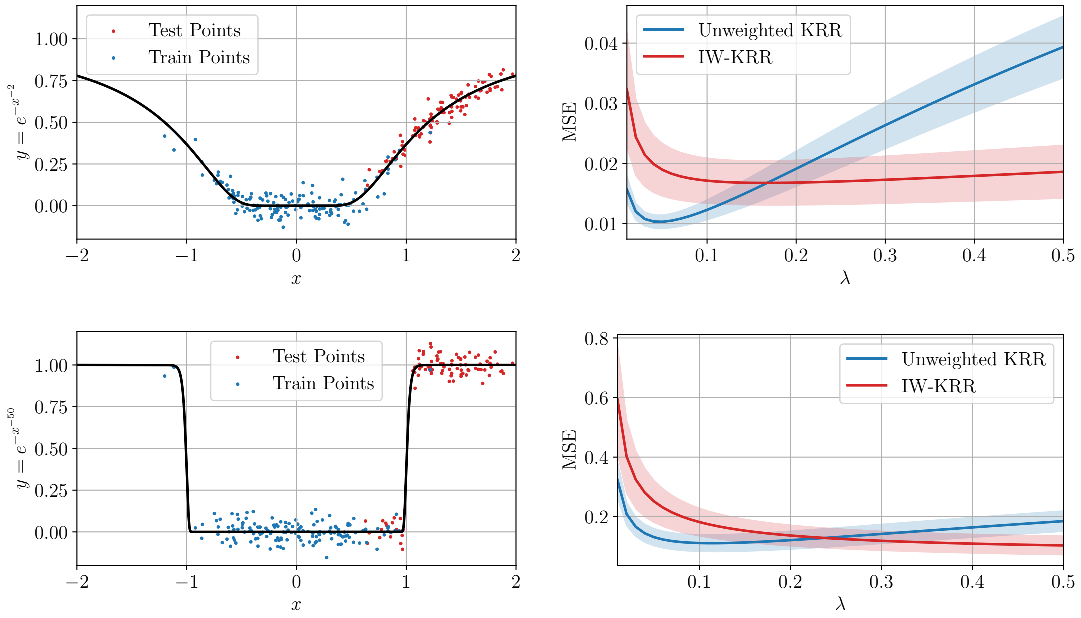


FIGURE 2.2: Comparison between unweighted KRR and IW-KRR for different regression functions. In the top left figure we consider a smooth regression from which training and testing inputs are distributed as $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(1.5, 0.3)$ respectively. On the top right panel we report the performance of the IW-KRR and unweighted KRR for different regularization parameter values. On the bottom row we repeat the experiment using relatively non-smooth regression function.

2.6 Simulations

We consider a simple one-dimensional regression problem with a regression function

$$f_\rho(x) = e^{-\frac{1}{x^{2k}}}, \quad k \in \mathbb{N}, \tag{2.37}$$

corrupted by homoscedastic Gaussian noise with mean $\mu = 0$ and standard deviation $\sigma = 0.05$. We assume that $x \sim \mathcal{N}(0, 0.5)$ at training time and $x \sim \mathcal{N}(1.5, 0.3)$ at testing time.

In the first experiment we compare the performance of importance weighted and unweighted KRR for two different values of k. We choose a Gaussian RBF kernel with length-scale parameter equal to one. On the top left panel of Figure 2.2 the regression function with $k = 1$ along with a randomly drawn set of points from training and testing measures are reported. On the top right panel the performance of the weighted and unweighted KRR for different values of the regularization parameter is presented. As one can see, importance weighting adaptation with optimal regularization parameter performs slightly worse than unweighted KRR with optimal λ. This can be explained by the fact that the underling regression function is sufficiently smooth and can be well approximated by the functions in RKHS with exponential kernel.

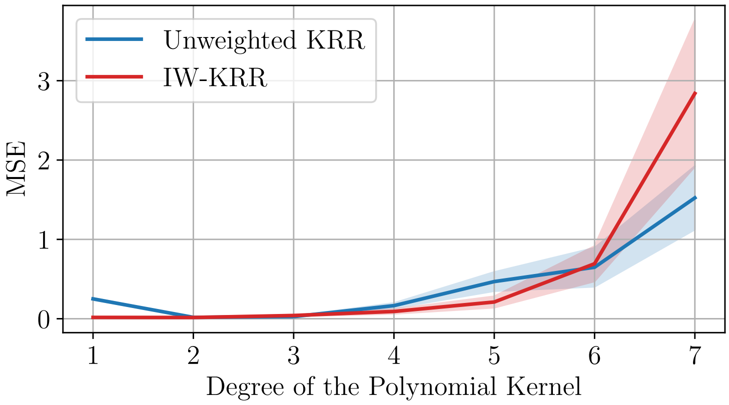


FIGURE 2.3: Performance of IW-KRR and unweighted KRR using a polynomial kernel with increasing degree. As the degree of the polynomial kernel increases, the role of the importance weights diminishes.

On the other hand when k in (2.37) is large, the regression function essentially becomes piece-wise constant. As it is well known, neither constants nor discontinuous functions belong to the RKHS associated with the Gaussian RBF kernel, so the increased values of k increases the level of misspecification. The regression function with $k = 25$ together with the randomly drawn training and testing points is reported in the bottom-left panel of Figure 2.2. Unlike the previous example, IW adaptation here can be beneficial as shown in the bottom-right panel of the figure. Notice that in both these examples, the optimal regularization parameter for unweighted KRR is smaller in comparison with the optimal λ of IW-KRR. This is well justified by the optimal choice of the regularization parameter in Theorem 2.9.

The relation between the approximation properties of the RKHS associated to a kernel and the performace of weighted KRR can be showcase using a polynomial kernel with an increasing kernel degree. In Figure 2.3 the performance of KRR with polynomial kernel is given for the function (2.37) with $k = 1$. For the degree-one kernel, the advantage of IW adaptation is apparent; the regression function can be well approximated by the linear function under the testing distribution. For the degree-two polynomial, the space of quadratic functions can approximate the true regression function in the supremum norm in a suitable chosen domain, and therefore the model trained on the training data with uniform weights gives a globally suitable model. With the degree of polynomial kernel increasing, IW adaptation does not provide a clear benefit.

2.7 Discussion

In this chapter, we studied the generalization properties of weighted KRR under covariate shift. For the bounded importance weights, we proved the minimax optimality of IW-KRR. We showed that slightly relaxing the boundedness condition leads to slower rates for IW-KRR. We believe that questions related to optimality are of potential interest, and we leave this to future investigations. By examining alternative re-weighting procedures we highlighted several important factors to be considered for good generalization properties. We demonstrated the importance of distinguishing well-specified and misspecified scenarios under covariate shift. Under the extra regularity conditions on the eigenfunctions, the optimality of clipped IW-KRR was shown for the family of covariate shifts with integrable (w.r.t ρ^{te}) importance weights.

2.8 Proofs

2.8.1 Upper bound for IW-KRR

In order to establish upper bounds, we split $\|f^{\mathrm{IW}}_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}}$ into two parts:

$$\|f^{\mathrm{IW}}_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \le \|f^{\mathrm{IW}}_{\mathbf{z},\lambda} - f_{\lambda}\|_{\rho_X^{\mathrm{te}}} + \|f_{\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}}, \tag{2.38}$$

the *estimation error* $\|f_{\mathbf{z},\lambda} - f_{\lambda}\|_{\rho_X^{\mathrm{te}}}$ and the *approximation error* $\|f_{\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}}$. A bound on the approximation error has already been given in Proposition 2.21.

Theorem 2.16. *Assume $\lambda \le \|T\|$ and*

$$n\lambda^{1+q} \ge 16(W+\sigma^2)\mathcal{N}(\lambda)^{1-q}\log^2\left(\frac{6}{\delta}\right) \tag{2.39}$$

for some $\delta \in (0,1)$. Then, with probability greater than $1-\delta$, it holds

$$\|f^{\mathrm{IW}}_{\mathbf{z},\lambda} - f_{\lambda}\|_{\rho_X^{\mathrm{te}}} \le 16\,(M+R)\left(\frac{W}{n\sqrt{\lambda}} + \sigma\sqrt{\frac{\mathcal{N}(\lambda)^{1-q}}{n\lambda^q}}\right)\log\left(\frac{6}{\delta}\right) \tag{2.40}$$

Proof. To bound the estimation error we first introduce more compact notations:

$$T_{\mathbf{x},\mathbf{w}} = S_{\mathbf{x}}^{\top} M_{\mathbf{w}} S_{\mathbf{x}}, \quad g_{\mathbf{z},\mathbf{w}} = S_{\mathbf{x}}^{\top} M_{\mathbf{w}} \mathbf{y} \quad \text{and} \quad g = Lf_{\mathcal{H}}.$$

The strategy to obtain an upper bound is fairly standard. We decompose analytically the estimation error in different terms, that will be bounded, via Bernstein inequality.

The decomposition relevant to us can be obtained by simple algebraic computations as follows

$$\begin{aligned}
f_{\mathbf{z},\lambda}^{\mathrm{IW}} - f_\lambda &= \left(T_{\mathbf{x},\mathbf{w}} + \lambda\right)^{-1} g_{\mathbf{z},\mathbf{w}} - (T+\lambda)^{-1} g \\
&= \left(T_{\mathbf{x},\mathbf{w}} + \lambda\right)^{-1} \left\{ \left(g_{\mathbf{z},\mathbf{w}} - g\right) + \left(T - T_{\mathbf{x},\mathbf{w}}\right)(T+\lambda)^{-1} g \right\} \\
&= \left(T_{\mathbf{x},\mathbf{w}} + \lambda\right)^{-1} (T+\lambda)^{\frac{1}{2}} \left\{ (T+\lambda)^{-\frac{1}{2}} \left(g_{\mathbf{z},\mathbf{w}} - g\right) + (T+\lambda)^{-\frac{1}{2}} \left(T - T_{\mathbf{x},\mathbf{w}}\right)(T+\lambda)^{-1} g \right\} \\
&= (T+\lambda)^{-\frac{1}{2}} \left\{ I - (T+\lambda)^{-\frac{1}{2}} \left(T - T_{\mathbf{x},\mathbf{w}}\right)(T+\lambda)^{-\frac{1}{2}} \right\}^{-1} \\
&\quad \left\{ (T+\lambda)^{-\frac{1}{2}} \left(g_{\mathbf{z},\mathbf{w}} - g\right) + (T+\lambda)^{-\frac{1}{2}} \left(T - T_{\mathbf{x},\mathbf{w}}\right) f_\lambda \right\}.
\end{aligned} \tag{2.41}$$

Assuming that

$$S_1 := \left\| (T+\lambda)^{-\frac{1}{2}} \left(T - T_{\mathbf{x},\mathbf{w}}\right)(T+\lambda)^{-\frac{1}{2}} \right\|_{\mathrm{HS}} < 1, \tag{2.42}$$

where $\|A\|_{\mathrm{HS}}^2 = \operatorname{Tr}\left(A^\top A\right)$, and using the Neumann series expansion we obtain

$$\begin{aligned}
\left\| \left\{ I - (T+\lambda)^{-\frac{1}{2}} \left(T - T_{\mathbf{x},\mathbf{w}}\right)(T+\lambda)^{-\frac{1}{2}} \right\}^{-1} \right\|_{\mathrm{op}} &= \left\| \sum_{n=0}^{\infty} \left[(T+\lambda)^{-\frac{1}{2}} \left(T - T_{\mathbf{x},\mathbf{w}}\right)(T+\lambda)^{-\frac{1}{2}} \right]^n \right\|_{\mathrm{op}} \\
&\leq \sum_{n=0}^{\infty} \left\| (T+\lambda)^{-\frac{1}{2}} \left(T - T_{\mathbf{x},\mathbf{w}}\right)(T+\lambda)^{-\frac{1}{2}} \right\|_{\mathrm{op}}^n \\
&\leq \sum_{n=0}^{\infty} \left\| (T+\lambda)^{-\frac{1}{2}} \left(T - T_{\mathbf{x},\mathbf{w}}\right)(T+\lambda)^{-\frac{1}{2}} \right\|_{\mathrm{HS}}^n \\
&\leq \frac{1}{1-S_1}.
\end{aligned}$$

From (2.41) by taking the $L^2(X, \rho_X^{\mathrm{te}})$ norm from both sides and using the isometry property we get

$$\left\| f_{\mathbf{z},\lambda}^{\mathrm{IW}} - f_\lambda \right\|_{\rho_X^{\mathrm{te}}} = \left\| T^{\frac{1}{2}} \left(f_{\mathbf{z},\lambda}^{\mathrm{IW}} - f_\lambda \right) \right\|_{\mathcal{H}} \leq \frac{S_2 + S_3}{1 - S_1}, \tag{2.43}$$

where

$$\begin{aligned}
S_2 &:= \left\| (T+\lambda)^{-\frac{1}{2}} \left(g_{\mathbf{z},\mathbf{w}} - g\right) \right\|_{\mathcal{H}}, \\
S_3 &:= \left\| (T+\lambda)^{-\frac{1}{2}} \left(T - T_{\mathbf{x},\mathbf{w}}\right) f_\lambda \right\|_{\mathcal{H}}.
\end{aligned}$$

We have to find an upper bound for each of S_i. To do so, notice that

$$S_i = \left\| \frac{1}{n} \sum_{k=1}^{n} \xi_k - \mathbb{E}[\xi_k] \right\|_F, \quad i = 1, 2, 3,$$

with appropriate choice of the random variable ξ and the norm $\|\cdot\|_F$. Indeed, in order to let the equality above hold, on the space $(X, \rho_X^{\mathrm{tr}})$ we define the operator valued random variable $\xi_1 : X \to \mathrm{HS}\,(\mathcal{H})$, where $\mathrm{HS}\,(\mathcal{H})$ is the space of Hilbert-Schmidt operators on

$\mathcal{H}$, as follows

$$\xi_1(x)[\cdot] = (T+\lambda)^{-\frac{1}{2}} w(x) K_x \langle K_x, \cdot \rangle_{\mathcal{H}} (T+\lambda)^{-\frac{1}{2}}.$$

Moreover, $\xi_2 : Z \to \mathcal{H}$ is defined on the space (Z, ρ^{tr}) by

$$\xi_2(x,y) = (T+\lambda)^{-\frac{1}{2}} w(x) K_x y.$$

Finally, $\xi_3 : X \to \mathcal{H}$ is defined on th space $(X, \rho_X^{\mathrm{tr}})$ by

$$\xi_3(x) = (T+\lambda)^{-\frac{1}{2}} w(x) K_x f_\lambda(x).$$

Application of Proposition 1.10 to each of S_i yields to the following bounds with probability at least $1-\delta/3$

$$S_i \leq \frac{2L_i \log(6/\delta)}{n} + \sigma_i \sqrt{\frac{2\log(6/\delta)}{n}} \tag{2.44}$$

where, as it can be straightforwardly verified, the constants L_i and σ_i are given by the expressions

$$L_1 = \frac{2W}{\lambda}, \quad \sigma_1 = 2\sigma \frac{\sqrt{\mathcal{N}(\lambda)}^p}{\lambda^{1+q}}, \tag{2.45}$$

$$L_2 = \frac{2MW}{\sqrt{\lambda}}, \quad \sigma_2 = 2\sigma M \sqrt{\frac{\mathcal{N}(\lambda)^p}{\lambda^q}}, \tag{2.46}$$

$$L_3 = \frac{2\|f_\lambda\|_{\mathcal{H}} W}{\sqrt{\lambda}}, \quad \sigma_3 = 2\sigma \|f_\lambda\|_{\mathcal{H}} \sqrt{\frac{\mathcal{N}(\lambda)^p}{\lambda^q}}, \tag{2.47}$$

where $p = 1-q$. Let us demonstrate for S_2. First, notice that

$$E\|\xi_2 - E\xi_2\|_{\mathcal{H}}^m \leq E_{\xi_2} E_{\xi_2'} \|\xi_2 - \xi_2'\|_{\mathcal{H}}^m \leq 2^{m-1} E_{\xi_2} E_{\xi_2'} \left(\|\xi_2\|_{\mathcal{H}}^m + \|\xi_2'\|_{\mathcal{H}}^m \right) \leq 2^m E\|\xi_2\|_{\mathcal{H}}^m,$$

where ξ_2' is an independent copy of ξ_2. Second,

$$\begin{aligned}
E\|\xi_2\|_{\mathcal{H}}^m &= E\langle \xi_2, \xi_2\rangle_{\mathcal{H}}^{m/2}\\
&= \int \langle (T+\lambda)^{-\frac{1}{2}} w(x) K_x y, (T+\lambda)^{-\frac{1}{2}} w(x) K_x y\rangle_{\mathcal{H}}^{m/2} d\rho^{\mathrm{tr}}(x,y)\\
&\le M^m \int \|K_x^\top (T+\lambda)^{-1} K_x\|_{\mathrm{op}}^{m/2} w^{m-1}(x) d\rho_X^{\mathrm{te}}(x)\\
&= M^m \int \|K_x^\top (T+\lambda)^{-1} K_x\|_{\mathrm{op}}^{m/2-p} \|K_x^\top (T+\lambda)^{-1} K_x\|^p w^{m-1}(x) d\rho_X^{\mathrm{te}}(x)\\
&\le M^m \left(\frac{1}{\lambda}\right)^{m/2-p} \int \|K_x^\top (T+\lambda)^{-1} K_x\|_{\mathrm{op}}^p w^{m-1}(x) d\rho_X^{\mathrm{te}}(x)\\
\text{Hölder inequality} \quad &\le M^m \left(\frac{1}{\lambda}\right)^{m/2-p} \left(\int \|K_x^\top (T+\lambda)^{-1} K_x\|_{\mathrm{op}} d\rho_X^{\mathrm{te}}(x)\right)^p \left(\int w^{(m-1)/q}(x) d\rho_X^{\mathrm{te}}(x)\right)^q\\
\|A\|_{\mathrm{op}} \le \mathrm{Tr}(A) \quad &\le M^m \left(\frac{1}{\lambda}\right)^{m/2-p} \left(\int \mathrm{Tr}((T+\lambda)^{-1} T_x) d\rho_X^{\mathrm{te}}(x)\right)^p \left(\int w^{(m-1)/q}(x) d\rho_X^{\mathrm{te}}(x)\right)^q\\
&= M^m \left(\frac{1}{\lambda}\right)^{m/2-p} \mathcal{N}(\lambda)^p \left(\int w^{(m-1)/q}(x) d\rho_X^{\mathrm{te}}(x)\right)^q\\
\text{Assumption 2} \quad &\le M^m \left(\sqrt{\frac{1}{\lambda}}\right)^{m-2} \left(\sqrt{\frac{1}{\lambda}}\right)^{2q} \mathcal{N}(\lambda)^p W^{m-2} \sigma^2\\
&\le \frac{1}{2} m! \left(MW\sqrt{\frac{1}{\lambda}}\right)^{m-2} \left(M\sqrt{\frac{1}{\lambda}}^q \sqrt{\mathcal{N}(\lambda)}^p \sigma\right)^2.
\end{aligned}$$

Let us verify (2.42). From the assumption (2.39), with probability greater than $1-\delta/3$, we have

$$\begin{aligned}
S_1 &\le 4\log\left(\frac{6}{\delta}\right)\left(\frac{W}{n\lambda} + \sqrt{\sigma^2 \frac{\mathcal{N}(\lambda)^p}{n\lambda^{1+q}}}\right)\\
&\le 4\frac{W\mathcal{N}(\lambda)^p}{n\lambda^{1+q}} \log^2\left(\frac{6}{\delta}\right) + \sqrt{\sigma^2 \frac{\mathcal{N}(\lambda)^p}{n\lambda^{1+q}} \log^2\left(\frac{6}{\delta}\right)}\\
&\le \frac{3}{4}.
\end{aligned} \tag{2.48}$$

Now, if we combine the estimate (2.43) with the bounds given in (2.44), then we get with probability at least $1-\delta$

$$\|f_{\mathbf{z},\lambda}^{\mathrm{IW}} - f_\lambda\|_{\rho_X^{\mathrm{te}}} \le 16\log\left(\frac{6}{\delta}\right)(M + \|f_\lambda\|_{\mathcal{H}})\left(\frac{W}{n\sqrt{\lambda}} + \sigma\sqrt{\frac{\mathcal{N}(\lambda)^p}{n\lambda^q}}\right).$$

The bound (2.63) finishes the proof. □

Proof of Theorem 2.5. Combining the decomposition in (2.38) with the bounds (2.40) and (2.62) we get

$$\begin{aligned}
\|f^{\mathrm{IW}}_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} &\le 16\log\left(\frac{6}{\delta}\right)(M+R)\left(\frac{W}{n\sqrt{\lambda}} + \sigma\sqrt{\frac{\mathcal{N}(\lambda)^p}{n\lambda^q}}\right) + \lambda^r R \\
&\le 16\log\left(\frac{6}{\delta}\right)(M+R)\left(\frac{W}{n\sqrt{\lambda}} + \frac{\sigma E_s^p}{\sqrt{n\lambda^{sp+q}}}\right) + \lambda^r R.
\end{aligned}$$

By choosing λ as in (2.21) we finally get

$$\begin{aligned}
\|f^{\mathrm{IW}}_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} &\le 16\log\left(\frac{6}{\delta}\right)(M+R)\left(\frac{W}{n\sqrt{\lambda}} + \frac{\sigma E_s^p}{\sqrt{n\lambda^{sp+q}}}\right) + \lambda^r R \\
&\le 2\,(M+R)\,\lambda^r\left(\frac{\sigma}{\sqrt{W}+\sigma} + \frac{W\lambda^{r+sp+q-0.5}}{8\log\left(\frac{6}{\delta}\right)(\sqrt{W}+\sigma)^2}\right) \\
&\le 3\,(M+R)\,\lambda^r, \quad \text{for} \quad s \in (0,1].
\end{aligned}$$

□

2.8.2 Upper bound for arbitrarily weighted KRR

For the proof we need the following proposition from (Rudi and Rosasco, 2017, Proposition 8)

Proposition 2.17. *Let $\mathcal{H}$ be a separable Hilbert space, let A, B two bounded self-adjoint positive linear operators on $\mathcal{H}$ and $\lambda > 0$. Then*

$$\left\|(A+\lambda I)^{-1/2}B^{1/2}\right\|_{\mathrm{op}} \le \left\|(A+\lambda I)^{-1/2}(B+\lambda I)^{1/2}\right\|_{\mathrm{op}} \le (1-\mu)^{-1/2}$$

with

$$\mu = \mu_{\max}\left[(B+\lambda I)^{-1/2}(B-A)(B+\lambda I)^{-1/2}\right]$$

Proof of Theorem 2.9. We decompose the excess risk as follows

$$\begin{aligned}
f'_{\mathbf{z},\lambda} - f_{\mathcal{H}} &= \left(T'_{\mathbf{x}}+\lambda\right)^{-1}\left\{\left(g'_{\mathbf{z}} - g'\right) + \left(T' - T'_{\mathbf{x}}\right)f_\lambda\right\} && (2.49)\\
&+ \left(T'_{\mathbf{x}}+\lambda\right)^{-1}T'\left(f'_{\mathcal{H}} - f_{\mathcal{H}}\right) && (2.50)\\
&+ \left(\left(T'_{\mathbf{x}}+\lambda\right)^{-1}\left(T-T'\right)+I\right)\left(f_\lambda - f_{\mathcal{H}}\right) && (2.51)
\end{aligned}$$

where $T'_{\mathbf{x}} = S_{\mathbf{x}}^{\top} M_{\mathbf{v}} S_{\mathbf{x}}$, $g'_{\mathbf{z}} = S_{\mathbf{x}}^{\top} M_{\mathbf{v}} \mathbf{y}$ and $g' = L' f'_{\mathcal{H}}$. Let us bound the $L^2(X, \rho_X^{\mathrm{te}})$ norm of each term in the decomposition. For (2.49) we have

$$\left\| \left(T'_{\mathbf{x}} + \lambda\right)^{-1} \left\{ \left(g'_{\mathbf{z}} - g'\right) + \left(T' - T'_{\mathbf{x}}\right) f_\lambda \right\} \right\|_{\rho_X^{\mathrm{te}}} \leq \|T^{1/2}(T' + \lambda)^{-1/2}\|_{\mathrm{op}} \frac{S'_1 + S'_2}{S'_3} \tag{2.52}$$

where S'_i, $i = 1, 2, 3$, is defined similarly as S_i in the proof of the Theorem 2.5 with testing measure ρ_X^{te} changed by ρ'_X. Using Proposition 2.17 and repeating the arguments used to bound $\frac{S_1 + S_2}{S_3}$ in Theorem 2.5, we get with probability $1 - \delta$

$$\left\| \left(T'_{\mathbf{x}} + \lambda\right)^{-1} \left\{ \left(g'_{\mathbf{z}} - g'\right) + \left(T' - T'_{\mathbf{x}}\right) f_\lambda \right\} \right\|_{\rho_X^{\mathrm{te}}} \leq 16 \log\left(\frac{6}{\delta}\right) \frac{(M + R)}{\sqrt{1 - \mu}} \left(\frac{V}{n\sqrt{\lambda}} + \gamma \sqrt{\frac{\mathcal{N}'(\lambda)^{1-q'}}{n\lambda^{q'}}} \right),$$

where $\mu = \mu_{\max}\left((T + \lambda I)^{-1/2}(T - T')(T + \lambda I)^{-1/2}\right)$.

To bound the (2.50), notice that

$$\begin{aligned}
\| \left(T'_{\mathbf{x}} + \lambda\right)^{-1} T' \|_{\mathrm{op}} =& \| \left(T'_{\mathbf{x}} + \lambda\right)^{-1} (T' + \lambda)(T' + \lambda)^{-1} T' \|_{\mathrm{op}} \\
\leq& \| \left(T'_{\mathbf{x}} + \lambda\right)^{-1} (T' + \lambda) \|_{\mathrm{op}} \\
=& \left\| (T' + \lambda)^{-\frac{1}{2}} \left\{ I - (T' + \lambda)^{-\frac{1}{2}} \left(T' - T'_{\mathbf{x}}\right) (T' + \lambda)^{-\frac{1}{2}} \right\}^{-1} (T' + \lambda)^{\frac{1}{2}} \right\|_{\mathrm{op}} \leq 4
\end{aligned}$$

where the last inequality follows form the same argument as (2.48). So $\left(T'_{\mathbf{x}} + \lambda\right)^{-1} T' \left(f'_{\mathcal{H}} - f_{\mathcal{H}}\right) \leq 4 \| \left(f'_{\mathcal{H}} - f_{\mathcal{H}}\right) \|_{\rho_X^{\mathrm{te}}}$. Similarly, for the (2.51) we have

$$\begin{aligned}
\| \left(T'_{\mathbf{x}} + \lambda\right)^{-1} \left(T - T'\right) + I \|_{\mathrm{op}} =& \| \left(T'_{\mathbf{x}} + \lambda\right)^{-1} \left(T' + \lambda\right) \left(T' + \lambda\right)^{-1} \left(T - T'\right) + I \|_{\mathrm{op}} \\
\leq& \| \left(T'_{\mathbf{x}} + \lambda\right)^{-1} \left(T' + \lambda\right) \|_{\mathrm{op}} \| \left(T' + \lambda\right)^{-1} \left(T - T'\right) + I \|_{\mathrm{op}} \\
\leq& 4 \| \left(T' + \lambda\right)^{-1} \left(T + \lambda - \left(T' + \lambda\right)\right) + I \|_{\mathrm{op}} \\
\leq& 4 \| \left(T' + \lambda\right)^{-1} \left(T + \lambda\right) \|_{\mathrm{op}} \\
\leq& \frac{4}{1 - \mu}.
\end{aligned}$$

Combining all together we get

$$\begin{aligned}
\|f'_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \leq& 16 \log\left(\frac{6}{\delta}\right) \frac{(M + R)}{\sqrt{1 - \mu}} \left(\frac{V}{n\sqrt{\lambda}} + \gamma \sqrt{\frac{\mathcal{N}'(\lambda)^{1-q'}}{n\lambda^{q'}}} \right) \\
&+ \frac{4}{1 - \mu} \|f_\lambda - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} + 4 \|f'_{\mathcal{H}} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \\
\leq& 16 \log\left(\frac{6}{\delta}\right) \frac{(M + R)}{1 - \mu} \left(\frac{V}{n\sqrt{\lambda}} + \gamma \sqrt{\frac{\mathcal{N}'(\lambda)^{1-q'}}{n\lambda^{q'}}} \right) \\
&+ \frac{4}{1 - \mu} \|f_\lambda - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} + 4 \|f'_{\mathcal{H}} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}}.
\end{aligned}$$

Using Proposition 2.21 and balancing the first and second terms of the right hand side of the above equation completes the proof. □

2.8.3 Upper bound for clipped IW-KRR

The following lemma bounds $\|T^{1/2}(T_D+\lambda)^{-1/2}\|_{\mathrm{op}}$ uniformly over λ.

Lemma 2.18. *Let $T_D = \int T_x w_D(x) d\rho^{\mathrm{tr}}$ and $D = 2\Sigma\mathcal{N}(\lambda)$. Then*

$$\left\|T^{1/2}(T_D+\lambda)^{-1/2}\right\|_{\mathrm{op}} \leq 2. \tag{2.53}$$

Proof. Observe that

$$\begin{aligned}\|T^{1/2}(T_D+\lambda)^{-1/2}\|_{\mathrm{op}} &\leq \|T(T_D+\lambda)^{-1}\|_{\mathrm{op}}^{1/2} \\ &= \left\|T(T+\lambda)^{-1}\left\{I-(T-T_D)(T+\lambda)^{-1}\right\}^{-1}\right\|_{\mathrm{op}} \\ &\leq \frac{1}{1-\|(T-T_D)(T+\lambda)^{-1}\|_{\mathrm{op}}}\end{aligned}$$

provided that $\|(T-T_D)(T+\lambda)^{-1}\|_{\mathrm{op}} < 1$. Let us show that it is indeed the case when $D = 2\mathcal{N}(\lambda)$. We have

$$\begin{aligned}\|(T-T_D)(T+\lambda)^{-1}\|_{\mathrm{op}} &= \left\|\int T_x(T+\lambda)^{-1}\left(w(x)-w_D(x)\right) d\rho_X^{\mathrm{tr}}(x)\right\|_{\mathrm{op}} \\ &\leq \int \left\|T_x(T+\lambda)^{-1}\right\|_{\mathrm{op}} \left(w(x)-w_D(x)\right) d\rho_X^{\mathrm{tr}}(x)\end{aligned}$$

An intermediate step in the proof of Lemma 13 of Fischer and Steinwart (2020) shows that

$$\left\|T_x(T+\lambda)^{-1}\right\|_{\mathrm{op}} = \sum_{i\geq 1}\frac{\mu_i}{\mu_i+\lambda}e_i^2(x),$$

where $\{\mu_i, e_i\}_{i\geq 1}$ is the eigenvalue-eigenvector pair of the covariance operator T. Consequently, we have

$$\|(T-T_D)(T+\lambda)^{-1}\|_{\mathrm{op}} \leq \sum_{i\geq 1}\frac{\mu_i}{\mu_i+\lambda}\int e_i^2(x)\left(w(x)-w_D(x)\right) d\rho_X^{\mathrm{tr}}(x). \tag{2.54}$$

For the integral we have

$$\int e_i^2(x)\left(w(x)-w_D(x)\right) d\rho_X^{\mathrm{tr}}(x) = \int_{w\geq D} e_i^2(x)\left(w(x)-D\right) d\rho_X^{\mathrm{tr}}(x) \leq \int_{w\geq D} e_i^2(x) d\rho_X^{\mathrm{te}}(x).$$

For the last integral by using the integrebility of IW, assumption (2.33) and Markov inequality we have

$$\int_{w\geq D} e_i^2(x)d\rho_X^{\mathrm{te}}(x) \leq \rho_X^{\mathrm{te}}(w \geq D) \leq \frac{\Sigma}{D},$$

therefore

$$\int e_i^2(x)\left(w(x) - w_D(x)\right) d\rho_X^{\mathrm{tr}}(x) \leq \frac{\Sigma}{D}. \tag{2.55}$$

Considering (2.55) in (2.54) we have

$$\|(T - T_D)(T + \lambda)^{-1}\|_{\mathrm{op}} \leq \frac{1}{D}\sum_{i\geq 1}\frac{\mu_i}{\mu_i + \lambda} = \frac{\Sigma\mathcal{N}(\lambda)}{D}$$

Choosing $D = 2\Sigma\mathcal{N}(\lambda)$ finishes the proof. □

For the proof we use the following decomposition of the error

$$\left\|f_{\mathbf{z},\lambda}^D - f_{\mathcal{H}}\right\|_{\rho_X^{\mathrm{te}}} \leq \left\|f_{\mathbf{z},\lambda}^D - f_\lambda^D\right\|_{\rho_X^{\mathrm{te}}} + \left\|f_\lambda^D - f_{\mathcal{H}}\right\|_{\rho_X^{\mathrm{te}}}, \tag{2.56}$$

where $f_\lambda^D = (T_D + \lambda I)^{-1}Tf_{\mathcal{H}}$. For the approximation part we have the following lemma.

Lemma 2.19. *Let $f_{\mathcal{H}}$ satisfies the Assumption 1 for some $r > 0$. Then, the following estimate holds*

$$\left\|f_\lambda^D - f_{\mathcal{H}}\right\|_{\rho_X^{\mathrm{te}}} \leq 2\lambda^r \left\|L^{-r}f_{\mathcal{H}}\right\|_{\rho_X^{\mathrm{te}}} \qquad \textit{if } r \leq 1. \tag{2.57}$$

Furthermore, for $r > 0.5$

$$\left\|f_\lambda^D\right\|_{\mathcal{H}} \leq \kappa^{-\frac{1}{2}+r}\left\|L^{-r}f_{\mathcal{H}}\right\|_{\rho_X^{\mathrm{te}}} \leq \left\|L^{-r}f_{\mathcal{H}}\right\|_{\rho_X^{\mathrm{te}}}. \tag{2.58}$$

Proof. By the identity $A(A+\lambda I)^{-1} = I - \lambda(A+\lambda)^{-1}$ valid for $\lambda > 0$ and any bounded self-adjoint positive operator, we have

$$\left(I - (T_D + \lambda)^{-1}T_D\right)f_{\mathcal{H}} = \lambda(T_D{+}\lambda)^{-1}f_{\mathcal{H}} = \lambda^r\left(\lambda^{1-r}(T_D + \lambda)^{-(1-r)}\right)\left((T_D + \lambda)^{-r}T^r\right)T^{-r}f_{\mathcal{H}}.$$

From the equality above by taking the norm we have

$$\left\|f_\lambda^D - f_{\mathcal{H}}\right\|_{\rho_X^{\mathrm{te}}} \leq \lambda^r \left\|\lambda^{1-r}(T_D + \lambda)^{-(1-r)}\right\|_{\mathrm{op}}\left\|(T_D + \lambda)^{-r}T^r\right\|_{\mathrm{op}}\left\|T^{-r}f_{\mathcal{H}}\right\|_{\rho_X^{\mathrm{te}}}.$$

Note that $\left\|\lambda^{1-r}(T_D + \lambda)^{-(1-r)}\right\|_{\mathrm{op}} \leq 1$ and $\|T^{-r}f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \leq R$ by Assumption 1. From proposition 2.20 and Lemma 2.18 we have $\|(T_D + \lambda)^{-r}T^r\|_{\mathrm{op}} \leq \left\|(T_D + \lambda)^{-1}T\right\|_{\mathrm{op}}^r \leq 2.$

Regarding the second estimate, if $r > 1/2$, since $\|T\| \leq 1$, we obtain,

$$\begin{aligned}\left\|f_\lambda^D\right\|_{\mathcal{H}} &= \left\|(T_D+\lambda)^{-1}T_D f_{\mathcal{H}}\right\|_{\mathcal{H}} \\ &= \left\|(T_D+\lambda)^{-1}T_D T^{r-\frac{1}{2}}T^{\frac{1}{2}-r}f_{\mathcal{H}}\right\|_{\mathcal{H}} \\ &\leq \|T\|_{\text{op}}^{r-\frac{1}{2}}\left\|T^{-r}f_{\mathcal{H}}\right\|_{\rho_X^{\text{te}}} \leq \left\|T^{-r}f_{\mathcal{H}}\right\|_{\rho_X^{\text{te}}}.\end{aligned}$$

□

To bound the estimation error we use the decomposition similar to (2.41). It is not difficult to show that

$$\left\|f_{\mathbf{z},\lambda}^D - f_\lambda^D\right\|_{\rho_X^{\text{te}}} \leq \frac{S_0}{1-S_1}(S_2+S_3), \tag{2.59}$$

where

$$\begin{aligned}S_0 &:= \left\|T^{1/2}(T_D+\lambda)^{-1/2}\right\|, \\ S_1 &:= \left\|(T_D+\lambda)^{-\frac{1}{2}}\left(T_D - T_{\mathbf{x},D}\right)(T_D+\lambda)^{-\frac{1}{2}}\right\|_{\text{HS}}, \\ S_2 &:= \left\|(T_D+\lambda)^{-\frac{1}{2}}\left(g_{\mathbf{z},D} - g_D\right)\right\|_{\mathcal{H}}, \\ S_3 &:= \left\|(T_D+\lambda)^{-\frac{1}{2}}\left(T_D - T_{\mathbf{x},D}\right)f_\lambda^D\right\|_{\mathcal{H}}.\end{aligned}$$

By Lemma 2.18, $S_0 \leq \left\|T(T_D+\lambda)^{-1}\right\|^{1/2} \leq 2$.

Application of Proposition 1.10 to each of S_i, $i = 1, 2, 3$, yields to the following bounds with probability at least $1-\delta/3$

$$S_i \leq \frac{2L_i\log(6/\delta)}{n} + \sigma_i\sqrt{\frac{2\log(6/\delta)}{n}} \tag{2.60}$$

where, as it can be straightforwardly verified, the constants L_i and σ_i are given by the expressions

$$\begin{aligned}L_1 &= 2\frac{D}{\lambda}, \quad \sigma_1 = 2\sqrt{\frac{\Sigma\mathcal{N}(\lambda)}{\lambda}}, \\ L_2 &= 2\frac{MD}{\sqrt{\lambda}}, \quad \sigma_2 = 2\sqrt{\Sigma\mathcal{N}(\lambda)}, \\ L_3 &= 2\frac{\|f_\lambda^D\|_{\mathcal{H}}D}{\sqrt{\lambda}}, \quad \sigma_3 = 2\|f_\lambda^D\|_{\mathcal{H}}\sqrt{\Sigma\mathcal{N}(\lambda)}.\end{aligned}$$

Now, by choosing $n\lambda \geq \Sigma\mathcal{N}(\lambda)\log^2\frac{6}{\delta}$, with probability greater than $1-\delta/3$, we have

$$S_1 \leq 2\left(\frac{2D}{n\lambda} + \sqrt{\frac{\Sigma\mathcal{N}(\lambda)}{n\lambda}}\right)\log\frac{6}{\delta}$$

$$\begin{aligned} D = 2\Sigma\mathcal{N}(\lambda) \qquad & \leq 4\frac{\Sigma\mathcal{N}(\lambda)\log^2\frac{6}{\delta}}{\lambda n} + \sqrt{\frac{\Sigma\mathcal{N}(\lambda)\log^2\frac{6}{\delta}}{\lambda n}} \\ & \leq \frac{1}{4} + \frac{1}{2} = \frac{3}{4}. \end{aligned}$$

Now, if we combine the estimate (2.59) with the bounds given in (2.60), then we get with probability at least $1-\delta$

$$\|f^D_{\mathbf{z},\lambda} - f^D_\lambda\|_{\rho_X^{\mathrm{te}}} \leq 16\log\left(\frac{6}{\delta}\right)(M+R)\left(\frac{D}{n\sqrt{\lambda}} + \sqrt{\frac{\Sigma\mathcal{N}(\lambda)}{n}}\right). \tag{2.61}$$

Form the decomposition (2.56) with the bounds (2.61) and (2.57) we get

$$\begin{aligned} \|f^D_{\mathbf{z},\lambda} - f_\mathcal{H}\|_{\rho_X^{\mathrm{te}}} & \leq 16\log\left(\frac{6}{\delta}\right)(M+R)\left(\frac{D}{n\sqrt{\lambda}} + \frac{E_s\Sigma}{\sqrt{n\lambda^s}}\right) + \lambda^r R \\ & \leq 16\log\left(\frac{6}{\delta}\right)(M+R)\lambda^r\left(\frac{D\lambda^{r+s-1/2}}{n\lambda^{2r+s}} + \frac{\Sigma E_s}{\sqrt{n\lambda^{2r+s}}}\right) + \lambda^r R \\ \text{(2.34)} \qquad & \leq (M+R)\lambda^r\left(\frac{\lambda^{s+r-1/2}}{8\log(6/\delta)E_s} + 1\right) + \lambda^r R \\ & \leq 2(M+R)\lambda^r, \quad \text{for} \quad s \in (0,1]. \end{aligned}$$

2.9 Auxiliary Results

Proposition 2.20. *Let A, B be to self-adjoint, positive operators on a Hilbert space. Then for any $s \in [0,1]$:*

$$\|A^s B^s\|_{\mathrm{op}} \leq \|AB\|^s_{\mathrm{op}}.$$

Proposition 2.21. *Let $f_\mathcal{H}$ satisfies the Assumption 1 for some $r > 0$. Then, the following estimate holds*

$$\|f_\lambda - f_\mathcal{H}\|_{\rho_X^{\mathrm{te}}} \leq \lambda^r \left\|L^{-r} f_\mathcal{H}\right\|_{\rho_X^{\mathrm{te}}} \qquad \textit{if } r \leq 1. \tag{2.62}$$

Furthermore, for $r > 0.5$

$$\|f_\lambda\|_\mathcal{H} \leq \kappa^{-\frac{1}{2}+r}\left\|L^{-r} f_\mathcal{H}\right\|_{\rho_X^{\mathrm{te}}} \leq \left\|L^{-r} f_\mathcal{H}\right\|_{\rho_X^{\mathrm{te}}}. \tag{2.63}$$

Chapter 3

Target Shift

3.1 Introduction

In this chapter, we consider the target shift (Lipton et al., 2018; Storkey, 2009; Zhang et al., 2013) problem, where it is assumed that the conditional distribution of the input x given the output y are identical in the source and target populations, but the marginal distribution of y differs:

$$\rho^{\mathrm{tr}}(x, y) = \rho(x|y)\rho_Y^{\mathrm{tr}}(y), \quad \rho^{\mathrm{te}}(x, y) = \rho(x|y)\rho_Y^{\mathrm{te}}(y). \tag{3.1}$$

Unlike the covariate shift discussed in the previous chapter, the target shift aligns with the *anticausal* setting in which the target y cause the input x (Schölkopf et al., 2012). Target shift arises in infectious disease diagnostic problems, where the features are observed symptoms and the target is the underlying disease state. In this setting, the distribution of the symptoms given the disease remains unchanged but we expect larger fraction of infected people during the pandemic.

In the literature of target shift correction the main emphases is in the estimation of the IW (Azizzadenesheli et al., 2019; Garg et al., 2020; Lipton et al., 2018), while the effectiveness of the importance-weighted risk minimizers remains under-explored. Having in mind the difficulties associated with the IW correction for covariate shift, three relevant questions need to be addressed for the target shift: What is the effect of importance weighting correction applied to the target shift scenario, and how is it different from the IW correction applied under covariate shift? What are the effects of large values of the weighting function on the generalization properties of the IW correction? How important is it to distinguish the misspecified and well-specified scenarios under target shift to choose a more accurate weighting function? The aim of this chapter is to gain a better theoretical understanding of the nature of the bias associated with the target shift scenario and the role of IW in the bias correction.

3.2 Importance weighting correction under the Target Shift

The crucial difference between covariate and target shift scenarios consists in the fact that under target shift the training and testing regression functions are different. So the learning problem under target shift is to estimate the testing regression function $f_{\rho^{\mathrm{te}}} : X \to Y$ defined from the testing conditional distribution $\rho^{\mathrm{te}}(y|x)$ as

$$f_{\rho^{\mathrm{te}}}(x) := \int_Y y \; d\rho^{\mathrm{te}}(y|x), \quad x \in X. \tag{3.2}$$

Importance weighted kernel ridge regression is defined similarly as in (2.14) but with IW function defined on the output space:

$$f^{\mathrm{IW}}_{\mathbf{z},\lambda} := \operatorname*{arg\,min}_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} w_Y(y_i)\left(f\left(x_i\right)-y_i\right)^2 + \lambda\|f\|^2_{\mathcal{H}} \right\}, \tag{3.3}$$

where $w_Y : Y \to \mathbb{R}_+$ is defined by

$$w_Y(y) = \frac{d\rho^{\mathrm{te}}_Y(y)}{d\rho^{\mathrm{tr}}_Y(y)},$$

assuming that $\rho^{\mathrm{te}}_Y \ll \rho^{\mathrm{tr}}_Y$. Throughout the chapter we assume that the importance weighting function is known and will be mainly concerned with the effects of weights on generalization properties of IW-KRR (3.3). When IW is not known, it can be efficiently estimated from the training and testing data. Let us mention that unlike the covariate shift, the weights $w_Y(y)$ cannot be directly estimated because ρ^{te}_Y is unknown on the test data.

The solution of the minimization problem (2.14) is unique for all $\lambda > 0$ and is given by

$$f_{\mathbf{z},\lambda} = \left(S^T_{\mathbf{x}} M_{\mathbf{w}_Y} S_{\mathbf{x}} + \lambda I\right)^{-1} S^T_{\mathbf{x}} M_{\mathbf{w}_Y} \mathbf{y} \tag{3.4}$$

where $M_{\mathbf{w}_Y}$ is the diagonal matrix with main diagonal entries $w_Y(y_i), i = 1, \dots, n$.

3.3 Learning Guarantees of IW-KRR under Target Shift

To provide the learning guarantees for IW-KRR under target shift, the condition similar to the Assumption 2 is needed.

Assumption 6. Let $w_Y = d\rho_Y^{\mathrm{te}}/d\rho_Y^{\mathrm{tr}}$. There exist positive constants W_Y and σ_Y such that for all integer $m \geq 2$

$$\int_Y w_Y^{m-1}(y)d\rho_Y^{\mathrm{te}}(y) \leq \frac{1}{2}m!W_Y^{m-2}\sigma_Y^2. \tag{3.5}$$

Before providing the generalization bound for (3.3) let us explain why IW correction is a reasonable approach. First consider the random variable $\xi := yK_x w_Y(y)$ on (Z, ρ^{tr}) with the values in the Hilbert space $\mathcal{H}$. Then, by the low of large numbers, the term $S_{\mathbf{x}}^T M_{\mathbf{w}_Y}\mathbf{y}$ in (3.4) converges to

$$\frac{1}{n}\sum_{i=i}^{n}\xi(z_i) \longrightarrow \int_Z yw_Y(y)K_x d\rho^{\mathrm{tr}}(y,x)$$

as the number of samples goes to infinity. But the integral is just the integral operator acting on the *testing* regression function $Lf_{\rho^{\mathrm{tr}}}$ (this can be easily verified by the decomposition of the measures and target shift assumption (3.1)). This shows that $S_{\mathbf{x}}^T M_{\mathbf{w}_Y}\mathbf{y}$ is a good approximation of $Lf_{\rho^{\mathrm{tr}}}$. Second with a function $f \in \mathcal{H}$, look at the random variable $\xi := w_Y(y)f(x)K_x$ on $\left(Z, \rho^{\mathrm{tr}}\right)$ with values in $\mathcal{H}$. Again we have

$$S_{\mathbf{x}}^{\top} M_{\mathbf{w}_Y} S_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\xi\left(z_i\right) \longrightarrow Tf$$

meaning that $S_{\mathbf{x}}^{\top} M_{\mathbf{w}_Y} S_{\mathbf{x}}$ is a good approximation of the covariance operator T. Thus $f_{\mathbf{z},\lambda}^{\mathrm{IW}}$ should approximate $f_\lambda = (T+\lambda I)^{-1}Lf_{\mathcal{H}}$ well, and one would expect good error analysis of $f_{z,\lambda}^{\mathrm{IW}} - f_\lambda$ in the space $\mathcal{H}$. This observation is made precise by the following theorem.

Theorem 3.1. *Let ρ^{te} and ρ^{tr} be the distributions on $X \times [-M, M]$, where $M > 0$ is some constant, satisfying Assumptions 1,3 and 6. Let, let n and λ satisfy the constraints $\lambda \leq \|T\|_{\mathrm{op}}$ and*

$$\lambda = \left(\frac{8E_s(\sqrt{W_Y}+\sigma_Y)\log\left(\frac{6}{\delta}\right)}{\sqrt{n}}\right)^{\frac{2}{2r+s}} \tag{3.6}$$

for $\delta \in (0,1)$, $r \geq 0.5$ and $s \in (0,1]$. Then, with probability greater than $1-\delta$, it holds

$$\|f_{\mathbf{z},\lambda}^{\mathrm{IW}} - f_{\mathcal{H}}\|_{\rho_X^{\mathrm{te}}} \leq 3\left(M+R\right)\left(\frac{8E_s(\sqrt{W_Y}+\sigma_Y)\log\left(\frac{6}{\delta}\right)}{\sqrt{n}}\right)^{\frac{2r}{2r+s}}. \tag{3.7}$$

A comparison with the convergence rate of uniformly weighted KRR in the absence of target shift (theorem 1.13) leads to the conclusion that IW-KRR under target shift is minimax optimal for a properly chosen regularisation parameter. For the IW-KRR under covariate shift, the same rates are achieved only for the bounded IW, while a weaker boundedness assumption on the weights results in slower rates. In particular, it means that the IW correction of target shift is less sensitive toward the large shift in the output space, whenever the output assumed to be bounded.

3.4 Effect of Using Incorrect Weights

We extend the IW correction framework by considering more general weighted risk minimization problem

$$f'_{\mathbf{z},\lambda} := \operatorname*{arg\,min}_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} v(y_i) \left(f\left(x_i\right) - y_i \right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\} \tag{3.8}$$

where now $v(y) = \frac{d\rho'_Y(y)}{d\rho^{\mathrm{tr}}_X(y)}$ for some measure $\rho'_Y \ll \rho^{\mathrm{tr}}_Y$.

As we have already discussed in the previous chapter, the rationale behind the IW correction in covariate shift is to eliminate the bias associated with the mismatch between the projections of the regression function. In the target shift, the main focus of the risk adjustment is to eliminate the difference between the regression function (2.2) and the function $\phi(x)/\psi(x)$ induced by the candidate distribution ρ', where

$$\phi(x) = \int_Y y \frac{v_Y(y)}{w_Y(y)} d\rho^{\mathrm{te}}(y|x), \quad \psi(x) = \int_Y \frac{v(y)}{w_Y(y)} d\rho^{\mathrm{te}}(y|x).$$

To understand the meaning behind the function $\phi(x)/\psi(x)$ consider the following two extremes: when $v(y) \equiv 1$ and when $v_Y = w_Y$. In the former case no correction to KRR is applied and $\phi(x)/\psi(x) = f_{\rho^{\mathrm{tr}}}(x) = \int y d\rho^{\mathrm{tr}}(y|x)$. In the latter case $\phi(x)/\psi(x)$ matches with the testing regression function as $\phi(x) = f_{\rho^{\mathrm{te}}}$ and $\psi(x) = 1$.

The theorem below provides the learning guarantees of W-KRR under the target shift.

Theorem 3.2. *Let ρ^{te} and ρ' be the distributions on $X \times [-M, M]$, where $M > 0$ is some constant, satisfying target shift condition and Assumptions 1,2,4. Furthermore, let n and λ satisfy the constraints $\lambda \leq \|T\|_{\mathrm{op}}$ and*

$$\lambda = \left(\frac{8DE_s(\sqrt{V_Y} + \sigma_Y) \log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{\frac{2}{2r+s}} \tag{3.9}$$

for $\delta \in (0,1)$, $s \in (0,1]$ and $D = \max\{1, 1/\inf \psi(x)\}$ Then, for $r \geq 0.5$, with probability greater than $1 - \delta$, it holds

$$\|f_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho^{\mathrm{te}}_X} \leq 6D\left(M + R\right) \left(\frac{8DE_s(\sqrt{V_Y} + \sigma_Y) \log\left(\frac{6}{\delta}\right)}{\sqrt{n}} \right)^{\frac{2r}{2r+s}} + 4 \left\| \frac{\phi}{\psi} - f_{\rho^{\mathrm{te}}} \right\|_{\rho^{\mathrm{te}}_X}. \tag{3.10}$$

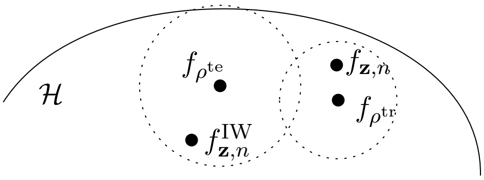


Figure 3.1: Performance of the IW-KRR and the uniformly weighted KRR under target shift: the uniformly weighted KRR approximates the training regression function which could significantly differ from the testing one.

In the case of uniform weights (when $\rho' = \rho^{\mathrm{tr}}$) the bias term of the bound (3.10) corresponds to the L_2 distance between the training and testing regression functions $\|f_{\rho^{\mathrm{te}}} - f_{\rho^{\mathrm{tr}}}\|_{\rho_X^{\mathrm{te}}}$ where

$$f_{\rho^{\mathrm{tr}}} = \int y d\rho^{\mathrm{tr}}(y|x).$$

Note that unlike the covariate shift case, there is no straightforward way to eliminate this bias even when the model is wellspecified (see Figure 3.1).

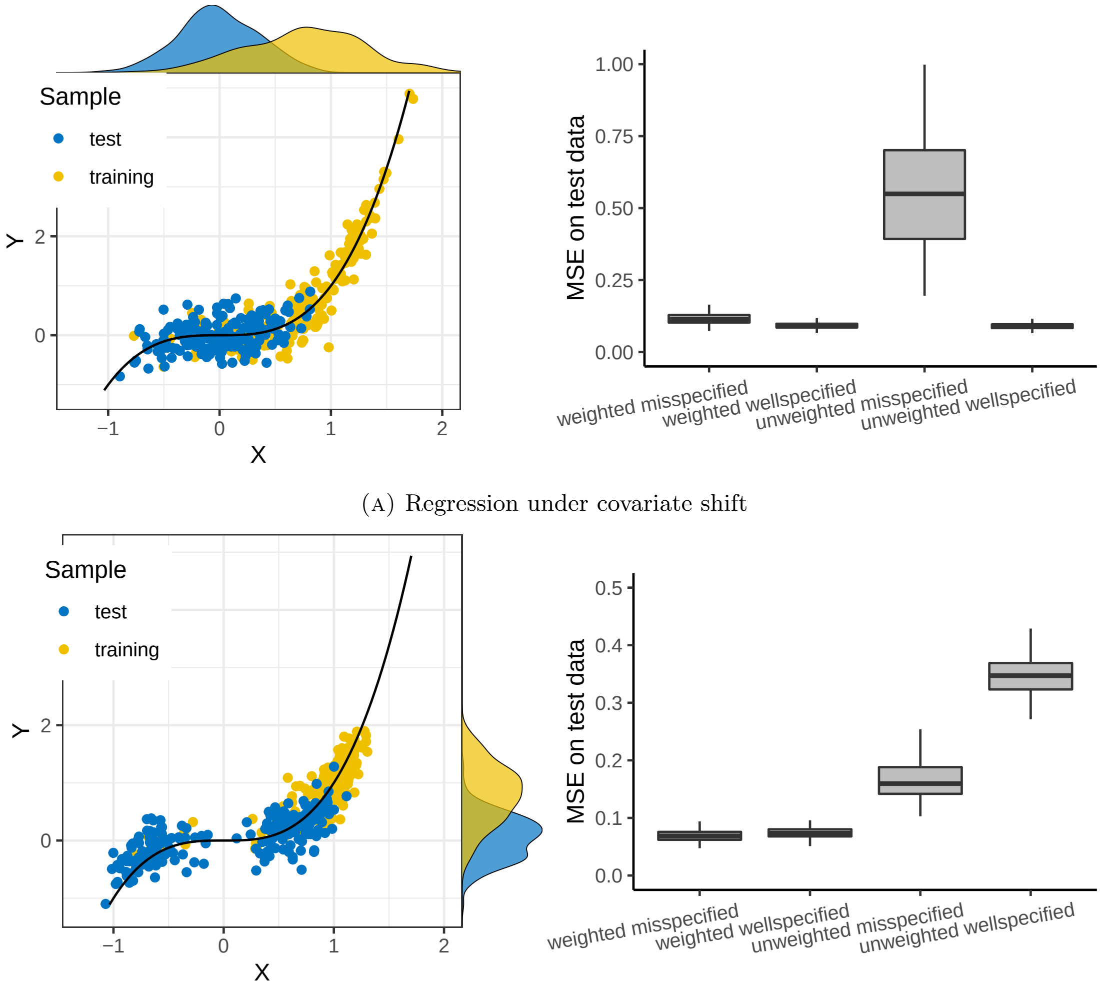
3.5 Simulations

We use simulations to justify the theoretical results of previous sections. The points we made in Theorems can be summarized as follows:

1. We can safely forget about the covariate shift for high capacity models.
2. Under covariate shift, IW correction is beneficial for low capacity models.
3. Under target shift importance weighting correction is beneficial regardless of the model capacity.

We consider a simple one-dimensional regression problem with the testing regression function being $f_{\rho^{\mathrm{te}}}(x) = x^3$.

Experimental setup - Covariate Shift We assume that $x \sim \mathcal{N}(0.8, 0.5)$ at training time and $x \sim \mathcal{N}(0, 0.35)$ at testing time. Output y assumed to be corrupted by homoscedastic Gaussian noise with mean $\mu = 0$ and standard deviation $\sigma = 0.3$. The regression function together with the training and test points generated in one random replication is shown on the left panel of Figure 3.2(a). The training and test sets consist of 200 data points each.



(A) Regression under covariate shift

(B) Regression under target shift

FIGURE 3.2: Performance of the least squares method for different shift scenarios. Left panels show the data points together with the regression function and right panels give the boxplot of the performances of different approaches for 200 random replications. "Misspecified" refers to the polynomial fit with degree 2. (a) Regression under covariate shift. (b) Regression under target shift

Experimental setup - Target Shift We assume that $y \sim \mathcal{N}(0, 0.5)$ at training time and $y \sim \mathcal{N}(1.5, 0.3)$ at testing time. The input is generated by $x = (y + \varepsilon)^{1/3}$, with $\varepsilon \sim \mathcal{N}(0, 0.3)$.

In the simulation we use the KRR with the polynomial kernel. For "misspecified" the degree of the kernel is two, while "wellspecified" corresponds to the cubic degree kernel. Left panel of Figure 3.2 shows the boxplot of the performances of all approaches, measured by the mean square error (MSE). Under covariate shift (a) the weighted models, as well as the unweighted misspecified model, perform equally well. Under target shift (b) deviation from the IW strategy leads to the larger test MSE regardless of the model capacity.

We do not report numerical experiments on the real data here as exhaustive experimental results on KRR, leading to similar conclusions, can already be found in Gretton et al. (2009); Zhang et al. (2013).

3.6 Proofs

Let us decompose the excess risk as follows

$$\begin{aligned} f'_{\mathbf{z},\lambda} - f_{\mathcal{H}} = &\underbrace{(T_{\mathbf{z}} + \lambda)^{-1} \{(g_{\mathbf{z}} - L\phi) + (LM_\psi - T_{\mathbf{z}}) f_\lambda\}}_{\text{I term}} \\ &+ \underbrace{(T_{\mathbf{z}} + \lambda)^{-1} \left\{L\left(\phi - f_{\rho^{\mathrm{te}}}\right) + (T - LM_\psi) f_\lambda\right\}}_{\text{II term}} \\ &+ (f_\lambda - f_{\mathcal{H}}). \end{aligned} \tag{3.11}$$

where $T_{\mathbf{z}} = S_{\mathbf{x}}^T M_{\mathbf{v}} S_{\mathbf{x}}$, $g_{\mathbf{z}} = S_{\mathbf{x}}^T M_{\mathbf{v}} \mathbf{y}$ and $M_\psi f = f\psi$ is a multiplication operator. Here $\mathbf{v} = (v_Y(y_1), \ldots, v_Y(y_n))$.

I term Similarly to the first term of the previous proof we can show

$$\|\text{I term}\|_{\rho_X^{\mathrm{te}}} \leq \frac{1}{1-\mu} \left(\frac{S_2' + S_3'}{1 - S_1'}\right)$$

where

$$\begin{aligned} S_2' &:= \left\|(T+\lambda)^{-\frac{1}{2}} (g_{\mathbf{z}} - L\phi)\right\|_{\mathcal{H}}, \\ S_3' &:= \left\|(T+\lambda)^{-\frac{1}{2}} (LM_\psi - T_{\mathbf{z}}) f_\lambda\right\|_{\mathcal{H}}, \\ S_1' &:= \left\|(LM_\psi+\lambda)^{-\frac{1}{2}} (LM_\psi - T_{\mathbf{z}}) (LM_\psi+\lambda)^{-\frac{1}{2}}\right\|_{\mathrm{HS}}, \end{aligned}$$

and

$$\mu = \mu_{\max}\left((T+\lambda I)^{-1/2}(T - LM_\psi)(T+\lambda I)^{-1/2}\right).$$

One can easily show that $\mu \leq 1 - \inf\psi$, therefore

$$\|\text{I term}\|_{\rho_X^{\mathrm{te}}} \leq D\left(\frac{S_2' + S_3'}{1 - S_1'}\right).$$

The following constants for S_1', S_2', S_3' can be straightforwardly verified,

$$\begin{aligned} L_1' = 2V_Y\frac{1}{\lambda}, &\quad \sigma_1' = 2\left(\sqrt{\frac{1}{\lambda}}\right)\sqrt{\mathcal{N}(\lambda)}\sigma_Y, \\ L_2' = 2MV_Y\sqrt{\frac{1}{\lambda}}, &\quad \sigma_2' = 2M\sigma_Y\sqrt{\mathcal{N}(\lambda)}, \\ L_3' = 2\|f_\lambda\|_{\mathcal{H}}V_Y\sqrt{\frac{1}{\lambda}}, &\quad \sigma_3' = 2\|f_\lambda\|_{\mathcal{H}}\sigma_Y\sqrt{\mathcal{N}(\lambda)}, \end{aligned}$$

Now, choosing $n\lambda \geq (V_Y + \sigma_Y^2)\mathcal{N}(\lambda)D^2$ with probability at least $1 - \delta/3$, we get

$$\begin{aligned} S_1' &= \left\| (LM_\psi + \lambda)^{-\frac{1}{2}} (LM_\psi - T_{\mathbf{x}}) (LM_\psi + \lambda)^{-\frac{1}{2}} \right\|_{\text{HS}} \\ &\leq D \left\| (T + \lambda)^{-\frac{1}{2}} (LM_\psi - T_{\mathbf{z}}) (T + \lambda)^{-\frac{1}{2}} \right\|_{\text{HS}} \\ &\leq 4 \log\left(\frac{6}{\delta}\right) \left(\frac{V_Y}{n\lambda(1-\mu)} + 2\sqrt{\frac{\sigma_Y^2 \mathcal{N}(\lambda)}{n\lambda(1-\mu)^2}} \right) \\ &\leq \frac{3}{4}. \end{aligned}$$

So, with probability at least $1 - \delta$, we have

$$\|\text{I term}\|_{\rho_X^{\text{te}}} \leq 16 \log\left(\frac{6}{\delta}\right) D\,(M + R)\,\lambda^r \left(\frac{V_Y \lambda^{r+s-0.5}}{n\lambda^{2r+s}} + \sigma_Y \frac{E_s}{\sqrt{n\lambda^{2r+s}}} \right). \tag{3.12}$$

II term For the second term notice that

$$\begin{aligned} \text{II term} &= (T_{\mathbf{z}} + \lambda)^{-1} \left\{ L\left(\phi - f_{\rho^{\text{te}}}\right) + (T - LM_\psi)\, f_\lambda \right\} \\ &= (T_{\mathbf{z}} + \lambda)^{-1} LM_\psi \left\{ \left(\frac{\phi}{\psi} - f_{\rho^{\text{te}}} \right) + \frac{1-\psi}{\psi}(f_\lambda - f_{\mathcal{H}}) \right\}. \end{aligned}$$

The argument used to bound the second term of the previous section allows us to conclude that $\|\,(T_{\mathbf{z}} + \lambda)^{-1} LM_\psi\|_{\text{op}} \leq 4$, therefore

$$\|\text{II term}\|_{\rho_X^{\text{te}}} \leq 4 \left\| \frac{\phi}{\psi} - f_{\rho^{\text{te}}} \right\|_{\rho_X^{\text{te}}} + 5D\|f_\lambda - f_{\mathcal{H}}\|_{\rho_X^{\text{te}}}. \tag{3.13}$$

Considering (3.12) and (3.13) in the decomposition (3.11) and using the proposition 2.21 we have with probability at least $1 - \delta$

$$\|f_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\text{te}}} \leq 16 \log\left(\frac{6}{\delta}\right) D\,(M + R)\,\lambda^r \left(\frac{V_Y \lambda^{r+s-0.5}}{n\lambda^{2r+s}} + \sigma_Y \frac{E_s}{\sqrt{n\lambda^{2r+s}}} \right) + 6D\lambda^r R \tag{3.14}$$

$$+ 4 \left\| \frac{\phi}{\psi} - f_{\rho^{\text{te}}} \right\|_{\rho_X^{\text{te}}}. \tag{3.15}$$

By choosing λ as in (3.9) we have

$$\|f_{\mathbf{z},\lambda} - f_{\mathcal{H}}\|_{\rho_X^{\text{te}}} \leq 6D(M + R)\lambda^r + 4 \left\| \frac{\phi}{\psi} - f_{\rho^{\text{te}}} \right\|_{\rho_X^{\text{te}}}$$

Substituting the expression (3.9) for λ in the inequality above, concludes the proof.

Chapter 4

Locally Smoothed Gaussian Process Regression

4.1 Introduction

In supervised learning tasks applied to a data set composed of observed input data and labels, the goal of function estimation is to establish a mapping between these two groups of observed quantities. Function estimation can be approached in various ways, and we can broadly divide algorithms in two categories, as *global* and *local*. Examples of global algorithms are Neural Networks (Neal, 1996) and kernel machines (Shawe-Taylor and Cristianini, 2004), which impose a functional form yielding a global representation of the function. The functional form is parameterized by a set of parameters which are optimized or inferred based on all the available data. The estimated model can later be used to query the function at any input points of interest. In local algorithms such as K-Nearest Neighbors (KNN), instead, the target point is fixed and the corresponding value of the function is estimated based on the closest data available.

Obviously, any global algorithm can be made local by training it only for the few training points located in the vicinity of the target test point. While it may seem that the idea of localizing global algorithms is not a very profound one, empirical evidence shows that localization could improve the performance of the best global models (Bottou and Vapnik, 1992). The idea of localization was therefore applied to global models such as SVMs (Blanzieri and Melgani, 2006, 2008). In addition to performance gains, by operating on smaller sets of data points, these local approaches enjoy computational advantages, which are particularly attractive for kernel machines for which scalability with the number of data points is generally an issue (Cheng et al., 2007; Segata and Blanzieri, 2010; Segata et al., 2012).

In this chapter, we develop novel ideas to implement a localization of Gaussian processes (GPs) in order to obtain performance gains, as well as computational ones. GPs are great candidates to benefit from computational speedups given that a naïve implementation requires expensive algebraic computations with the covariance matrix; denoting by n the number of input data, such operations cost $\mathcal{O}(n^3)$ operations and require storing $\mathcal{O}(n^2)$ elements, hindering the applicability of GPs to data sets beyond a few thousand data points Quinonero-Candela and Rasmussen (2005). Another issue with GPs is how to choose a suitable kernel for the problem at hand so as to avoid problems of model misspecification. Both of these issues have been addressed in various ways, by proposing scalable approximations based on inducing points Hensman et al. (2013) and random features Cutajar et al. (2017a); Rahimi and Recht (2008), and by composing GPs to obtain a rich and flexible class of equivalent kernels Wilson et al. (2016).

In this chapter we explore an alternative way to address scalability and kernel design issues by localizing GPs. In particular, we show how the localization operation leads to a particular form for the localized GP, and what is the effect on the kernel of this model. Furthermore, the localization makes it apparent how to implement the model with considerable gains compared to other approaches to approximate GPs. We demonstrate such performance gains on regression tasks on standard UCI benchmarks Asuncion and Newman (2007).

4.1.1 Related work

Local learning algorithms were introduced by Bottou and Vapnik (1992), with the main objective of estimating the optimal decision function for each single testing point. Examples of local learning algorithms include the well-known K-Nearest Neighbor regression (Altman, 1992) and local polynomial regression (Fan and Gijbels, 2018). These methods provide simple means for solving regression problems for the cases where training data are nonstationary or their size is prohibitively large for building a global model. However, neither of these methods provides ways to quantify uncertainty in predictions, which is a highly desirable feature in cost-sensitive applications.

Gaussian Process Regression (GPR) (Rasmussen and Williams, 2006) is a popular nonparametric regression method based on Bayesian principles which provides uncertainty estimates for its predictions. Similarly to other kernel methods (e.g., SVMs and KRR), GPR is a global method, meaning that it takes into account the whole dataset at prediction time. Thus, GPR inherits the computational complexity of global kernel methods, which is prohibitive for large datasets. Among the large class of scalable approximations for GPR, successful ones are based on Random Fourier Features (Rahimi and Recht,

2008) and on sparsification of the Gram matrix induced by the kernel (Rasmussen and Williams, 2006).

Random feature approximation of the kernel proposed in Rahimi and Recht (2008) is based on Bochner theorem and allows representing the kernel function as a dot product of (possibly infinite) feature maps applied to the input data. In practice, infinite feature maps are replaced by a finite Monte Carlo approximation. The disadvantage of this approach is that it is necessary to construct a specific random feature mapping for each type of kernel. While random feature approximations are known for popular kernels such as RBF (Rahimi and Recht, 2008) and polynomial (Pennington et al., 2015), there is no straightforward application of this method to approximate arbitrary kernels.

The Gram matrix sparsification approach is based on the idea of introducing so-called inducing points in order to approximate the full Gram matrix. One of the most popular methods in this family is the Nyström approximation (Rasmussen and Williams, 2006). The main drawback of this approach is that a low number of inducing points might lead to a poor approximation of the original model, which affects predictive performance. An important advancement within this family of approaches which provides a scalable variational formulation was proposed by Titsias (2009).

While providing good performance and scalability for large datasets, these approaches still require some design choices for the kernel. For stationary kernels, they assume that the same kernel is suitable for all the regions of input space and if data are nonstationary, this may harm the predictive performance. The literature has a wide range of proposals to address kernel design by incorporating ideas from deep learning Cutajar et al. (2017b); Wilson et al. (2016).

Recently, partitioning strategies have also gained some attention. The main idea is to divide the input space in regions where local estimators are defined Carratino et al. (2021); Meister and Steinwart (2016); Mücke (2019); Tandon et al. (2016). In partition-based methods, the main challenge is to define an effective partitioning of the space.

There are several approaches that use the idea of local learning for training GP models. The method proposed by Meier et al. (2014b) and extended by Meier et al. (2014a) mostly focuses on Bayesian parametric linear regression. The methods in these papers build an ensemble of local models centered at several fixed points, where each training point is weighted accordingly to the distance from the center of the model. Predictions are computed as a weighted sum of the local models. The authors claim that their approach extends to GPR, but in this case each local model considers the full training set. This means that these methods use localization to address nonstationarity, but poorly scale to large datasets. The method proposed by Snelson and Ghahramani (2007) proposes to build local GP models that use only subsets of the training data, but it lacks a mechanism that assigns importance weight for the training points for each local model according to the distance from the center of the model. That is why the model can make overconfident predictions for the points that lay far away from the centers of the local models. In (Gramacy and Apley, 2015) in order to obtain fast approximate prediction at a target point, the Authors propose a forward step-wise variable selection procedure to find the optimal sub-design.

This chapter is organized as follows: Section 4.2 introduces GPs and the idea of localization, along with a discussion on the connection between our localized GPs and local kernel ridge regression. The experimental campaign in Section 4.3 reports results on a variety of benchmarks for regression.

4.2 Gaussian Processes, Kernel Ridge Regression, and Localization

4.2.1 Gaussian Process Regression

A zero-mean Gaussian process (GP) $\{f(x) : x \in X \subset \mathbb{R}^d\}$ is a set of random variables $f(x)$ indexed by the input set X such that for each finite subset $\mathbf{x} = \{x_1, \ldots, x_m\} \subset X$ the random vector $(f(x_1), \ldots, f(x_m))$ is a zero-mean multivariate normal. The finite-dimensional distribution of such a process is determined by the covariance function $K : X \times X \to \mathbb{R}$, defined by

$$K(x, x') = E[f(x)f(x')]. \tag{4.1}$$

The fact that $f(x)$ is Gaussian process with covariance kernel K is commonly denoted by

$$f(x) \sim \mathcal{GP}(0, K(x, x')).$$

Given a data set $\mathcal{D}$ comprising a set of input-label pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1,\ldots,n}$, GPs can be used as a prior over functions to model the relationship between inputs and labels. The likelihood function for GPs applied to regression tasks can be derived from assuming that

$$y_i = f(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{4.2}$$

The GP prior, together with the Gaussian assumption on the relationship between $f(x)$ and y_i induces a multivariate normal distribution over $\mathbf{y} = (y_1, \ldots, y_n)^\top$ as

$$\mathbf{y} \sim \mathcal{N}(0, K_{\mathbf{xx}} + \sigma^2 I), \tag{4.3}$$

where $(K_{\mathbf{xx}})_{ij} = K(x_i, x_j)$.

In GP regression it is possible to compute the predictive distribution of the function values at any arbitrary input points $x \in X$ given $\mathcal{D}$ by means of well-known formulas for conditional distributions of Gaussian random vectors. Given the data set $\mathcal{D}$, it is straightforward to show (Rasmussen and Williams, 2006) that the posterior over the function is also a GP

$$f|\mathbf{y}, X \sim \mathcal{GP}(m(x), \mathcal{K}(x, x'))$$

with mean

$$m(x) = K_{x\mathbf{x}} \left(K_{\mathbf{xx}} + \sigma^2 I\right)^{-1} \mathbf{y}, \quad x \in X, \tag{4.4}$$

and covariance function

$$\mathcal{K}\left(x, x'\right) = K\left(x, x'\right) - K_{x\mathbf{x}} \left(K_{\mathbf{xx}} + \sigma^2 I\right)^{-1} K_{\mathbf{x}x'}, \quad x, x' \in X, \tag{4.5}$$

where $K_{\mathbf{x}x} = K_{x\mathbf{x}}^\top = \left(K\left(x_1, x\right), \ldots, K\left(x_n, x\right)\right)^\top$.

The problem with the expressions above is that they require solving a linear system involving a matrix of size $n \times n$. Direct methods to solve these operations require $\mathcal{O}\left(n^3\right)$ operations and storing $\mathcal{O}\left(n^2\right)$. Iterative solvers, instead, can reduce these complexities by relying exclusively on matrix-vector products, which require $\mathcal{O}\left(n^2\right)$ operation per iteration and do not need to store the Gram matrix (Cutajar et al., 2016; Filippone and Engler, 2015). However, a quadratic time complexity may still prohibitive for large-scale problems.

There is a rich literature on approaches that recover tractability by introducing approximations. One popular line of work introduces m so-called inducing-points as a means to approximate the whole GP prior Quinonero-Candela and Rasmussen (2005). This treatment of GPs was later extended within a scalable variational framework (Krauth et al., 2017; Titsias, 2009), making the complexity cubic in the number of inducing points m. Another approach, proposes ways to linearize GPs by obtaining an explicit set of features so as to obtain a close approximation to the original kernel-based model. Within this framework, a popular approach is based on random features (Rahimi and Recht, 2008). Denoting by Φ the $n \times D$ matrix obtained by applying a set of D random basis functions to the inputs in $x_1 \ldots, x_n$, these approximations are so that $\Phi\Phi^\top$ approximates in an unbiased way the original kernel matrix $K_{\mathbf{xx}}$, that is $E[\Phi\Phi^\top] = K_{\mathbf{xx}}$. For the Gaussian kernel, for example, a Fourier analysis shows that the basis functions that satisfy this property are trigonometric functions with random frequencies (Rahimi and Recht, 2008). This approach has been applied to GPs in Lázaro-Gredilla et al. (2010) and later made scalable by operating on mini-batches in Cutajar et al. (2017a).

4.2.2 Locally Smoothed Gaussian Process Regression

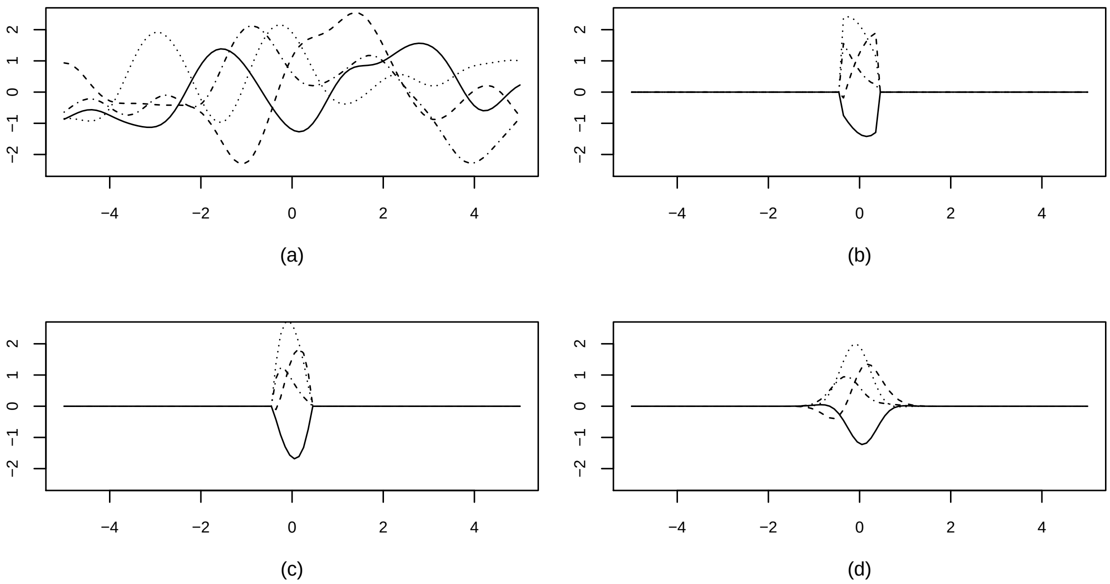
GP regression as formulated above is an example of a *global learner*; for a fixed training data $\mathcal{D}$ it builds a posterior distribution over functions that can be used to calculate the predictive distribution for any test inputs. To construct a predictive distribution for *any* input point it is necessary to use all the information available from the data $\mathcal{D}$, resulting in the need to do algebraic operations with the matrix $\left(K_{\mathbf{xx}} + \sigma^2 I\right)$. However, if we focus on the prediction problem locally, at a *given* target input x_0, most of the information carried by the (potentially large) covariance matrix might be neglected with little loss of information. The main idea behind localized GPs is to down-weight the contribution of the data points far from x_0, so that the structure of the covariance matrix is more adapted to the prediction task at a given point. To make this general idea work, we need to tackle two challenges. First, we need to specify what it means to be far or close to a given point; second, the change of the structure of the covariance matrix must give us a valid covariance matrix, i.e., the resulting covariance matrix should be symmetric and positive definite. Note that a simple truncation of the covariance function to obtain a compact-support covariance function may generally destroy positive definiteness Kaufman et al. (2008).

We accomplish localization of GPs in a straightforward manner as follows. We localize the target and the prior in the model (4.2) by multiplying them by the square root of the weighting function

$$k_h(x, x_0) := \frac{1}{h^d} k\left(\frac{\|x - x_0\|}{h}\right),$$

where $k : X \subset \mathbb{R}^d \to \mathbb{R}$ is a non-negative, integrable function satisfying $\int K(x)dx = 1$ and $\|\cdot\|$ is Euclidean norm on $\mathbb{R}^d$. Considering the square root of the weighting function will be convenient later when we discuss the link between local GPs and local Kernel Ridge Regression. Some classical examples of the weighting functions are given in Fig. 1. Because of the linearity of the weighting operation, the resulting model is another zero-mean Gaussian process $\tilde{f}(x)$ with covariance function given by

$$\tilde{K}(x, x'; x_0) = k_h^{\frac{1}{2}}(x, x_0) K(x, x') k_h^{\frac{1}{2}}(x', x_0). \tag{4.6}$$



(a)

(b)

(c)

(d)

FIGURE 4.1: (a) Samples from a global GP prior with exponential kernel. (b) Samples from a GP prior with exponential kernel localized by rectangular smoother centered at $x_0 = 0$. (c) Samples from a GP prior with exponential kernel localized by Epanechnikov smoother centered at $x_0 = 0$. (d) Samples from a GP prior with exponential kernel localized by Gaussian smoother centered at $x_0 = 0$.

In this formulation, we have localized the relationship between noisy targets and function realizations as

$$\tilde{y}_i = \tilde{f}(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{4.7}$$

with $\tilde{y}_i = \sqrt{k_h(x_i, x_0)}y_i$, and the prior is given by a zero-mean GP with the localized covariance kernel (4.6). The model (4.7) can be alternatively written as a model with heteroscedastic noise

$$y_i = f(x_i) + \frac{1}{\sqrt{k_h(x_i, x_0)}}\varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Making the noise parameter location-dependent can significantly improve the performance for problems where the assumption of a homoscedastic noise is not satisfied.

Theorem 4.1. *Let $I = \{i : \|x_i - x_0\| \leq h\}$, $\mathbf{x}_I = \{x_i : i \in I\}$ and $\mathbf{y}_I = \{y_i\}_{i \in I} \in \mathbb{R}^{|I|}$. Assume that (4.7) holds for the fixed target point $\mathbf{x}_0$. Then $f(\mathbf{x}_0) \mid \mathbf{y}_I$ is a Gaussian random variable with mean and variance given by*

$$\tilde{m}(x_0) = K_{x_0\mathbf{x}_I}\left(K_{\mathbf{x}_I\mathbf{x}_I} + \sigma^2 W_{x_0}^{-1}\right)^{-1}\mathbf{y}_I \tag{4.8}$$

$$\tilde{\mathcal{K}}\left(x_0, x_0\right) = K\left(x_0, x_0\right) - K_{x_0\mathbf{x}_I}\left(K_{\mathbf{x}_I\mathbf{x}_I} + \sigma^2 W_{x_0}^{-1}\right)^{-1} K_{\mathbf{x}_I x_0} \tag{4.9}$$

where W_{x_0} is the diagonal matrix with main diagonal entries $k_h(x_i, x_0)$, $x_i \in \mathbf{x}_I$.

Proof: Let x_0 be any fixed target point. Then the observations $\mathbf{y}_I \in \mathbb{R}^{|I|}$ and GP-function value at target point $f_0 = f\left(x_0\right) \in \mathbb{R}$ are jointly Gaussian such that

$$\left[\begin{array}{c} y \\ f_0 \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} \mathbf{0}_I \\ 0 \end{array}\right], \left(\begin{array}{cc} K_{\mathbf{x}_I\mathbf{x}_I} + \sigma^2 W_{x_0}^{-1} & K_{\mathbf{x}_I x_0} \\ K_{x_0\mathbf{x}_I} & K(x_0, x_0) \end{array}\right)\right).$$

Then the proposition follows from the basic formula for conditional distributions of Gaussian random vectors (see, e.g., Rasmussen and Williams (2006), Appendix A.2).

ALGORITHM 4.2.1

```
**Requires:** Data $Z$, target point $x_0$, localization parameter $h$, local kernel $k$, global kernel $K$ and noise variance $\sigma^2$.
**Outputs:** Posterior mean $\tilde{m}(x_0)$ and variance $\tilde{\mathcal{K}}(x_0, x_0)$.
1. Define the set of active observations $I := \{i : \|x_i - x_0\| \leq h\}$
2. Define the active input-output pair $\mathbf{x}_I := \{x_i : i \in I\}$ and $\mathbf{y}_I := \{y_i : i \in I\}$
3. Calculate the Grammian $K_{\mathbf{x}_I\mathbf{x}_I}$ and diagonal matrix $W_{x_0} := \operatorname{diag}(\{k_h(\mathbf{x}_i, x_0) : i \in I\})$
4. Calculate the Cholesky decomposition $L := \text{Cholesky}\left(K_{\mathbf{x}_I\mathbf{x}_I} + \sigma^2 W_{x_0}^{-1}\right)$
5. Calculate posterior mean $\tilde{m}(x_0) := K_{x_0\mathbf{x}_I}(L^{-1})^\top L^{-1}\mathbf{y}$
6. Calculate posterior variance $\tilde{\mathcal{K}}(x_0, x_0) = K(x_0, x_0) - K_{x_0,\mathbf{x}_I}(L^{-1})^\top L^{-1} K_{\mathbf{x}_I,x_0}$
```

Compared to global GPs, in the local formulation, in order to compute the posterior mean and variance, we need to invert $\left(K_{\mathbf{x}_I\mathbf{x}_I} + \sigma^2 W_{x_0}^{-1}\right)$. This might give a key advantage when dealing with large data sets, as the localization by compactly supported kernel (local) could significantly sparsify the Gram matrix corresponding to $K_{\mathbf{xx}}$. Denoting by s_0 the number of inputs for which the localizing weights are nonzero for a test point $\mathbf{x}_0$, the complexity of performing such an inversion is $\mathcal{O}(s_0^3)$. Another interesting observation is that the kernel function $\tilde{K}\left(x, x'\right)$ is potentially more flexible than the original kernel function; this is due to the multiplication by the localizing weighting function, which may introduce some interesting nonstationarity even for kernel functions which are stationary, depending on the choice of the weighting function.

The calculation of the predictive distribution in with Locally Smoothed Gaussian Process Regression (LSGPR) is described in Algorithm 4.2.2 The parameter selection in probabilistic models given by GPs is based on the *marginal log-likelihood* maximization, which, in our local formulation, can be defined as follows

$$\log p(\mathbf{y}_I|\mathbf{x}_I) = -\frac{1}{2}\mathbf{y}_I^\top \left(K_{\mathbf{x}_I\mathbf{x}_I} + \sigma^2 W_{x_0}^{-1}\right)^{-1} \mathbf{y}_I - \frac{1}{2}\log\left|K_{\mathbf{x}_I\mathbf{x}_I} + \sigma^2 W_{x_0}^{-1}\right| - \frac{n}{2}\log(2\pi)$$

Unfortunately, gradient-based optimization cannot be used to find the optimal localization parameter h, as the marginal log-likelihood is not continuously differentiable w.r.t. this parameter when compactly supported local kernels are used. The simplest way to resolve this problem is by using grid search for the localization parameter h, while kernel parameters can be optimized by gradient-based methods for any given h.

4.2.3 Local Kernel Ridge Regression

For every Gaussian process $f(x)$ with covariance function $K(x, x')$ there is a unique corresponding Hilbert space $\mathcal{H}$. This is commonly referred to as a *reproducing kernel Hilbert space* (RKHS) and constructed as a completion of the linear space of all functions

$$x \mapsto \sum_{i=1}^{k} \alpha_i K\left(a_i, x\right), \quad \alpha_1, \ldots, \alpha_k \in \mathbb{R}, a_1, \ldots, a_k \in X, k \in \mathbb{N}$$

relative to the norm induced by the inner product

$$\left\langle \sum_{i=1}^{k} \alpha_i K\left(s_i, \cdot\right), \sum_{j=1}^{l} \beta_j K\left(t_j, \cdot\right) \right\rangle_{\mathcal{H}} = \sum_{i=1}^{k} \sum_{j=1}^{l} \alpha_i \beta_j K\left(s_i, t_j\right).$$

It is well know that the posterior mean of Gaussian process regression can be alternatively derived by minimizing the regularized empirical risk over the RKHS (Kimeldorf and Wahba, 1970); see, e.g., Kanagawa et al. (2018) for a recent review. For local GPs, this corresponds to a weighted least square minimization over the RKHS with the weights given by $k_h(x, x_0)$, that is

$$\tilde{m}(x) = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 k_h\left(x_i, x_0\right) + \frac{\sigma^2}{n} \|f\|_{\mathcal{H}}^2. \tag{4.10}$$

Note that in the local formulation, for a given point x_0 one has to estimate both the parameters of the reproducing kernel and the width of the local kernel h. Here are two examples of well-known classical local methods which are the solution of the empirical risk minimization problem (4.10).

K-nearest neighbors. This model corresponds to the noise-free case ($\sigma = 0$) with a positive constant reproducing kernel and a rectangular local kernel whose width is adjusted to contain exactly k data points. The solution of the minimization problem (4.10) is the mean of the outputs corresponding to the k closest to x_0 input points.

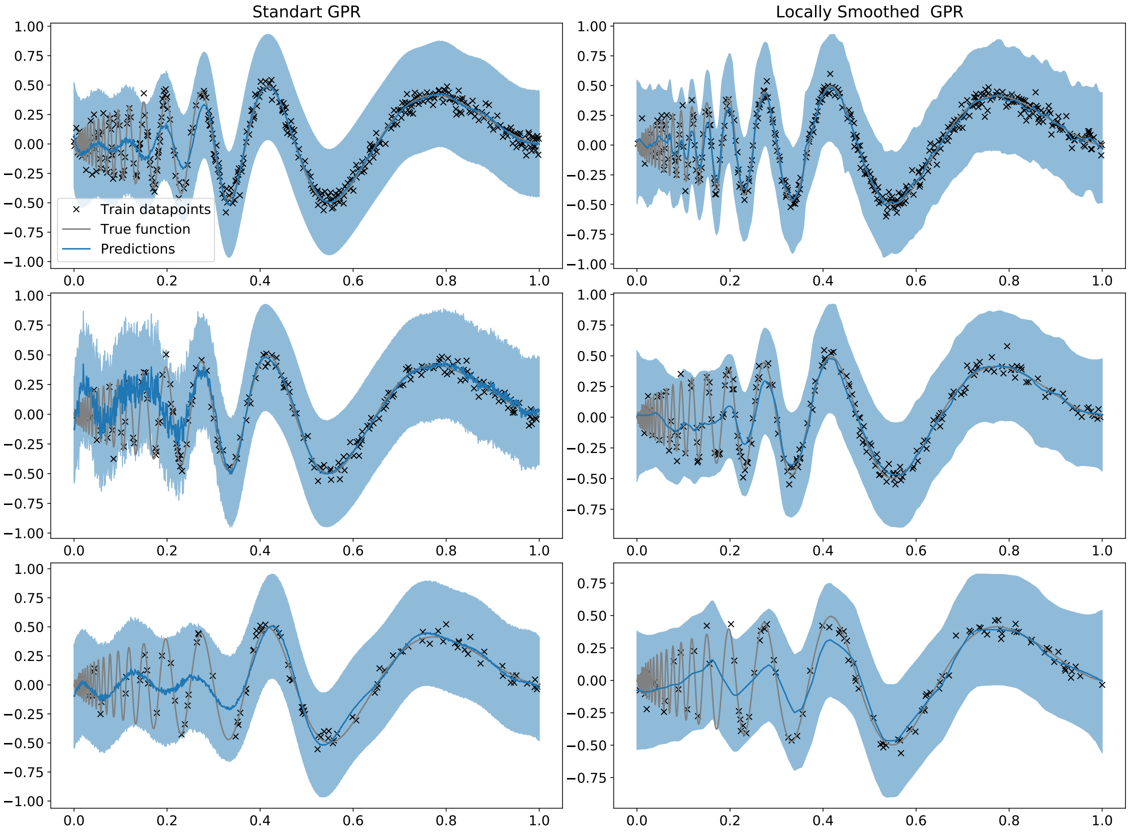


Figure 4.2: Illustration of the predictive distribution of GP (left) and LSGP (right) applied to data sampled from the Doppler function for 400 training points (top), 200 training points (middle) and 100 training points.

Local polynomial regression. If we use the polynomial kernel $K(x, x') = (1 + xx')^k$ for the space $\mathcal{H}$, and use any smooth local kernel (i.e. exponential), then in the noise-free case the solution of the minimization problem (4.10) is so called local polynomial regression Tsybakov (2009). In this special case, when the degree of the polynomial is 0, we have Nadaraya-Watson regression, which is the minimizer of the local squared loss over the constant function.

Table 4.1: UCI datasets used for evaluation.

Dataset	Training instances	Dimensionality
Yacht	308	6
Boston	506	13
Concrete	1030	8
Kin8nm	8192	8
Powerplant	9568	4
Protein	45730	9

4.3 Experiments

4.3.1 Toy dataset

In order to illustrate the behavior of the proposed Locally Smoothed Gaussian Process (LSGP), we start from a toy dataset generated from the Doppler function (Fig. 4.2).

$$y(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x+0.05}\right) + \varepsilon, \quad 0 \le x \le 1, \quad \varepsilon \sim \mathcal{N}(0, 0.1) \tag{4.11}$$

For this experiment, we used the RBF kernel and the Epanechnikov localizing kernel

$$k(x) = \frac{3}{4}(1 - |x|^2)\mathbb{I}(|x| \le 1). \tag{4.12}$$

We tuned the lengthscale parameter of the RBF kernel by optimizing the marginal log-likelihood of the model. We used the L-BFGS algorithm for gradient optimization Pytlak (2008). We chose the value of the parameter h of the localizing kernel that gave the best Mean Squared Error (MSE) on a validation set. The LSGP model used on average 7 training points to make a prediction. We compared the predictions of LSGP with the predictions of standard GP regression.

As we can see from Fig. (4.2), the Gaussian Process with RBF kernel is unable to make reasonable predictions in the region where the target function contains high frequency components. While some nonstationary covariance functions might be appropriate for this example, the combination of a standard stationary covariance function with the localization approach offers substantial modeling improvements.

4.3.2 UCI datasets

We evaluated the performance of the LSGP method on several problems from the UCI datasets collection and compared it against standard GPR, Deep GPs approximated with random features (Cutajar et al., 2017a), and k-nearest neighbors (KNN) regression. In particular, we aim to compare the predictive performance offered by the localization

Table 4.2: Comparison in terms of test set MSE between LSGPR a standard GPR.

Dataset	LSGPR Hilbert	LSGPR Epan.	GP	DeepGP	KNN
Yacht	**0.63±0.12**	2.02±0.58	1.09±0.05	0.93±0.13	57.80±16.65
Boston	14.78±0.88	15.30±1.28	17.94±0.71	**7.92±0.14**	23.30±2.58
Concrete	**34.79±1.07**	40.43±3.16	37.81±0.61	130.94±3.93	94.23±7.89
Kin8nm	0.01±0.000	0.01±0.00	0.01±0.000	0.06±0.00	0.01±0.00
Powerplant	14.65±0.34	**14.40±0.67**	16.85±0.69	14.57±0.15	15.35±0.49
Protein	36.85±2.15	**12.50±0.26**	17.03±0.57	16.94 ±0.16	19.89±4.26

against the baseline of exact GPR, and to verify that any performance gains are not just due to localization, meaning that we expect to outperform KNN. Because our model is more flexible than standard GPR, we also added to the comparison Deep GP models based on random features expansion. The size and dimensionality of these problems is outlined in Table 4.1. Since the Euclidean norm used in the local kernels depends on the units in each coordinate, all the datasets except the Protein were scaled within the $[0, 1]$ range. We used a standardization procedure for the Protein dataset because the baseline model worked much better with this type of preprocessing. For the Deep GP model we used the hyperparameters and the data preprocessing described in the original paper.

The LSGPR method requires creating a new local model with its own set of hyperparameters for each input point where the prediction has to be made. During the optimization, the kernel parameters of each model were constrained to be equal among all local models. Considering the hyperparameter h, we found that it is hard to find values of h which perform well across all regions of the input space. Thus, for each input point of interest, we chose values of h that ensured that the localizer considers at least m neighboring training points. In this experiment, we used 3-fold cross-validation to choose the noise variance σ^2, the lengthscale of the GP kernel and the parameter m of the localizing kernel. We report the results on the held-out test set.

In this experiment, we also used the Hilbert localizing kernel (Belkin et al., 2019; Devroye et al., 1998; Shepard, 1968)

$$k(x) = \|x\|^{-1}\mathbb{I}(\|x\| \leq 1), \tag{4.13}$$

which showed good performance for most of the datasets. In Table 4.2 locally smoothed Gaussian Process Regression based on Hilbert kernel is referred to as LSGPR Hilbert, while the same model based on Epanechnikov kernel is referred as LSGPR Epanechnikov. GP regression, Deep GP regression and KNN regression are referred to as GP, DeepGP and KNN, respectively.

The results indicate that LSGP offers competitive performance with respect to GPR and Deep GP baselines. The results also clearly show that LSGP offers superior performance to KNN, suggesting that the localization alone is not enough to obtain good performance, and that this works well in combination with the GP model. To make a comparison with the baselines, we used the one-sided Wilcoxon test Wilcoxon (1945). For each method, we measured its performance on 10 data splits, and we used exactly the same splits for testing performance of each method, so we had matched samples of the MSEs. Then we used the test to compare methods in pairs, where the alternative hypothesis was that the MSE of any given method is smaller than the MSE of a competitors. We used confidence level $\alpha = 0.05$. In Table 4.2 the results that are statistically better than the competitors are marked in bold.

4.4 Conclusions

In this chapter we developed a novel framework to localize Gaussian processes. We focused in particular on Gaussian Process Regression, and we derived the GP model after applying the localization operation through the down-weighting of contributions from input points which are far away from a given test point. The form of the localized GP maintains positive definiteness of the covariance, and it allows for considerable speedups compared to standard global GPR due to the sparsification effect of the Gram matrix.

The proposed method requires cross-validation to tune the scale parameter of localizing kernel, while others GP-based techniques use a less expensive marginal log-likelihood (MLL) gradient optimization to tune these types of parameters. We found MLL gradient optimization problematic because of the discontinuity of local kernel with respect to the scale parameter, which in turn makes MLL function discontinuous with respect to this parameter. It would be interesting to investigate ways to extend the idea of localization for GPR to other tasks, such as classification.

Chapter 5

Optimization of a Regression Function in a passive design

5.1 Introduction

Estimating the minimum value and the minimizer of an unknown function from observation of its noisy values on a finite set of points is a key problem in many applications. Let $D = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ be a design set and let Θ be a compact and convex subset of $\mathbb{R}^d$. Assume that we observe noisy values of an unknown regression function $f : \mathbb{R}^d \to \mathbb{R}$ at points of the design set:

$$y_i = f(x_i) + \xi_i, \quad i = 1, \dots, n, \tag{5.1}$$

where ξ_i's are independent zero mean errors with $\mathbf{E}[\xi_i^2] \leq \sigma^2$. Our goal is to estimate the minimum value of the regression function $f^* = \min_{x \in \Theta} f(x)$ and its location $x^* = \arg\min_{x \in \Theta} f(x)$ when x^* is unique. As accuracy measures of an estimator $\hat{x}_n$ of x^* we consider the expected optimization error $\mathbf{E}(f(\hat{x}_n) - f^*)$ and the quadratic risk $\mathbf{E} \left\| \hat{x}_n - x^* \right\|^2$, where $\| \cdot \|$ denotes the Euclidean norm. The accuracy of an estimator T_n of f^* will be measured by the risk $\mathbf{E}(T_n - f^*)^2$. We will assume that f belongs to the class of β-Hölder smooth and strongly convex functions with $\beta \geq 2$.

The existing literature considers two different assumptions on the choice of the design. Under the *passive* design setting, the points x_i are sampled independently from some probability distribution. Under the *active* (or sequential) design setting, for each i the statistician can plan the experiment by selecting the point x_i depending on the previous queries and the corresponding responses $x_1, y_1, \dots, x_{i-1}, y_{i-1}$. The accuracy of estimation under the active design is at least as good as under the passive design but it can be strictly better, which is the case for the problems considered here.

Active design, estimation of x^*. Active (or sequential) scheme has a long history starting at least from the seminal work of Kiefer and Wolfowitz (1952) where an analog of the Robins-Monro algorithm was introduced to estimate the minimizer x^* of a univariate function f. The idea of the Kiefer-Wolfowitz (KW) method is to approximate the derivative of f using first order differences of y_i's and plug this estimator in the gradient algorithm. Kiefer and Wolfowitz (1952) proved convergence in probability of the KW algorithm under some regularity conditions on the regression function. A multivariate extension of the KW algorithm was proposed by Blum (1954). Convergence rates of the KW algorithm for $d = 1$ were investigated in Dupač (1957) proving an upper bound on the quadratic risk of the order $n^{-2/3}$ for $\beta = 3$. By using suitably chosen linear combinations of first order differences to approximate the gradient, Fabian (1967) proved the existence of a method that attains, for odd integers $\beta \geq 3$, the quadratic risk of the order $n^{-(\beta-1)/\beta}$ for functions f with bounded βth partial derivatives. The method of Fabian (1967) uses $(\beta - 1)/2$ evaluations y_i at every step of the algorithm in order to approximate the gradient. Chen (1988) and Polyak and Tsybakov (1990) have established minimax lower bounds for the estimation risk on the class of $\beta-$Hölder smooth and strongly convex functions f, for all $\beta \geq 2$. For the quadratic risk, these bounds are of the order $n^{-(\beta-1)/\beta}$. Polyak and Tsybakov (1990) proposed a new class of methods using smoothing kernels and randomization to approximate the gradient. This constitutes an alternative to the earlier used deterministic schemes derived from finite differences. Polyak and Tsybakov (1990) proved that such randomized methods attain the minimax optimal rate $n^{-(\beta-1)/\beta}$ on the above classes for all $\beta \geq 2$ and not only for odd integers $\beta \geq 3$. An additional advantage over Fabian's algorithm is the computational simplicity of these methods. In particular, they require at each step only one or two evaluations of the function. For subsequent developments on similar methods, we refer to Akhavan et al. (2020, 2021); Bach and Perchet (2016); Dippon (2003), where one can find further references.

Active design, estimation of f^*. The problem of estimating f^* under the active scheme was first considered by Mokkadem and Pelletier (2007) who suggested a recursive estimator and proved its asymptotic normality with $\sqrt{n}$ scaling. Belitser et al. (2012) defined an estimator of f^* via a multi-stage procedure whose complexity increases exponentially with the dimension d, and showed that this estimator achieves (asymptotically, for n greater than an exponent of d) the $O_p(1/\sqrt{n})$ rate when f is β-Hölder and strongly convex with $\beta > 2$. Akhavan et al. (2020) improved upon this result by constructing a simple computationally feasible estimator $\hat{f}_n$ such that $\mathbf{E}|\hat{f}_n - f^*| = O(1/\sqrt{n})$ for $\beta \geq 2$. It can be easily shown that the rate $1/\sqrt{n}$ cannot be further improved when estimating

f^*. Indeed, using the oracle that puts all the queries at the unknown true minimizer x^* one cannot achieve better rate under the Gaussian noise.

Passive design, estimation of x^*. The problem of estimating the minimizer x^* under the i.i.d. passive design was probably first studied in Härdle and Nixdorf (1987), where some consistency and asymptotic normality results were discussed. Tsybakov (1990b) proposed to estimate x^* by a recursive procedure using local polynomial approximations of the gradient. Considering the class of strongly convex and β-Hölder ($\beta \geq 2$) regression functions f, Tsybakov (1990b) proves that the minimax optimal rate of estimating x^* on the above class of functions is $n^{-(\beta-1)/(2\beta+d)}$, and shows that the proposed estimator attains this optimal rate. However, in order to define this estimator, one needs to know of the marginal density of the design points that may be inaccessible in practice.

There was also some work on estimating x^* in different passive design settings. Several papers are analyzing estimation of x^* in a passive scheme, where x_i's are given non-random points in $[0,1]$ (Müller (1985, 1989)) or in $[0,1]^d$ (Facer and Müller (2003)). Another line of work (Härdle and Nixdorf (1987); Nazin et al. (1989, 1992)) considers the problem of estimating the zero of a nonparametric regression function under i.i.d. design, also called passive stochastic approximation when recursive algorithms are used. Nazin et al. (1989, 1992) establish minimax optimal rates for this problem and propose passive stochastic approximation algorithms attaining these rates. Application to transfer learning is recently developed in Krishnamurthy and Yin (2022), where one can find further references on passive stochastic approximation.

Passive design, estimation of f^*. To the best of our knowledge, the problem of estimating f^* under i.i.d. passive design was not studied. However, there was some work on a related and technically slightly easier problem of estimating the maximum of a function observed under the Gaussian white noise model in dimension $d = 1$ (Ibragimov and Khas'minskii (1982); Lepski (1993)). Extrapolating these results to the regression model and general d suggests that the optimal rate of convergence for estimating f^* on the class of β-Hölder regression functions f is of the order $(n/\log n)^{-\beta/(2\beta+d)}$, cf. the conjecture stated in Belitser et al. (2021)[1]. We are not aware of any results on estimation of f^* on the class of β-Hölder and strongly convex regression functions f, which is the main object of study in the current work.

Finally, we review some results on a related problem of estimating the mode of a probability density function. There exists an extensive literature on this problem. In the univariate case, Parzen (1962a) proposed the maximizer of kernel density estimator

[1] The upper bound with the rate $(n/\log n)^{-\beta/(2\beta+d)}$ is straightforward (cf. Belitser et al. (2021)). The lower bound, though not explicitly reported in the literature, can be routinely obtained using the same ideas as in Ibragimov and Khas'minskii (1982).

(KDE) as an estimator for the mode. Direct estimate of the mode based on order statistics was proposed by Grenander (1965), where the consistency of the proposed method was shown. Other estimators of the mode in the univariate case were considered by (Chernoff, 1964; Dalenius, 1965; Venter, 1967). The minimax rate of mode estimation on the class of β-Hölder densities that are strongly concave near the maximum was shown to be $n^{-(\beta-1)/(2\beta+d)}$ in Tsybakov (1990a), where the optimal recursive algorithm was introduced. It generalizes an earlier result of Khas'minskii (1979) who considered the special case $d = 1, \beta = 2$ and derived the minimax lower bound of the order $n^{-1/5}$ matching the upper rate provided by Parzen (1962a). Klemelä (2005) proposed to use the maximizer of KDE with the smoothing parameter chosen by the Lepski method (Lepskii, 1991), and showed that this estimator achieves optimal adaptive rate of convergence. Dasgupta and Kpotufe (2014) proposed minimax optimal estimators of the mode based on k-nearest neighbor density estimators, emphasizing the implementation ease of the method. Computational complexity of mode estimation was investigated by Arias-Castro et al. (2022) showing the impossibility of a minimax optimal algorithm with sublinear computational complexity. It was shown that the maximum of a histogram, with a proper choice of bandwidth, achieves the minimax rate while running in linear time. Bayesian approach to the mode estimation was developed by Yoo and Ghosal (2019).

Contributions. In this chapter, we consider the model described at the beginning of this section under the passive observation scheme. We assume that f belongs to the class of β-Hölder and strongly convex regression functions. The contributions of the present work can be summarized as follows.

- We construct a recursive estimator of the minimizer x^* adaptive to the unknown marginal density of x_i's and achieving the minimax optimal rate $n^{-(\beta-1)/(2\beta+d)}$ up to a logarithmic factor.
- We show that the minimax optimal rate for the problem of estimating the minimum value f^* of function f on the above class of functions is of the order $n^{-\beta/(2\beta+d)}$, and we propose an algorithm achieving this optimal rate. Thus, the additional strong convexity assumption only allows for a logarithmic improvement in the rate compared to the optimal rate $(n/\log n)^{-\beta/(2\beta+d)}$ on the class of β-Hölder functions without strong convexity (see the discussion above).

Given our results, we have the following table summarizing the minimax optimal rates for estimation under the active and passive design.

	rate of quadratic risk, estimation of x^*	rate of estimating f^*
`passive scheme`	$n^{-\frac{2(\beta-1)}{2\beta+d}}$	$n^{-\frac{\beta}{2\beta+d}}$
`active scheme`	$n^{-\frac{\beta-1}{\beta}}$	$n^{-\frac{1}{2}}$

Table 5.1: Comparisons between the rates of convergence for passive and active schemes

5.2 Definitions and assumptions

One of the main purposes of this work is to investigate the performance of the proposed algorithms over a family of functions that enjoy a higher order smoothness assumption. In the following definition, we characterize such a class of functions.

We first introduce the class of β-Hölder functions that will be used throughout the paper. For $\beta, L > 0$, by $\mathcal{F}_\beta(L)$ we denote the class of $\ell = \lfloor \beta \rfloor$ times differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying the the following inequality

$$\left| f(x) - \sum_{|\boldsymbol{m}| \leq l} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(x)(x - x')^{\boldsymbol{m}} \right| \leq L\|x - x'\|^\beta, \quad \forall x, x' \in \mathbb{R}^d.$$

Our estimators will be based on kernels satisfying the following assumption.

Assumption 7. The kernel $k : \mathbb{R}^d \to \mathbb{R}$ has a compact support Supp(k) contained in the unit Euclidean ball, and satisfies the following conditions

$$k(u) \geq 0, \qquad \int k(u)\,\mathrm{d}u = 1, \qquad \sup_{u \in \mathbb{R}^d} k(u) < \infty \ .$$

Furthermore, for special requirements of our analysis, we assume that k is a L_k-Lipschitz function, i.e. for any $x, y \in \mathbb{R}^d$, we have

$$|k(x) - k(y)| \leq L_k \|x - y\| \ .$$

Assumption 8. It holds for all $i, i' \in [n]$, that: (i) ξ_i and $x_{i'}$ are independent; (ii) $\mathbf{E}[\xi_i] = 0$; (iii) $\mathbf{E}[\xi_i^2] \leq \sigma^2$, where $\sigma^2 \geq 0$.

Assumption 9. We consider the model (5.1) with $f : \mathbb{R}^d \to \mathbb{R}$ satisfying the following assumptions

(i) The function f attains its minimum at $x^* \in \Theta$.

(ii) The function f belongs to Hölder functional class $\mathcal{F}_\beta(L)$ with $\beta \geq 2$.

(iii) There exists $\alpha > 0$ such that the function f is α-strongly convex on Θ i.e. for any $x, y \in \Theta$, it satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \left\| x - y \right\|^2 \quad .$$

(iv) The function f is uniformly bounded on the set $\Theta' = \{x + y : x \in \Theta \quad \text{and} \quad \|y\| \leq 1\}$ such that $\sup_{x \in \Theta'} f(x) \leq M$.

By $\mathcal{F}_{\beta,\alpha}(L)$ we denote the class of regression functions f satisfying Assumption 9.

Next, we introduce our assumptions on the design distribution.

Assumption 10. The design distribution ρ_X is absolutely continuous with respect to the Lebesgue measure with a density $p(x)$ such that

$$0 < p_{\min} \leq p(x) \leq p_{\max} < \infty, \quad \forall x \in \Theta'.$$

Throughout this chapter, we call $\mathtt{A} > 0$ a numerical constant, if $\mathtt{A}$ can only depend on d, Θ, β, L, M, $p_{\max}$, $p_{\min}$, K, and σ, where the dependence on d is at most of a polynomial order with the degree of polynomial only depending on β. We note that dependence on the strong convexity parameter α is not included in the numerical constant since we specify it explicitly in our upper bounds.

5.3 Estimating the location of the minimum

The form of gradient $g_{j,\lambda}$ in Algorithm 5.3 explained by the local polynomial method.

```
ALGORITHM 5.3.1

Requires   Kernel $k$, and parameters $h_j = \left(\frac{\log(j)}{j}\right)^{\frac{1}{2\beta+d}}$ and $\lambda_j = \left(\frac{\log(j)}{j}\right)^{\frac{\beta}{2\beta+d}}$,
for $j \in [n]$.
Initialization   Choose $z_1 \in \Theta$, and assign $\eta_j = \frac{1}{\alpha j}$, for $j \in [n]$.
For $j \in [n]$
    1.  Let $g_{j,\lambda}(z_j) = h_j^{-1}\left(AB_{j,\lambda}^{-1}(z_j)D_j(z_j)\right)$ ,
    2.  Update $z_{j+1} = \mathrm{Proj}_{\Theta}\left(z_j - \eta_j g_{j,\lambda}(z_j)\right)$ .
```

where x is sufficiently close to z and

$$U(u) = \left(\frac{u^{\boldsymbol{m}^{(1)}}}{\boldsymbol{m}^{(1)}}, \ldots, \frac{u^{\boldsymbol{m}^{(S)}}}{\boldsymbol{m}^{(S)}}\right)^{\top}, \quad \theta(z) \quad = \left(h^{|\boldsymbol{m}^{(1)}|} D^{\boldsymbol{m}^{(1)}} f(z), \ldots, h^{|\boldsymbol{m}^{(S)}|} D^{\boldsymbol{m}^{(S)}} f(z)\right)^{\top} \quad .$$

The exposition in section 1.9 suggests that an estimator for $\nabla f(z)$ is

$$g_j(z) = \frac{1}{h}\left(D^1 U(0)^\top\right)\hat{\theta}_n(z) = \frac{1}{h}A\hat{\theta}_n(z) \ , \tag{5.2}$$

where we introduced $A = \left(D^1 U(0)^\top\right)$, i.e.

$$A_{i,m} = \begin{cases} 1, & \text{if } m = i+1 \\ 0, & \text{otherwise} \end{cases}$$

for $i \in [d]$, and $m \in [S]$. Our "statistical" version of the projected gradient descent algorithm 5.3 is based on estimate (5.2) with two important modifications. First, to avoid invertibility issues of the matrix $B_n(z)$ we add the regularization constant $\lambda > 0$ to the diagonal entries. Second, at round j of Algorithm 5.3, we consider the matrix $B_{j,\lambda} = B_j + \lambda I$ and the vector D_j, computed based on the first j observations (ordered in an arbitrary way). This leads us to the regularized estimator of the gradient

$$g_{j,\lambda}(z) = \frac{1}{h}A\hat{\theta}_{j,\lambda}(z) := \frac{1}{h}A(B_j(z) + \lambda I)^{-1}D_j(z) \ . \tag{5.3}$$

The following lemma provides finite sample guarantees for the *bias* of the regularized gradient estimate (5.3).

Lemma 5.1. *For $j \in [n]$, let $g_{j,\lambda}$ be defined by (5.3). Under Assumptions 8, 9, 10 and if the bandwidth and regularizer are chosen to be*

$$h = h_j = j^{-\frac{1}{2\beta+d}}, \quad \lambda = \lambda_j = j^{-\frac{\beta}{2\beta+d}},$$

the following upper bound holds for any $x \in \Theta$

$$\mathbf{E}\left\| g_{j,\lambda}(x) - \nabla f(x) \right\| \le \mathtt{A} j^{-\frac{\beta-1}{2\beta+d}} \ , \tag{5.4}$$

where $\mathtt{A} > 0$ is a numerical constant.

Unregularized version of the lemma above is well know in the literature (see for instance Stone (1982)). The rate $j^{-(\beta-1)/(2\beta+d)}$ know to be optimal in the minimax sense and holds without strong convexity assumption. Note that the bound (5.4) is non-asymptotic and holds for the stochastic design.

Theorem 5.2. *Assume that f satisfies Assumptions 8–10. Then, for z_n that is generated by Algorithm 5.3, we have*

$$\mathbf{E}\left\| z_n - x^* \right\|^2 \le \mathtt{A} \min\left(1, \left(\frac{\log(n)}{n}\right)^{\frac{2(\beta-1)}{2\beta+d}} \alpha^{-2}\right) \ , \tag{5.5}$$

where $\mathtt{A} > 0$ *is a numerical constants.*

Proof. We use the definition of Algorithm 5.3 and strong convexity of f to obtain an upper bound for $\mathbf{E}[\|z_{j+1} - x^*\| | z_j]$, which depends on the bias term $\|\mathbf{E}[g_j(z_j)|z_j] - \nabla f(z_j)\|$ and on the stochastic error term $\mathbf{E}\left[\|g_j(z_j)\|^2 | z_j\right]$. Lemmas 6 and 7 bounds the bias and stochastic error terms uniformly (this is why the bound includes an additional log factor) over Θ leading to the bound claimed in the theorem. □

We consider the logarithmic factor appearing in (5.5) as a price to pay for the fact that our algorithm is adaptive to the marginal density of x_i's. Indeed, Tsybakov (1990b) considered estimators that can depend on the marginal density of x_i's and achieve the optimal rate $n^{-(\beta-1)/(2\beta+d)}$, while Algorithm 5.3 is free of such dependence. Note also that our algorithm can be realized in online mode with the data that arrive progressively. We conjecture that the extra logarithmic factor can be eliminated if we estimate x^* by the minimizer of the local polynomial estimator of f. However, such a method needs the whole sample and cannot be realized in online mode. It remains an open question whether there exists an algorithm combining all the three advantages, that is, online realization, adaptivity to the marginal density and convergence with the sharp optimal rate $n^{-(\beta-1)/(2\beta+d)}$.

The rate in the above theorem leads us to the conclusion that the estimation of the minimizer of the strongly convex and smooth function is as hard as the estimation of the gradient at a *fixed* point. Notice that the course of dimensionality can not be resolved in the optimization problems for the passive design which distinguishes it from the active design optimization where the optimal rate $n^{-(\beta-1)/(2\beta)}$ is independent of the ambient dimension d (see Akhavan et al. (2020); Novitskii and Gasnikov (2021)). Note also that the slow rate is inherited from the gradient estimation part (5.3) and is not related to the projected gradient descent itself.

5.4 Estimating a minimum value of the regression function

In this section, we apply the above results to estimation of the minimum value $f^* = \min_{x\in\Theta} f(x)$ for functions f in the class $\mathcal{F}_{\beta,\alpha}(L)$. Observe that $f(x^*)$ is not an estimator since it depends on the unknown f, so Theorem 5.2 does not provide a result about estimation of f^*. The estimation of f^* proceeds by estimating the minimizer and the value of the function separately on the equally and randomly divided data. Function estimation step can be done by *any* optimal algorithm. Here we adopt the framework of the local regression already used to estimate the gradient in the Algorithm 5.3. *Regularized local polynomial estimator* of the function f at point z is defined as

$$f_n(z) = U^\top(0)\hat{\theta}_{n,\lambda}(z). \tag{5.6}$$

Minimum value estimation by local polynomial estimator is outlined in Algorithm 5.4.

```
ALGORITHM 5.4.1
Requires: Algorithm 5.3, kernel $k : [-1,1]^d \to \mathbb{R}$, parameters $h_n = n^{-\frac{1}{2\beta+d}}$ and $\lambda_n = n^{-\frac{\beta}{2\beta+d}}$.
1. Randomly split the data $D$ in two equal parts $D_1$ and $D_2$
2. On the subsample $D_1$ construct the minimizer $z_n$ by the Algorithm 5.3:
   $z_n \leftarrow$ Algorithm 5.3$(D_1)$
3. On the second subsample $D_2$, construct an estimator $f_n(z_n) = U^\top(0)\hat{\theta}_{n,\lambda}(z_n)$ at the point $z_n$
```

Local polynomial estimator is known to be optimal (Stone, 1982), however, to the best of our knowledge, all the rates considered in the literature are asymptotic and hold only for sufficiently big n. Below we provide an upper rate for (5.6) which is non-asymptotic.

Theorem 5.3. *Under Assumptions 8, 9, and 10, for any $x \in \Theta$, we have*

$$\mathbf{E}\left[(f_n(x) - f(x))^2\right] \leq \left(B_{bias}^2 + B_{var}\right) n^{-\frac{2\beta}{2\beta+d}} \ .$$

Note that this theorem may be of independent interest since, to the best of our knowledge, the non-asymptotic rates of convergence of a regularized local polynomial estimator have not been studied in the literature.

The following theorem gives the rate of convergence for the estimator $T_n = f_n(z_n)$.

Theorem 5.4. *Assume that f satisfies Assumptions 8, 9, and 10. Then, we have*

$$\mathbf{E}\left|T_n - f\left(x^*\right)\right| \leq \begin{cases} C_1(\log(n)/n)^{\frac{2}{4+d}} & \text{if } \beta = 2 \\ C_2 n^{-\frac{\beta}{2\beta+d}} & \text{if } \beta > 2 \end{cases} \tag{5.7}$$

where $C_1, C_2 > 0$ are numerical constants.

Proof. Using the fact that, for any fixed x the estimator $f_n(x)$ is measurable with respect to the second half of the sample and z_n is measurable with respect to its first half we get

$$\begin{aligned}\mathbf{E}\left[|T_n - f(x^*)|\right] &\leq \mathbf{E}\left(\mathbf{E}\left[|f_n(z_n) - f(z_n)|\big|z_n\right]\right) + \mathbf{E}\left[|f(z_n) - f(x^*)|\right] \\ &\leq \mathbf{E}\left[\left(\mathbf{E}\left[(f_n(z_n) - f(z_n))^2\,|z_n\right]\right)^{\frac{1}{2}}\right] + L^{'}\mathbf{E}\left[\|z_n - x^*\|^2\right] \\ &\leq \left(\mathbf{E}\left(f_n(x) - f(x)\right)^2\right)^{\frac{1}{2}} + L^{'}\mathbf{E}\left[\|z_n - x^*\|^2\right].\end{aligned}$$

By Theorems 5.2 and 5.3 we deduce

$$\begin{aligned}\mathbf{E}\left[|T_n - f(x^*)|\right] &\leq \max\left(1, \alpha^{-1}\right)\left(\mathtt{C}_3 n^{-\frac{\beta}{2\beta+d}} + \mathtt{C}_4\left(\frac{\log(n)}{n}\right)^{\frac{2(\beta-1)}{2\beta+d}}\right) \\ &\leq \begin{cases} C_1(\log(n)/n)^{\frac{2}{4+d}} & \text{if } \beta = 2 \\ C_2 n^{-\frac{\beta}{2\beta+d}} & \text{if } \beta > 2 \end{cases},\end{aligned}$$

where $\mathtt{C}_1, \mathtt{C}_2, \mathtt{C}_3, \mathtt{C}_4 > 0$ are numerical constants. □

Theorem 5.4 shows that estimation of f^* for smooth and strongly convex functions under passive design is realized with the same rate as function estimation. The lower bound (5.10) below shows that the slow rate in (5.7) cannot be improved in a minimax sense and it corresponds to the rate of a smooth function estimation at a *fixed* point. We show below that the rate $(n/\log n)^{-\beta/(2\beta+d)}$ is optimal for β-smooth regression functions without strong convexity assumption. It corresponds to the rates of function estimation in supremum norm. The strong convexity assumption allows us to reduce the global function reconstruction problem to a simpler, point estimation, leading to the rates without extra logarithmic factor. Note that the rate $n^{-\beta/(2\beta+d)}$ cannot be improved even when x^* is known as the function estimation at the point of minimum is still required.

Note that, for $\beta > 2$, the convergence rate of Algorithm 5.3 used at the first stage to estimate the minimizer is more than needed to achieve the rate (5.7). The optimal estimate of f^* can be obtained by estimating the minimizer at a slower rate, namely, $n^{-\beta/(2\beta+d)}$ for the optimization risk. Therefore, it is not necessary to have z_n as an estimator at the first step - it can be replaced by some suboptimal estimators. This could be beneficial considering the fact that suboptimal algorithms may be computationally less costly.

In an active design setting, much faster rate can be obtained, see Table 5.1. Specifically, f^* can be estimated with the parametric rate $Cn^{-1/2}$ where $C > 0$ is a constant, which is independent of the dimension d and smoothness β for any $\beta > 2$ and all n large enough

(Akhavan et al., 2020). Clearly, the rate $n^{-1/2}$ cannot be improved even by using the ideal but non-realizable oracle that makes all queries at point x^*.

5.5 Lower bound

The following theorem provides lower bounds for the minimax risks of arbitrary estimators on the class $\mathcal{F}_{\beta,\alpha}(L)$. Let $w(\cdot)$ be a monotone non-decreasing function on $[0,\infty)$ such that $w(0) = 0$ and $w \not\equiv 0$.

Theorem 5.5. *Let $x_1, \dots, x_n$ be i.i.d. random vectors with a bounded Lebesgue density on $\mathbb{R}^d$. Assume that the random variables ξ_i are i.i.d. having a density $p_\xi(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}$ such that*

$$\exists I_* > 0, v_0 > 0 : \quad \int \left(\sqrt{p_\xi(u)} - \sqrt{p_\xi(u+v)}\right)^2 du \le I_* v^2 \ , \tag{5.8}$$

for $|v| \le v_0$. Then, for any $\beta, \alpha, L > 0$ we have

$$\inf_{x_n} \sup_{f \in \mathcal{F}_{\beta,\alpha}(L)} \mathbf{E}_f w(n^{\frac{\beta-1}{2\beta+d}} \|x_n - x^*\|) \ge c_1, \tag{5.9}$$

and

$$\inf_{f_n} \sup_{f \in \mathcal{F}_{\beta,\alpha}(L)} \mathbf{E}_f w(n^{\frac{\beta}{2\beta+d}} |f_n - f^*|) \ge c_1', \tag{5.10}$$

where $\inf_{x_n}$ and $\inf_{f_n}$ denote the infimum over all estimators of the minimizer and over all estimators of the minimum value of f, respectively, and $c_1 > 0, c_1' > 0$ are constants that depend only on $\beta, \alpha, L, \Theta, I_*, v_0$, and $w(\cdot)$.

Condition (5.8) is rather general. It is satisfied, for example, for the Gaussian distribution and also for a large class of regular densities, cf. Ibragimov and Has'minskii (1981). The lower bound (5.9) was proved in Tsybakov (1990b) under a more restrictive condition on the density p_ξ.

The proof of Theorem 5.5 is given in Section 5.8. It is based on a reduction to the problem of testing two hypotheses.

Considering the bounds (5.10), (5.9) with $w(u) = u^2$ and $w(u) = u$, respectively, and combining them with Theorems 5.2 and 5.4 we obtain that the estimator T_n is minimax optimal for f^*, and z_n is minimax optimal up to a logarithmic factor for x^* on the class of functions $\mathcal{F}_{\beta,\alpha}(L)$.

In the next theorem, we provide a minimax lower bound on estimation of f^* over the class of β-Hölder functions $\mathcal{F}_\beta(L)$ when there is no strong convexity assumption.

Theorem 5.6. *Let $x_1, \dots, x_n$ be i.i.d. random vectors with a bounded Lebesgue density on $\mathbb{R}^d$, and let ξ_i be i.i.d. Gaussian random variables with zero mean and variance σ^2. Assume that Θ contains an open subset of $\mathbb{R}^d$. Then, for any $\beta > 0, L > 0$, we have*

$$\inf_{f_n} \sup_{f \in \mathcal{F}_\beta(L)} \mathbf{E}_f w \left(\left(\frac{n}{\log n} \right)^{\frac{\beta}{2\beta+d}} |f_n - f^*| \right) \geq c_3, \tag{5.11}$$

where $\inf_{f_n}$ denotes the infimum over all estimators of the minimum value of f and $c_3 > 0$ is a constant that depends only on $\beta, \alpha, L, \Theta, \sigma^2$, and $w(\cdot)$.

Theorem 5.6 implies that $\left(\frac{n}{\log n}\right)^{-\frac{\beta}{2\beta+d}}$ is the minimax rate of estimating the minimum value f^* on the class $\mathcal{F}_\beta(L)$. Indeed, the matching upper bound with the rate $\left(\frac{n}{\log n}\right)^{-\frac{\beta}{2\beta+d}}$ is obtained in a trivial way if we estimate f^* by the minimum of any rate optimal (in supremum norm) nonparametric estimator of f, for example, by the local polynomial estimator as in Stone (1982).

Thus, if we drop the assumption of strong convexity, the minimax rate deteriorates only by a logarithmic factor. It suggests that strong convexity is not a crucial advantage in estimation of the minimum value of a function under the passive design.

5.6 Discussion

In this chapter, we have considered the problem of estimating the minimizer and the minimum value of the regression function from the i.i.d data with a special focus on highly smooth and strongly convex regression functions. We provide upper bounds for the proposed algorithms. We show that the rates of estimation of the minimizer is the same as the rate for estimating the gradient of the regression function. To estimate the minimum value we have used two-stage procedure where in the first step we estimate the location of the minimum followed by the estimation of the function value at the estimated in the first step point. We obtain optimal nonparametric rates of convergence for our two-stage procedure.

An interesting open question is to make our algorithms adaptive to the unknown smoothness β, that is, to develop a data-driven choice of the smoothing parameter h and of the regularization parameter λ. When considering adaptation to the unknown smoothness of function f, the optimal rates for estimation of f^* will be presumably slower than the minimax rates by a logarithmic factor.

5.7 Proofs

In this section, we provide the proofs of our two sets of upper bounds on estimator of the minimizer and minimum value. Section 5.7.1 is devoted to the proof of Theorem 5.2 on upper bounds of the Algorithm 5.3, whereas Section 5.7.2 is devoted to the proofs of Theorem 5.3. In section 5.8 we provide the proof of the lower bound (theorem 5.5).

5.7.1 Proof of theorem 5.2

For the proof of theorem 5.2 we need some preliminary lemmas.

Lemma 5.7. *For $j \in [n]$, let $g_{j,\lambda}$ be defined by (5.3). Under Assumptions 8, 9, 10 and if the bandwidth and regularizer are chosen to be*

$$h = h_j = \left(\frac{\log(j)}{j}\right)^{\frac{1}{2\beta+d}} \quad \textit{and} \quad \lambda = \lambda_j = \left(\frac{\log(j)}{j}\right)^{\frac{\beta}{2\beta+d}} ,$$

the following upper bound holds for any $x \in \Theta$

$$\mathbf{E} \left\| g_{j,\lambda}(x) - \nabla f(x) \right\| \leq \mathtt{A} \left(\frac{1}{j}\right)^{\frac{\beta-1}{2\beta+d}} \quad , \tag{5.12}$$

where $\mathtt{A} > 0$ is a numerical constant.

Proof. We introduce the shorthand notations. For any $j \in [n]$, and $i \in [j]$, let

$$M_{i,j}(x) = U\left(\frac{x_i - x}{h_j}\right) U^\top \left(\frac{x_i - x}{h_j}\right) k\left(\frac{x_i - x}{h_j}\right) \quad \text{and} \quad R_{i,j}(x) = U\left(\frac{x_i - x}{h_j}\right) k\left(\frac{x_i - x}{h_j}\right) \quad . \tag{5.13}$$

Also, denote $C_j(x) = \frac{1}{jh_j^d} \sum_{i=1}^{j} R_{i,j}(x) f(x_i)$, $D_j = \frac{1}{kh_j^d} \sum_{i=1}^{j} R_{i,j}(x) y_i$, and note that $\mathbf{E}\left[C_j(x)\right] = \mathbf{E}\left[D_j(x)\right]$.

First, for the sake of simplicity, denote $B = \mathbf{E}[B_j(x)]$. By letting $\phi_j = g_j(x) - h_j^{-1}\left(AB^{-1}B_j(x)\boldsymbol{c}_j(f,x)\right)$, we can write

$$\mathbf{E}\left[\phi_j\right] = \mathbf{E}\left[g_j(x)\right] - h_j^{-1} A\boldsymbol{c}_j(f,x) = \mathbf{E}\left[g_j(x)\right] - \nabla f(x) \ \ ,$$

where $\boldsymbol{c}_j(f,x) = \left(h_j^{|\boldsymbol{m}^{(1)}|} D^{\boldsymbol{m}^{(1)}} f(x), \ldots, h_j^{|\boldsymbol{m}^{(S)}|} D^{\boldsymbol{m}^{(S)}} f(x)\right)^\top$. Also, note that by Assumption 8, $\mathbf{E}[g_j(x)] = \mathbf{E}\left[h_j^{-1}\left(AB_{j,\lambda}^{-1}(x) C_j(x)\right)\right]$. To conclude the proof, we need to provide an upper bound for the term $\|\mathbf{E}\left[\phi_j\right]\|$. Let

$$\begin{aligned}\psi_{1,j} &= h_j^{-1}\left(AB^{-1}C_j(x)\right) \quad ,\\ \psi_{2,j} &= h_j^{-1}\left(A(B+\lambda_j I)^{-1}C_j(x)\right) \quad .\end{aligned}$$

Then we have

$$\begin{aligned}\|\mathbf{E}[\phi_j]\| &= \left\|\mathbf{E}\left[\left(\psi_{1,j} - h_j^{-1}\left(AB^{-1}B_j(x)\boldsymbol{c}_j(f,x)\right)\right) + (\psi_{2,j}-\psi_{1,j}) + (g_j(x)-\psi_{2,j})\right]\right\|\\ &\le \underbrace{\left\|\mathbf{E}[\psi_{1,j} - h_j^{-1}\left(AB^{-1}B_j(x)\boldsymbol{c}_j(f,x)\right)]\right\|}_{\text{term I}} + \underbrace{\|\mathbf{E}[\psi_{2,j}-\psi_{1,j}]\|}_{\text{term II}} + \underbrace{\|\mathbf{E}[g_j(x)-\psi_{2,j}]\|}_{\text{term III}} \quad .\end{aligned}$$

We provide adequately tight upper bounds, for each term in the above, separately. For term I, we can write

$$\begin{aligned}\text{term I} &= h_j^{-1}\left\|AB^{-1}\mathbf{E}\left[\frac{1}{jh_j^d}\sum_{i=1}^{j} R_{i,j}(x)\left(f(x_i) - U^\top\left(\frac{x_i-x}{h_j}\right)\boldsymbol{c}_j(f,x)\right)\right]\right\|\\ &\le h_j^{-1}\left\|AB^{-1}\right\|_{\text{op}}\left\|\mathbf{E}\left[\frac{1}{kh_j^d}\sum_{i=1}^{j} R_{i,j}(x)\left(f(x_i) - U^\top\left(\frac{x_i-x}{h_j}\right)\boldsymbol{c}_j(f,x)\right)\right]\right\| \quad .\end{aligned}$$

Since $\|A\|_{\text{op}} \le 1$, by Lemma 5.14(iii), we deduce that $\left\|AB^{-1}\right\|_{\text{op}} \le \mu_{\min}^{-1}$. Then we can write

$$\text{term I} \le h_j^{-1}\mu_{\min}^{-1}\left(\frac{1}{jh_j^d}\sum_{i=1}^{j}\mathbf{E}\left[\left\|R_{i,j}(x)\left(f(x_i) - U^\top\left(\frac{x_i-x}{h_j}\right)\boldsymbol{c}_j(f,x)\right)\right\|\right]\right) \quad .$$

Since by Assumption 9 $f \in \mathcal{F}_\beta(L)$, for any $i \in [j]$, we have

$$|f(x_i) - U^\top\left(\frac{x_i-x}{h_j}\right)\boldsymbol{c}_j(f,x)| \le L\left\|x-x_i\right\|^\beta \quad ,$$

and we can write

$$\begin{aligned}\text{term I} &\le Lh_j^{-1}\mu_{\min}^{-1}\left(\frac{1}{jh_j^d}\sum_{i=1}^{j}\mathbf{E}\left[\|R_{i,j}(x)\|\left\|x-x_i\right\|^\beta\right]\right)\\ &= Lh_j^{-d-1}\mu_{\min}^{-1}\int_{\mathbb{R}^d}\|x-u\|^\beta\left\|U\left(\frac{u-x}{h_j}\right)k\left(\frac{u-x}{h_j}\right)\right\|p(u)\,\mathrm{d}u\\ &= Lh_j^{\beta-1}\mu_{\min}^{-1}\int_{\mathbb{R}^d}\|w\|^\beta\left\|U(w)k(w)\right\|p(x+h_jw)dw \le \mathtt{A}_1 h_j^{\beta-1} \quad ,\end{aligned}$$

where we introduced $\mathtt{A}_1 = L\mu_{\min}^{-1}p_{\max}\kappa_\beta$, and $\kappa_\beta = \int_{\mathbb{R}^d} \|u\|^\beta \|U(u)k(u)\| \, \mathrm{d}u$. For term II, we deduce that

$$\begin{aligned}\text{term II} &= h_j^{-1}\left\|A\left((B+\lambda_j I)^{-1} - B^{-1}\right)\mathbf{E}\left[C_j(x)\right]\right\| \\ &\leq Mh_j^{-1}\lambda_j \|A\|_{\mathrm{op}} \left\|B^{-1}\right\|_{\mathrm{op}} \left\|(B+\lambda_j I)^{-1}\right\|_{\mathrm{op}} \mathbf{E}\left[\|C_j(x)\|\right] \ .\end{aligned}$$

By Assumption 9(iii), we have $\sup_{x\in\Theta'} f(x) \leq M$. Also, $\mathbf{E}\left[\|C_j(x)\|\right] \leq \mathbf{E}\left[\sup_{x\in\Theta}\|C_j(x)\|\right]$, where by Lemma 5.14(ii) we get $\mathbf{E}\left[\sup_{x\in\Theta}\|C_j(x)\|\right] \leq Mp_{\max}\nu_{1,1}$. Moreover, by Lemma 5.14(iii), we can write $\left\|B^{-1}\right\|_{\mathrm{op}}\left\|(B+\lambda_j I)^{-1}\right\|_{\mathrm{op}} \leq \mu_{\min}^{-2}$. Therefore, we deduce that

$$\text{term II} \leq \mathtt{A}_2 h_j^{-1}\lambda_j \ ,$$

with $\mathtt{A}_2 = Mp_{\max}\nu_{1,1}\mu_{\min}^{-2}$. Finally, we need to bound term III

$$\begin{aligned}\text{term III} &\leq h_j^{-1}\left\|\mathbf{E}\Big[A\left(B_{j,\lambda}^{-1}(x) - (\mathbf{E}[B_{j,\lambda}(x)])^{-1}\right)(C_j(x) - \mathbf{E}\left[C_j(x)\right])\Big]\right\| \\ &\quad + h_j^{-1}\left\|\mathbf{E}\Big[A\left(B_{j,\lambda}^{-1}(x) - (\mathbf{E}[B_{j,\lambda}(x)])^{-1}\right)\mathbf{E}\left[C_j(x)\right]f(x_j)\Big]\right\| \\ &\leq h_j^{-1}\mathbf{E}\left[\left\|B_{j,\lambda}^{-1}(x) - (\mathbf{E}[B_{j,\lambda}(x)])^{-1}\right\|_{\mathrm{op}}\|C_j(x) - \mathbf{E}\left[C_j(x)\right]\|\right] \\ &\quad + h_j^{-1}\mathbf{E}\left[\left\|B_{j,\lambda}^{-1}(x) - (\mathbf{E}[B_{j,\lambda}(x)])^{-1}\right\|_{\mathrm{op}}\right]\mathbf{E}\left[\sup_{x\in\Theta}\|C_j(x)\|\right] \ .\end{aligned}$$

For the first term in the r.h.s., we use Lemma 5.19, and we get

$$\text{term III} \leq \mathtt{A}_3 j^{-1}h_j^{-d-1} + h_j^{-1}\mathbf{E}\left[\left\|B_{j,\lambda}(x)^{-1} - (\mathbf{E}[B_{j,\lambda}(x)])^{-1}\right\|_{\mathrm{op}}\right]\mathbf{E}\left[\sup_{x\in\Theta}\|C_j(x)\|\right] \ ,$$

where $\mathtt{A}_3 > 0$ is the numerical constant that appears in Lemma 5.19. For the second term on the r.h.s., by invoking Lemma 5.14(i), can be bounded by the following term

$$\begin{aligned}&\mathtt{A}_4 h_j^{-1}\mathbf{E}\Big[\|B_{j,\lambda}(x)\|_{\mathrm{op}}^{-1}\|B_{j,\lambda}(x) - (\mathbf{E}[B_{j,\lambda}(x)])\|_{\mathrm{op}}\Big] \leq \\ &\qquad\qquad \mathtt{A}_4 h_j^{-1}\left(\mathbf{E}\left[\|B_{j,\lambda}(x)\|_{\mathrm{op}}^{-2}\right]\mathbf{E}\left[\|B_{j,\lambda}(x) - (\mathbf{E}[B_{j,\lambda}(x)])\|_{\mathrm{op}}^2\right]\right)^{\frac{1}{2}} \ ,\end{aligned}$$

where for the last display we used Cauchy-Schwarz inequality and we introduced $\mathtt{A}_4 = Mp_{\max}\nu_{1,1}\mu_{\min}^{-1}$. Now by using Jensen's inequality and Lemma 5.16, we get

$$\text{term III} \leq \mathtt{A}_3 j^{-1}h_j^{-d-1} + \mathtt{A}_4 j^{-\frac{1}{2}}h_j^{-\frac{d}{2}-1} \leq \mathtt{A}_5 j^{-\frac{1}{2}}h_j^{-\frac{d}{2}-1},$$

where $\mathtt{A}_5 = 3\mathtt{A}_3 + \mathtt{A}_4$. By combining all of these, we have

$$\|\mathbf{E}[g_j(x)] - \nabla f(x)\| \leq \mathtt{A}_6 \left(h_j^{\beta-1} + h_j^{-1}\lambda_j + h_j^{-1-\frac{d}{2}} j^{-\frac{1}{2}} \right) \quad , \tag{5.14}$$

where we introduce $\mathtt{A}_6 = \max(\mathtt{A}_1, \mathtt{A}_2, \mathtt{A}_5)$. Finally, by substituting $h_j = \left(\frac{\log(j)}{j}\right)^{\frac{1}{2\beta+d}}$, and $\lambda_j = \left(\frac{\log(j)}{j}\right)^{\frac{\beta}{2\beta+d}}$, we deduce

$$\|\mathbf{E}[g_j(x)] - \nabla f(x)\| \leq \mathtt{A}_{\text{bias}} \left(\frac{\log(j)}{j}\right)^{\frac{\beta-1}{2\beta+d}} \quad ,$$

where $\mathtt{A}_{\text{bias}} = 3\mathtt{A}_6$. □

The following lemma provides the bound of the variance uniformly over Θ.

Lemma 5.8. *Let g_j be defined by Algorithm 5.3, and assume that Assumptions 8, 9, and 10 hold. Then, we have*

$$\mathbf{E}\left[\sup_{x\in\Theta} \|g_j(x) - \mathbf{E}\left[g_j(x)\right]\|^2\right] \leq \mathtt{A}_{var} \left(\frac{\log(j)}{j}\right)^{\frac{2(\beta-1)}{2\beta+d}} \quad ,$$

where $\mathtt{A}_{var} > 0$ is a numerical constant.

Proof. Let $G_j(x) = \frac{1}{jh_j^d}\sum_{i=1}^{j} R_{i,j}(x)\xi_i$, and recall that $C_j(x) = \frac{1}{jh_j^d}\sum_{i=1}^{j} R_{i,j}(x) f(x_i)$. Then, we have

$$\begin{aligned}\mathbf{E}\left[\sup_{x\in\Theta} \|g_j(x) - \mathbf{E}\left[g_j(x)\right]\|^2\right] \leq & \underbrace{2\mathbf{E}\left[\sup_{x\in\Theta} \left\| h_j^{-1} A B_{j,\lambda}^{-1}(x) G_j(x) \right\|^2\right]}_{\text{term I}} \\ & + \underbrace{2\mathbf{E}\left[\sup_{x\in\Theta} \left\| h_j^{-1} A B_{j,\lambda}^{-1}(x) C_j(x) - \mathbf{E}\left[h_j^{-1} A B_{j,\lambda}^{-1}(x) C_j(x)\right] \right\|^2\right]}_{\text{term II}} .\end{aligned}$$

For term I, we have

$$\begin{aligned}\text{term I} \leq & \underbrace{4h_j^{-2}\mathbf{E}\left[\sup_{x\in\Theta} \left\| A\left(B_{j,\lambda}^{-1}(x) - (\mathbf{E}[B_{j,\lambda}(x)])^{-1}\right) G_j(x) \right\|^2\right]}_{\text{term III}} \\ & + \underbrace{4h_j^{-2} \left\| (\mathbf{E}[B_{j,\lambda}(x)])^{-1} \right\|_{\text{op}}^2 \mathbf{E}\left[\sup_{x\in\Theta} \|G_j(x)\|^2\right]}_{\text{term IV}} .\end{aligned}$$

For term III, by using the property of Assumption 8, we can write

$$\text{term III} \le 4\sigma^2\mu_{\min}^{-2}\lambda_j^{-2}h_j^{-2}\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)-\mathbf{E}[B_{j,\lambda}(x)]\|_{\mathrm{op}}^2\left(j^{-2}h_j^{-2d}\sum_{i=1}^{j}\|R_{i,j}(x)\|^2\right)\right] \ .$$

Now, by invoking Cauchy-Schwarz inequality we get

$$\text{term III} \le 4\sigma^2\mu_{\min}^{-2}\lambda_j^{-2}h_j^{-2}\left(\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)-\mathbf{E}[B_{j,\lambda}(x)]\|_{\mathrm{op}}^4\right]\mathbf{E}\left[\sup_{x\in\Theta}\left(j^{-2}h_j^{-2d}\sum_{i=1}^{j}\|R_{i,j}(x)\|^2\right)^2\right]\right)^{\frac{1}{2}}$$

By utilizing Lemma 5.15, we get

$$\begin{aligned}\text{term III} &\le 4\sigma^2\mu_{\min}^{-2}\lambda_j^{-2}j^{-1}h_j^{-d-2}\log(j)\left(\mathbf{E}\left[\sup_{x\in\Theta}\left(j^{-2}h_j^{-2d}\sum_{i=1}^{j}\|R_{i,j}(x)\|^2\right)^2\right]\right)^{\frac{1}{2}}\\ &\le 4\sigma^2\mu_{\min}^{-2}\lambda_j^{-2}j^{-1}h_j^{-d-2}\log(j)\Bigg(j^{-4}h_j^{-4d}\Bigg(\sum_{i=1}^{j}\mathbf{E}[\sup_{x\in\Theta}\|R_{i,j}(x)\|^4]\\ &\qquad+\sum_{i,m=1}^{j}\mathbf{E}[\sup_{x\in\Theta}\|R_{i,j}(x)\|^2]\mathbf{E}[\sup_{x\in\Theta}\|R_{m,j}(x)\|^2]\Bigg)\Bigg)^{\frac{1}{2}}\\ &\le 4\sigma^2\mu_{\min}^{-2}\lambda_j^{-2}j^{-1}h_j^{-d-2}\log(j)\left(j^{-3}h_j^{-3d}p_{\max}\nu_{1,4}+j^{-2}h_j^{-2d}p_{\max}^2\nu_{1,2}^2\right)^{\frac{1}{2}} \ ,\end{aligned}$$

where the last inequality is a result of Lemma 5.14(i). Now, by using the inequalities $1\le jh_j^d$ and $j\ge\lambda_j^{-2}h_j^{-d}\log(j)$, we can write

$$\text{term III} \le \mathtt{A}_1 j^{-1}h_j^{-d-2} \ ,$$

where we introduce $\mathtt{A}_1 = 4\sigma^2\mu_{\min}^{-2}(p_{\max}\nu_{1,4}+p_{\max}^2\nu_{1,2}^2)^{\frac{1}{2}}$. For term IV, we have

$$\text{term IV} \le 4\sigma^2\mu_{\min}^{-2}j^{-1}h_j^{-2d-2}\mathbf{E}\left[\sup_{x\in\Theta}\|R_{1,j}\|^2\right] \le \mathtt{A}_2 j^{-1}h_j^{-d-2} \ ,$$

where the last inequality is obtained by Lemma 5.14 (i), with $\mathtt{A}_2 = 4\sigma^2\mu_{\min}^{-2}p_{\max}\nu_{1,2}$. Therefore, we deduce that

$$\text{term I} \le \mathtt{A}_3 j^{-1}h_j^{-d-2} \ ,$$

with $\mathtt{A}_3 = \mathtt{A}_1 + \mathtt{A}_2$. We proceed the proof by providing an adequate tight upper bound for term II.

$$\text{term II} \leq \underbrace{4h_j^{-2}\mathbf{E}\left[\sup_{x\in\Theta}\left\|B_{j,\lambda}^{-1}(x) - (\mathbf{E}[B_{j,\lambda}(x)])^{-1}\right\|_{\text{op}}^2 \|C_j(x)\|^2\right]}_{\text{term V}}$$
$$+ \underbrace{4h_j^{-2}\left\|(\mathbf{E}[B_{j,\lambda}(x)])^{-1}\right\|_{\text{op}}^2 \mathbf{E}\left[\sup_{x\in\Theta}\|C_j(x) - \mathbf{E}[C_j(x)]\|^2\right]}_{\text{term VI}} \enspace .$$

Similar to term III, for term V we have

$$\text{term V} \leq 4M^2\mu_{\min}^{-2}\lambda_j^{-2}j^{-1}h_j^{-d-2}\log(j)\left(j^{-3}h_j^{-3d}p_{\max}\nu_{1,4} + j^{-2}h_j^{-2d}p_{\max}^2\nu_{1,2}^2\right)^{\frac{1}{2}} \leq \mathtt{A}_4 j^{-1}h_j^{-d-2} \enspace ,$$

where $\mathtt{A}_4 = 4M^2\mu_{\min}^{-2}(p_{\max}\nu_{1,4} + p_{\max}^2\nu_{1,2}^2)^{\frac{1}{2}}$. Finlay, for term VI, by Lemma 5.18 we can write

$$\text{term VI} \leq \mathtt{A}_5 j^{-1}h_j^{-d-2}\log(j) \enspace ,$$

where $\mathtt{A}_5 > 0$ is a numerical constant. Thus, we deduce that

$$\text{term II} \leq \mathtt{A}_6 j^{-1}h_j^{-d-2}\log(j) \enspace ,$$

with $\mathtt{A}_6 = \mathtt{A}_4 + \mathtt{A}_5$. We conclude the proof, by letting $\mathtt{A}_{\text{var}} = \mathtt{A}_3 + \mathtt{A}_6$, and substituting the parameters $h_j = \left(\frac{\log(j)}{j}\right)^{\frac{1}{2\beta+d}}$, and $\lambda_j = \left(\frac{\log(j)}{j}\right)^{\frac{\beta}{2\beta+d}}$. □

Lemma 5.9. *Let g_j be defined by Algorithm 5.3, and assume that Assumptions 8, 9, and 10 hold. Then, we have*

$$\mathbf{E}\left[\sup_{x\in\Theta}\|g_j(x) - \nabla f(x)\|^2\right] \leq \mathtt{A}_{error}\left(\frac{\log(j)}{j}\right)^{\frac{2(\beta-1)}{2\beta+d}} \enspace ,$$

where $\mathtt{A}_{error} > 0$ is a numerical constant.

Proof. We can write

$$\mathbf{E}\left[\sup_{x\in\Theta}\|g_j(x) - \nabla f(x)\|^2\right] \leq \mathbf{E}\left[\sup_{x\in\Theta}\|g_j(x) - \mathbf{E}\left[g_j(x)\right]\|^2\right] + \sup_{x\in\Theta}\|\mathbf{E}\left[g_j(x)\right] - \nabla f(x)\|^2 \enspace .$$

We conclude the proof by using Lemmas 5.7 and 5.8, and letting $\mathtt{A}_{\text{error}} = \mathtt{A}_{\text{bias}}^2 + \mathtt{A}_{\text{var}}$. □

Lemma 5.10. *Let g_k be defined by Algorithm 5.3, and assume that Assumptions 8, 9, and 10 hold. Then, we have*

$$\mathbf{E}\left[\sup_{x\in\Theta}\|g_j(x)\|^2\right] \leq \mathtt{A}_{sm} j^{\frac{2+d}{2\beta+d}} \log(j)^{\frac{2(\beta-1)}{2\beta+d}} \quad ,$$

where $\mathtt{A}_{sm} > 0$ is a numerical constant.

Proof. Let $G_j(x) = \frac{1}{jh_j^d}\sum_{i=1}^{j} R_{i,j}(x)\xi_i$, and recall that $C_j(x) = \frac{1}{jh_j^d}\sum_{i=1}^{j} R_{i,j}(x)f(x_i)$. By the definition of g_j, we can write

$$\begin{aligned}\mathbf{E}[\sup_{x\in\Theta}\|g_j(x)\|^2] &\leq h_j^{-2}\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)\|_{\mathrm{op}}^{-2}\|D_j(x)\|^2\right]\\ &\leq \underbrace{2h_j^{-2}\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)\|_{\mathrm{op}}^{-2}\|C_j(x)\|^2\right]}_{\text{term I}} + \underbrace{2h_j^{-2}\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)\|_{\mathrm{op}}^{-2}\|G_j(x)\|^2\right]}_{\text{term II}} \ ,\end{aligned}$$

where the last inequality is obtained by $(u+v)^2 \leq 2u^2+2v^2$, for any $u, v \geq 0$. Now, for term I, we can write

$$\begin{aligned}\text{term I} \leq &\underbrace{4h_j^{-2}\left(\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)\|_{\mathrm{op}}^{-2}\sup_{x\in\Theta}\|C_j(x)-\mathbf{E}\left[C_j(x)\right]\|^2\right]\right)}_{\text{term III}}\\ &+\underbrace{4h_j^{-2}\left(\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)\|_{\mathrm{op}}^{-2}\right]\mathbf{E}\left[\sup_{x\in\Theta}\|C_j(x)\|^2\right]\right)}_{\text{term IV}} \ ,\end{aligned}$$

To provide the upper bound for term III, by using Cauchy-Schwarz inequality we have

$$\text{term III} \leq 4h_j^{-2}\left(\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)\|_{\mathrm{op}}^{-4}\right]\mathbf{E}\left[\sup_{x\in\Theta}\|C_j(x)-\mathbf{E}\left[C_j(x)\right]\|^4\right]\right)^{\frac{1}{2}} \leq \mathtt{A}_1 j^{-1}h_j^{-d-2}\log(j) \ ,$$

where we utilized Lemmas 5.16 and 5.18, with $\mathtt{A}_1 > 0$ as a numerical constant. For term IV, by invoking Lemmas 5.14(i) and 5.16, we deduce that

$$\text{term IV} \leq \mathtt{A}_2 h_j^{-d-2} \ ,$$

where $\mathtt{A}_2 > 0$ is a numerical constant. Finally, it is enough to provide an upper bound for term II.

$$\text{term II} = 2h_j^{-2}\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)\|_{\mathrm{op}}^{-2}\|G_j(x)\|^2\right] \leq 2\sigma^2\lambda_j^{-2}h_j^{-2-2d}j^{-1}\mathbf{E}\left[\sup_{x}\|R_{1,j}(x)\|^2\right] \ ,$$

where the last inequality is a result of Assumption 8. Thanks to Lemma 5.14(i), and the fact that $j \geq \lambda_j^{-2} h_j^{-d} \log(j)$, we can write

$$\text{term II} \leq 2\sigma^2 h_j^{-2} \log(j) \ . \tag{5.15}$$

Now it is straightforward to see that summation of the terms are dominated by $\mathtt{A}_{\mathrm{sm}} h_j^{-d-2} \log(j)$, where $\mathtt{A}_{\mathrm{sm}} = \mathtt{A}_1 + \mathtt{A}_2 + 2\sigma^2$. By substituting $h_j = \left(\frac{\log(j)}{j}\right)^{\frac{1}{2\beta+d}}$, we conclude the proof. □

Now we are ready to proof theorem 5.2.

Proof. By the definition of the algorithm and contracting property of the Euclidean projection, for any $j \in [n]$, we have

$$\|z_{j+1} - x^*\|^2 \leq \|z_j - x^*\|^2 - \frac{2}{\alpha j} \underbrace{(z_j - x^*)^\top \mathbf{E}\left[g_j(z_j)|z_j\right]}_{\text{term I}} + \frac{1}{\alpha^2 j^2} \mathbf{E}\left[\|g_j(z_j)\|^2 |z_j\right] \ .$$

By adding and subtracting $\nabla f(z_n)$, for term I we get

$$\begin{aligned} \mathbf{E}[\delta_{j+1}|z_j] \leq \delta_j - \frac{2}{\alpha j}\langle z_j - x^*, \nabla f(z_j)\rangle + \frac{2}{\alpha j} \|z_j - x^*\| \, \|\mathbf{E}[g_j(z_j)|z_j] - \nabla f(z_j)\| \\ + \frac{1}{\alpha^2 j^2} \mathbf{E}\left[\|g_j(z_j)\|^2 |z_j\right] \ , \end{aligned} \tag{5.16}$$

where we introduced $\delta_j = \|z_j - x^*\|^2$. Since f is an α-strongly function, we have

$$\alpha \delta_j \leq \langle z_j - x^*, \nabla f(z_j)\rangle \ . \tag{5.17}$$

Combining (5.16) and (5.17), yields

$$\mathbf{E}[\delta_{j+1}|z_j] \leq \left(1 - \frac{2}{j}\right)\delta_j + \frac{2}{\alpha j} \underbrace{\|z_j - x^*\| \, \|\mathbf{E}[g_j(z_j)|z_j] - \nabla f(z_j)\|}_{\text{term II}} + \frac{1}{\alpha^2 j^2} \mathbf{E}\left[\|g_j(z_j)\|^2 |z_j\right] \ .$$

Note that for any $a, b \in \mathbb{R}$ and $\gamma > 0$, we have $2a \cdot b \leq \gamma a^2 + \frac{b^2}{\gamma}$. For term II in (5.18), we can write

$$\|z_j - x^*\| \, \|\mathbf{E}[g_j(z_j)|z_j] - \nabla f(z_j)\| \leq \frac{\alpha}{2}\delta_j + \frac{1}{2\alpha} \|\mathbf{E}[g_j(z_j)|z_j] - \nabla f(z_j)\|^2 \ .$$

Plugging-in the above upper bound for term II, and taking total expectation, yields

$$\begin{aligned}\tilde{\delta}_{j+1} &\leq \left(1-\frac{1}{j}\right)\tilde{\delta}_j + \frac{1}{\alpha^2 j}\mathbf{E}\left[\|\mathbf{E}[g_j(z_j)|z_j] - \nabla f(z_j)\|^2\right] + \frac{1}{\alpha^2 j^2}\mathbf{E}\left[\|g_j(z_j)\|^2\right]\\ &\leq \left(1-\frac{1}{j}\right)\tilde{\delta}_j + \frac{1}{\alpha^2 j}\mathbf{E}\left[\sup_{x\in\Theta}\|g_j(x) - \nabla f(x)\|^2\right] + \frac{1}{\alpha^2 j^2}\mathbf{E}\left[\sup_{x\in\Theta}\|g_j(x)\|^2\right] \quad , \quad (5.18)\end{aligned}$$

where $\tilde{\delta}_j = \mathbf{E}[\delta_j]$. By invoking Lemmas 5.9 and 5.10, we deduce

$$\tilde{\delta}_{j+1} \leq \left(1-\frac{1}{j}\right)\tilde{\delta}_j + \mathtt{A}_1 j^{-1-\frac{2(\beta-1)}{2\beta+1}} \log(j)^{\frac{2(\beta-1)}{2\beta+d}} \alpha^{-2} \quad , \quad (5.19)$$

where $\mathtt{A}_1 = \mathtt{A}_{\text{error}} + \mathtt{A}_{\text{sm}}$, is a numerical constant. At the end, by utilizing (Akhavan et al., 2020, Lemma D.1.), we conclude the proof

$$\mathbf{E}\left\|z_n - x^*\right\|^2 \leq \left(\frac{2\text{diam}(\Theta)}{n} + \mathtt{A}_2 n^{-\frac{2(\beta-1)}{2\beta+d}}\alpha^{-2}\right)\log(n)^{\frac{2(\beta-1)}{2\beta+d}} \quad ,$$

where $\mathtt{A}_2 = \frac{2\beta+2}{d+2}\mathtt{A}_1$, and $\text{diam}(\Theta) = \sup_{x,y\in\Theta}\|x-y\|^2$. □

5.7.2 Proof of the Theorem 5.3

Lemma 5.11. *Under Assumption 8, 9, and 10, for any $x \in \Theta$ we have*

$$\left|\mathbf{E}\left[f_n(x)\right] - f(x)\right| \leq B_{bias} n^{-\frac{\beta}{2\beta+d}} \quad ,$$

where $B_{bias} > 0$ is a numerical constant.

Proof. Let $B = \mathbf{E}[B_n(x)]$, and $\phi_n = f_n(x) - \boldsymbol{A}B^{-1}B_n\boldsymbol{c}_n(f,x)$. It is straightforward to see that

$$\mathbf{E}\left[\phi_n\right] = \mathbf{E}\left[f_n(x)\right] - A\boldsymbol{c}_n(f,x) = \mathbf{E}\left[f_n(x)\right] - f(x) \quad .$$

Therefore we need to provide an upper bound, for the term $\left|\mathbf{E}\left[\phi_n\right]\right|$. Let

$$\begin{aligned}\psi_{1,n} &= \boldsymbol{A}B^{-1}C_n(x) \quad ,\\ \psi_{2,n} &= \boldsymbol{A}\left(B + \lambda_n I\right)^{-1}C_n(x) \quad ,\end{aligned}$$

where $C_n(x) = \frac{1}{nh_n^d}\sum_{k=1}^n R_k(x)f(x_k)$. Now, we can write

$$|\mathbf{E}[\phi_n| \leq \underbrace{|\mathbf{E}[\psi_{1,n} - h_n^{-1}\left(AB^{-1}B_n(x)\boldsymbol{c}_n(f,x)\right)]|}_{\text{term I}} + \underbrace{|\mathbf{E}[\psi_{2,n} - \psi_{1,n}|}_{\text{term II}} + \underbrace{|\mathbf{E}[f_n(x) - \psi_{2,n}]|}_{\text{term III}} \quad .$$

By following similar steps as in the proof of Lemma 5.7, we get

$$\text{term I} \leq \mathtt{B}_1 h_n^{\beta} \quad , \quad \text{term II} \leq \mathtt{B}_2 \lambda_n \quad \text{and} \quad \text{term III} \leq \mathtt{B}_3 h_n^{-\frac{d}{2}} n^{-\frac{1}{2}} \quad ,$$

where $\mathtt{B}_1, \mathtt{B}_2, \mathtt{B}_3 > 0$, are numerical constants. Therefore, we deduce that

$$|\mathbf{E}[\phi_n]| \leq \mathtt{B}_4 \left(h_n^{\beta} + \lambda_n + h_n^{-\frac{d}{2}} n^{-\frac{1}{2}} \right) \quad ,$$

with $\mathtt{B}_4 = \max\left(\mathtt{B}_1, \mathtt{B}_2, \mathtt{B}_3\right)$. We concluding the proof, by substituting $h_n = n^{-\frac{1}{2\beta+d}}$, and $\lambda_n = n^{-\frac{\beta}{2\beta+d}}$. □

Lemma 5.12. *Let Assumptions 8, 9, and 10 hold. Then, for any $x \in \Theta$ we have*

$$\mathbf{E}\left[(f_n(x) - \mathbf{E}\left[f_n(x)\right])\right]^2 \leq \mathtt{B}_{var} n^{-\frac{2\beta}{2\beta+d}} \quad ,$$

where $\mathtt{B}_{var} > 0$ is a numerical constant.

Proof. Similar to the proof of Lemma 5.10, let $G_n(x) = \frac{1}{nh_n^d}\sum_{k=1}^{n} R_k(x)\xi_k$, and $C_n(x) = \frac{1}{nh_n^d}\sum_{k=1}^{n} R_k(x) f(x_k)$. Then, we have

$$\mathbf{E}[(f_n(x) - \mathbf{E}\left[f_n(x)\right])^2] \leq \underbrace{2\mathbf{E}\left[\left\|B_{n,\lambda}^{-1}(x) G_n(x)\right\|^2\right]}_{\text{term I}} + \underbrace{2\mathbf{E}\left[\left\|B_{n,\lambda}^{-1}(x) C_n(x) - \mathbf{E}\left[B_{n,\lambda}^{-1}(x) C_n(x)\right]\right\|^2\right]}_{\text{term II}} \; .$$

For term I, we can write

$$\text{term I} \leq \underbrace{4\mathbf{E}\left[\left\|\left(B_{n,\lambda}^{-1}(x) - (\mathbf{E}[B_{n,\lambda}])^{-1}\right) G_n(x)\right\|^2\right]}_{\text{term III}} + \underbrace{4\left\|\mathbf{E}[B_{n,\lambda}(x)]\right\|^{-2} \mathbf{E}\left[\|G_n(x)\|^2\right]}_{\text{term IV}} \; .$$

By Assumption 8, for term III we get

$$\text{term III} \leq 4\sigma^2 \mu_{\text{min}}^{-2} \lambda_n^{-2} \mathbf{E}\left[\|B_{n,\lambda}(x) - \mathbf{E}[B_{n,\lambda}(x)]\|_{\text{op}}^2 \left(n^{-2} h_n^{-2d} \sum_{k=1}^{n} \|R_k(x)\|^2\right)\right] \; .$$

Now, by using Cauchy-Schwarz inequality, we can write

$$\text{term III} \leq 4\sigma^2 \mu_{\text{min}}^{-2} \lambda_n^{-2} \left(\mathbf{E}\left[\|B_{n,\lambda}(x) - \mathbf{E}[B_{n,\lambda}(x)]\|_{\text{op}}^4\right] \mathbf{E}\left[\left(n^{-2} h_n^{-2d} \sum_{k=1}^{n} \|R_k(x)\|^2\right)^2\right]\right)^{\frac{1}{2}} \; .$$

Utilizing Lemma 5.14, implies

$$\begin{aligned}\text{term III} &\leq 4\sigma^2\mu_{\text{min}}^{-2}\lambda_n^{-2}n^{-1}h_n^{-d}\left(\mathbf{E}\left[\left(n^{-2}h_n^{-2d}\sum_{k=1}^{n}\|R_k(x)\|^2\right)^2\right]\right)^{\frac{1}{2}}\\ &\leq 4\sigma^2\mu_{\text{min}}^{-2}\lambda_n^{-2}n^{-1}h_n^{-d}\left(n^{-4}h_n^{-4d}\left(\sum_{k=1}^{n}\mathbf{E}[\|R_k(x)\|^4]+\sum_{j,k=1}^{n}\mathbf{E}[\|R_j(x)\|^2]\mathbf{E}[\|R_k(x)\|^2]\right)\right)^{\frac{1}{2}} .\end{aligned}$$

By invoking Lemma 5.14(i), and the fact that $1 \leq nh_n$, we deduce that

$$\text{term III} \leq 4\sigma^2\mu_{\text{min}}^{-2}\lambda_n^{-2}n^{-2}h_n^{-2d}\left(p_{\text{max}}\nu_{1,4}+(p_{\text{max}}\nu_{1,2})^2\right)^{\frac{1}{2}} .$$

Now, by using the inequality $n \geq \lambda_n^{-2}h_n^{-d}$, we get

$$\text{term III} \leq \mathtt{B}_1 n^{-1}h_n^{-d} ,$$

where $\mathtt{B}_1 = 4\sigma^2\mu_{\text{min}}^{-2}(p_{\text{max}}\nu_{1,4}+(p_{\text{max}}\nu_{1,2})^2)^{\frac{1}{2}}$. For term IV, we can write

$$\text{term IV} \leq 4\sigma^2\mu_{\text{min}}^{-2}n^{-1}h_n^{-2d}\mathbf{E}\left[\|R_n(x)\|^2\right] \leq \mathtt{B}_2 n^{-1}h_n^{-d} ,$$

where the last inequality is a result of Lemma 5.14(i), with $\mathtt{B}_2 = 4\sigma^2\mu_{\text{min}}^{-2}p_{\text{max}}\nu_{1,2}$. as a numerical constant. For term II, we have

$$\begin{aligned}\text{term II} \leq &\underbrace{4\mathbf{E}\left[\left\|B_{n,\lambda}^{-1}(x)-(\mathbf{E}[B_{n,\lambda}(x)])^{-1}\right\|_{\text{op}}^2\|C_n(x)\|^2\right]}_{\text{term V}}\\ &+\underbrace{4\left\|(\mathbf{E}[B_{n,\lambda}(x)])^{-1}\right\|_{\text{op}}^2\mathbf{E}\left[\|C_n(x)-\mathbf{E}[C_n(x)]\|^2\right]}_{\text{term VI}} .\end{aligned}$$

$$\text{term V} \leq 4M^2\mu_{\text{min}}^{-2}\lambda_n^{-2}k^{-1}h_n^{-d}\left(n^{-3}h_n^{-3d}p_{\text{max}}\nu_{1,4}+n^{-2}h_n^{-2d}p_{\text{max}}^2\nu_{1,2}^2\right)^{\frac{1}{2}} \leq \mathtt{B}_3 n^{-1}h_n^{-d} ,$$

where $\mathtt{B}_3 = 4n^2\mu_{\text{min}}^{-2}(p_{\text{max}}\nu_{1,2}+(p_{\text{max}}\nu_{1,2})^2)^{\frac{1}{2}}$. For term VI, by using Lemma 5.17, we have

$$\text{term VI} \leq \mathtt{B}_4 n^{-1}h_n^{-d} ,$$

with $\mathtt{B}_4 > 0$ as a numerical constant. Finally, by combing all of these, we get

$$\mathbf{E}\left[(f_n(x)-\mathbf{E}[f_n(x)])^2\right] \leq \mathtt{B}_{\text{var}}n^{-1}h_n^{-d} ,$$

where we introduced $\mathtt{B}_{\mathrm{var}} = \mathtt{B}_1 + \mathtt{B}_2 + \mathtt{B}_3 + \mathtt{B}_4$. We conclud the proof by substituting $h_n = n^{-\frac{1}{2\beta+d}}$. □

Proof. We deduce that

$$\mathbf{E}\left[(f_n(x) - f(x))^2\right] = (\mathbf{E}\left[f_n(x)\right] - f(x))^2 + \mathbf{E}\left[(f_n(x) - \mathbf{E}\left[f_n(x)\right])^2\right] \ .$$

We conclude the proof by utilizing Lemmas 5.11 and 5.12. □

5.8 Proof of Theorem 5.5

We first prove (5.10). We apply the scheme of proving lower bounds for estimation of functionals described in Section 2.7.4 in Tsybakov (2009). Moreover, we use its basic form when the problem is reduced to testing two simple hypotheses (that is, the mixture measure μ from Section 2.7.4 in Tsybakov (2009) is the Dirac measure). The functional we are estimating is $F(f) = f^* = \min_{x\in\Theta} f(x)$, where Θ is a sufficiently large Euclidean ball centered at 0. We choose the two hypotheses as the probability measures $P_1^{\otimes n}$ and $P_2^{\otimes n}$, where P_j stands for the distribution of a pair (x_i, y_i) satisfying (5.1) with $f = f_j$, $j = 1, 2$. For $r > 0$, $\delta > 0$, we set

$$f_1(x) = \alpha(1+\delta)\|x\|^2/2, \quad f_2(x) = f_1(x) + rh_n^{\beta}\Phi\left(\frac{x - x^{(n)}}{h_n}\right),$$

where $h_n = n^{-1/(2\beta+d)}$, $x^{(n)} = (h_n/8, 0, \dots, 0) \in \mathbb{R}^d$ and $\Phi(x) = \prod_{i=1}^d \Psi(x_i)$ with

$$\Psi(t) = \int_{-\infty}^{t} (\eta(y + 1/2) - \eta(y))\, \mathrm{d}y,$$

where $\eta(\cdot)$ is an infinitely many times differentiable function on $\mathbb{R}^1$ such that

$$\eta(x) \geq 0, \quad \eta(x) = \begin{cases} 0, & x \notin [0, 1/2] \\ 1, & x \in [1/8, 3/8] \end{cases}.$$

It is shown in Tsybakov (1990b) that if r is small enough the functions f_1 and f_2 are α-strongly convex and belong to $\mathcal{F}_\beta(L)$. Thus, $f_j \in \mathcal{F}_{\beta,\alpha}(L), j = 1, 2$. It is also not hard to check (cf. Tsybakov (1990b)) that for the function $\eta_1(y) = \eta(y+1/2) - \eta(y)$ we have

$$\eta_1\left(-\frac{r\Psi^{d-1}(0)h_n^{\beta-2}}{\alpha(1+\delta)} - \frac{1}{8}\right) = 1$$

when $r < \alpha(1+\delta)/4$. Using this remark we get that the minimizers $x_j^* = \arg\min_{x\in\Theta} f_j(x)$ have the form

$$x_1^* = (0, 0, \dots, 0) \quad \text{and} \quad x_2^* = \left(-\frac{r\Psi^{d-1}(0)h_n^{\beta-1}}{\alpha(1+\delta)}, 0, \dots, 0 \right).$$

The values of the functional F on f_1 and f_2 are $F(f_1) = 0$ and

$$\begin{aligned}
F(f_2) &= f_2(x_2^*) \\
&= \frac{r^2\Psi^{2(d-1)}(0)}{2\alpha(1+\delta)} h_n^{2(\beta-1)} + r\Psi^{d-1}(0)\Psi\left(-\frac{r\Psi^{d-1}(0)h_n^{\beta-2}}{\alpha(1+\delta)} - \frac{1}{8} \right) h_n^{\beta} \\
&\geq \frac{r^2\Psi^{2(d-1)}(0)}{2\alpha(1+\delta)} h_n^{2(\beta-1)} + r\Psi^{d-1}(0)\Psi(-1/4)h_n^{\beta} \quad (\text{for } r \text{ small enough}) \\
&\geq r\Psi^{d-1}(0)\Psi(-1/4)h_n^{\beta}.
\end{aligned}$$

Here, $\Psi(0) = \int_{-\infty}^{\infty} \eta(y)\,\mathrm{d}y > 0$ and $\Psi(-1/4) = \int_{-\infty}^{1/4} \eta(y)\,\mathrm{d}y > 0$.

Note that assumption (i) of Theorem 2.14 in Tsybakov (2009) is satisfied with $\beta_0 = \beta_1 = 0$, $c = 0$ and $s = r\Psi^{d-1}(0)\Psi(-1/4)h_n^{\beta}/2$. Therefore, by Theorem 2.15 (ii) in Tsybakov (2009), (5.10) will be proved if we show that

$$\mathrm{H}^2\left(P_1^{\otimes n}, P_2^{\otimes n}\right) \leq a < 2, \tag{5.20}$$

where $\mathrm{H}^2(P, Q)$ denotes the Hellinger distance between the probability measures P and Q. Using assumption (5.8) we obtain

$$\begin{aligned}
\mathrm{H}^2\left(P_1^{\otimes n}, P_2^{\otimes n}\right) &= 2\left(1 - \left(1 - \frac{\mathrm{H}^2(P_1, P_2)}{2}\right)^n\right) \\
&\leq n\,\mathrm{H}^2(P_1, P_2) \quad (\text{as } (1-x)^n \geq 1 - xn,\ x \in [0,1]) \\
&= n \int \left(\sqrt{p_\xi(y)} - \sqrt{p_\xi\left(y + (f_1(x) - f_2(x))\right)} \right)^2 p(x)\,\mathrm{d}x\,\mathrm{d}y \\
&\leq nI_* \int (f_1(x) - f_2(x))^2\, p(x)\,\mathrm{d}x \\
&= nI_* r^2 h_n^{2\beta+d} \int \Phi^2(u) p\left(x^{(n)} + uh_n\right)\,\mathrm{d}u \\
&\leq p_{\max} I_* r^2 \int \Phi^2(u)\,\mathrm{d}u, \quad \text{for } r \leq v_0,
\end{aligned}$$

where $p_{\max}$ is the maximal value of the density $p(\cdot)$ of x_i. Choosing $r \leq \sqrt{a/\left(p_{\max} I_* \int \Phi^2(u)\,\mathrm{d}u\right)}$, with $a < 2$ we obtain (5.20). This completes the proof of (5.10).

In order to prove (5.9), it suffices to use the same construction of two hypotheses as above, apply the Hellinger version of Theorem 2.2 from Tsybakov (2009), and to notice that $\|x_1^* - x_2^*\| \geq cn^{-(\beta-1)/(2\beta+d)}$, where $c > 0$ is a constant.

5.9 Proof of Theorem 5.6

We apply again the scheme of proving lower bounds for estimation of functionals from Section 2.7.4 in Tsybakov (2009). However, we use a different construction of the hypotheses. Without loss of generality, assume that $n \geq 2$, that Θ contains the cube $[0,1]^d$. Define $h_n = (n/\log(n))^{-1/(2\beta+d)}$, $N = (1/h_n)^d$, and assume without loss of generality that N is an integer. For $r > 0$, we set

$$f_j(x) = -rh_n^\beta \Phi\left(\frac{x - \mathbf{t}^{(j)}}{h_n}\right), \quad j = 1, \dots, N,$$

where $\Phi(x) = \prod_{i=1}^d \Psi(x_i)$, where $\Psi(\cdot)$ is an infinitely many times differentiable function on $\mathbb{R}$ taking positive values on its support $[-1/2, 1/2]$, and we denote by $\mathbf{t}^{(1)}, \dots, \mathbf{t}^{(N)}$ the N points of the equispaced grid on $[0,1]^d$ with step h_n over each coordinate, such that the supports of all f_j's are included in $[0,1]^d$ and are disjoint. It is not hard to check that for r small enough all the functions f_j, $j = 1, \dots, N$, belong to $\mathcal{F}_\beta(L)$.

We consider the product probability measures $P_0^{\otimes n}$ and $P_1^{\otimes n}, \dots P_N^{\otimes n}$, where P_0 stands for the distribution of a pair (x_i, y_i) satisfying (5.1) with $f \equiv 0$, and P_j stands for the distribution of (x_i, y_i) satisfying (5.1) with $f = f_j$. Consider the mixture probability measure $\mathbb{P}_\mu = \frac{1}{N}\sum_{j=1}^N P_j^{\otimes n}$, where μ denotes the uniform distribution on $\{1, \dots, N\}$.

Note that, for each $j = 1, \dots, N$, we have $F(f_j) = -rh_n^\beta \Phi_{\max}$, where $F(f) = f^* = \min_{x\in\Theta} f(x)$, and $\Phi_{\max} > 0$ denotes the maximal value of function $\Phi(\cdot)$. Let

$$\chi^2(P', P) = \int (\mathrm{d}P'/\mathrm{d}P)^2 \,\mathrm{d}P - 1$$

denote the chi-square divergence between two mutually absolutely continuous probability measures P' and P. We will use the following lemma, which is a special case of Theorem 2.15 in Tsybakov (2009).

Lemma 5.13. *Assume that there exist $v > 0, b > 0$ such that $F(f_j) = -2v$ for $j = 1, \dots, N$ and $\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) \leq b$, Then*

$$\inf_{\hat{f}_n} \sup_{j=0,1,\dots,N} P_j^{\otimes n}\big(|\hat{f}_n - F(f_j)| \geq v\big) \geq \frac{1}{4}\exp(-b),$$

where $\inf_{\hat{f}_n}$ denotes the infimum over all estimators.

In our case, the first condition of this lemma is satisfied with $v = rh_n^\beta \Phi_{\max}/2$. We now check that the second condition $\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) \le b$ holds with some constant $b > 0$ independent of n. Using a standard representation of the chi-square divergence of a Gaussian mixture from the pure Gaussian noise measure (see, for example, Lemma 8 in Carpentier et al. (2019)) we obtain

$$\begin{aligned}
\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) &= \frac{1}{N^2}\sum_{j,j'=1}^{N} \mathbf{E}\exp\left(\frac{\sum_{i=1}^n f_j(x_i) f_{j'}(x_i)}{\sigma^2}\right) - 1 \\
&= \frac{1}{N^2}\sum_{j,j'=1}^{N} \mathbf{E}\exp\left(\frac{\sum_{i=1}^n f_j(x_i) f_{j'}(x_i)}{\sigma^2}\right) - 1 \\
&= \frac{1}{N^2}\sum_{j=1}^{N} \mathbf{E}\exp\left(\frac{\sum_{i=1}^n f_j^2(x_i)}{\sigma^2}\right) + \frac{N(N-1)}{N^2} - 1 \\
&\le \frac{1}{N^2}\sum_{j=1}^{N} \mathbf{E}\exp\left(\frac{\sum_{i=1}^n f_j^2(x_i)}{\sigma^2}\right) \\
&= \frac{1}{N^2}\sum_{j=1}^{N} \left[\mathbf{E}\exp\left(\frac{f_j^2(x_1)}{\sigma^2}\right)\right]^n,
\end{aligned}$$

where the equality in the third line is due to the fact that if $j \ne j'$ then f_j and $f_{j'}$ have disjoint supports and thus $f_j(x_i)f_{j'}(x_i) = 0$. Note that $\max_{x\in\mathbb{R}^d} f_j^2(x) \le r^2\Phi_{\max}^2$, for all $j = 1, \dots, N$. Choose r such that $r \le \sigma/\Phi_{\max}$. Then $\frac{f_j^2(x_1)}{\sigma^2} \le 1$, and using the elementary inequality $\exp(u) \le 1+2u, u \in [0,1]$, we obtain that $\exp\left(\frac{f_j^2(x_1)}{\sigma^2}\right) \le 1+\frac{2f_j^2(x_1)}{\sigma^2}$ for all $j = 1, \dots, N$. Substituting this bound in the last display and noticing that $\mathbf{E}(f_j^2(x_1)) = \int f_j^2(x)p(x)dx \le p_{\max} r^2 h_n^{2\beta+d} \int \Phi^2(x)\,\mathrm{d}x = c_* \frac{\log n}{n}$, where $c_* = p_{\max} r^2 \int \Phi^2(x)\,\mathrm{d}x$, we obtain:

$$\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) \le \frac{1}{N}\left[1 + \frac{2\mathbf{E}(f_j^2(x_1))}{\sigma^2}\right]^n \le \frac{1}{N}\left[1 + \frac{2c_*\log n}{\sigma^2 n}\right]^n \le \frac{1}{N}\exp\left(\frac{2c_*\log n}{\sigma^2}\right) = \frac{n^{c_0}}{N},$$

where $c_0 = 2c_*/\sigma^2 = 2p_{\max} r^2 \int \Phi^2(x)\,\mathrm{d}x/\sigma^2$. Since $N = (n/\log n)^{\frac{d}{2\beta+d}}$ we finally get

$$\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) \le n^{c_0 - \frac{d}{2\beta+d}} (\log n)^{\frac{d}{2\beta+d}}.$$

By choosing r small enough to have $c_0 \le \frac{d}{2(2\beta+d)}$ we obtain that $\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) \le \left(\frac{\log n}{\sqrt{n}}\right)^{\frac{d}{2\beta+d}} \le \left(\frac{\log 2}{\sqrt{2}}\right)^{\frac{d}{2\beta+d}} := b$. Thus, the second condition of Lemma 5.13 holds if r is chosen as a small enough constant. Notice that, in Lemma 5.13, the rate v is of the desired order $(n/\log n)^{-\frac{\beta}{2\beta+d}}$. The result of the theorem now follows from Lemma 5.13 and the standard argument to obtain the lower bounds, see Section 2.7.4 in Tsybakov (2009).

5.10 Proof of lemmas

Recall that $\Theta' = \{x + y : x \in \Theta \quad \text{and} \quad y \in \mathrm{Supp}(k)\}$.

Lemma 5.14. *For any $q \geq 1$, let*

$$\nu_{1,q} = \int_{\mathbb{R}^d} \|U(u)k(u)\|^q \, \mathrm{d}u \ , \qquad \nu_{2,q} = \int_{\mathbb{R}^d} \left\| U\left(u\right) U^\top \left(u\right) k\left(u\right) \right\|_{\mathrm{op}}^q \mathrm{d}u \ ,$$

and $p_{\max} = \max_{y \in \Theta'} p(y)$. Then, under Assumption 10, for any $x \in \Theta$, $j \in [n]$, and $i \in [j]$, we have

(i) $h_j^{-d}\mathbf{E}\left[\sup_{x\in\Theta} \|R_{i,j}(x)\|^q\right] \leq p_{\max}\nu_{1,q}$.

(ii) $h_j^{-d}\mathbf{E}\left[\sup_{x\in\Theta} \|M_{i,j}(x)\|_{\mathrm{op}}^q\right] \leq p_{\max}\nu_{2,q}$.

(iii) There exists $\mu_{\min} > 0$, such that $\inf_{x\in\Theta} \mu_{\min}\left(\mathbf{E}\left[B_j(x)\right]\right) \geq \mu_{\min}$.

Proof. We have

$$\begin{aligned} h_j^{-d}\mathbf{E}\left[\left\|\sup_{x\in\Theta} R_{i,j}(x)\right\|^q\right] &= h_j^{-d}\int_{\mathbb{R}^d} \sup_{x\in\Theta}\left\| U\left(\frac{y-x}{h_j^d}\right) k\left(\frac{y-x}{h_j^d}\right)\right\|^q p(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} \|U(u)k(u)\|^q \sup_{x\in\Theta} p(x+h_j u)\,\mathrm{d}u \leq p_{\max}\nu_{1,q} \ . \end{aligned}$$

For (ii) we can write

$$\begin{aligned} h_j^{-d}\mathbf{E}\left[\sup_{x\in\Theta}\|M_{i,j}(x)\|_{\mathrm{op}}^q\right] &= h_j^{-d}\int_{\mathbb{R}^d}\sup_{x\in\Theta}\left\|U\left(\frac{y-x}{h_j^d}\right)U^\top\left(\frac{y-x}{h_j^d}\right)k\left(\frac{y-x}{h_j^d}\right)\right\|_{\mathrm{op}}^q p(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d}\left\|U\left(u\right)U^\top\left(u\right)k\left(u\right)\right\|_{\mathrm{op}}^q \sup_{x\in\Theta} p(x+h_j u)\,\mathrm{d}u \leq p_{\max}\nu_{2,q} \ . \end{aligned}$$

Similarly for (iii), we get

$$\begin{aligned} \mathbf{E}\left[B_j(x)\right] &= h_j^{-d}\mathbf{E}\left[U\left(\frac{x_1-x}{h_j^d}\right)U^\top\left(\frac{x_1-x}{h_j^d}\right)k\left(\frac{x_1-x}{h_j^d}\right)\right] \\ &= \int_{\mathbb{R}^d} U\left(u\right)U^\top\left(u\right)k\left(u\right)p(x+h_j u)\,\mathrm{d}u \ . \end{aligned}$$

By denoting $H = \int_{\mathbb{R}^d} U\left(u\right)U^\top\left(u\right)k\left(u\right)$, we deduce that $\inf_{x\in\Theta}\mu_{\min}\left(\mathbf{E}\left[B_j(x)\right]\right) \geq p_{\min}\mu_{\min}\left(H\right)$. By (Tsybakov, 1986, Lemma 1), we have $\mu_{\min}\left(H\right) > 0$. We conclude the proof by letting $\mu_{\min} = p_{\min}\mu_{\min}\left(H\right)$. □

Lemma 5.15. *Let $j \in [n]$, with $j \geq 3$ and $h_j = \left(\frac{\log(j)}{j}\right)^{\frac{1}{2\beta+d}}$, and assume that Assumption 10 holds. Then, for any $x \in \Theta$, we have*

$$\mathbf{E}\left[\sup_{x\in\Theta} \|B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right]\|_{\mathrm{op}}^4\right] \leq \mathtt{A}_1 h_j^{-2d} j^{-2}\log(j)^2 \ . \tag{5.21}$$

Furthermore, for $j \geq \lambda_j^{-2} h_j^{-d}\log(j)$, we have $\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)\|_{\mathrm{op}}^{-4}\right] \leq \mathtt{A}_2\mu_{\min}^{-4}$, where $\mathtt{A}_1, \mathtt{A}_2 > 0$ are numerical constants.

Proof. In order to prove (5.21), first we show that $\|B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right]\|_{\mathrm{op}}$ is upper bounded by a Lipschitz function. Let $Q_{i,j}(x) = h_j^{-d}M_{i,j}(x) - h_j^{-d}\mathbf{E}\left[B_{j,\lambda}(x)\right]$.

1. There exists a Lipschitz upper bound: First note that

$$\|B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right]\|_{\mathrm{op}} \leq \sum_{s=1}^{S}\left|j^{-1}h_j^{-d}\sum_{i=1}^{j}\left(Q_{i,j}(x) - \mathbf{E}\left[Q_{i,j}(x)\right]\right)_s\right| \ ,$$

where $\left(Q_{i,j}(x) - \mathbf{E}\left[Q_{i,j}(x)\right]\right)_s$ is the (s,s)-entry of the matrix $Q_{i,j}(x) - \mathbf{E}\left[Q_{i,j}(x)\right]$. Recall that the kernel function k is L_k-Lipschitz. Furthermore, for $s \in [S]$ let $G^{(s)} : \mathbb{R}^d \to \mathbb{R}$, such that

$$G^{(s)}(u) = \Big(U\left(u\right)U^{\top}\left(u\right)\Big)_s \ ,$$

it is straightforward to check that $G^{(s)}$ is a continuously differentiable function. Denote Ω as a convex and compact subset of $\mathbb{R}^d$, such that $\mathrm{Supp}(k) \subseteq \Omega$, and let $L_G^{(s)} = \max_{u\in\Omega}\left\|\nabla G^{(s)}(u)\right\|$, and $L_G = \max_{s\in[S]} L_s$. Now, it is clear to see that for any $s \in [S]$, $G^{(s)}$ is a L_G-Lipschitz function on $\mathrm{Supp}(k)$. Moreover, for any $s \in [S]$, and $x, y \in \Theta$, we can write

$$\begin{aligned}&\left|j^{-1}h_j^{-d}\sum_{i=1}^{j}\left((Q_{i,j}(x) - \mathbf{E}[Q_{i,j}(x)])_s - (Q_{i,j}(y) - \mathbf{E}\left[Q_{i,j}(y)\right])_s\right)\right| \\ &\qquad \leq j^{-1}h_j^{-d}\underbrace{\sum_{i=1}^{j}|(Q_{i,j}(x))_s - (Q_{i,j}(y))_s|}_{\text{term I}} \\ &\qquad + j^{-1}h_j^{-d}\underbrace{\sum_{i=1}^{j}\mathbf{E}\left[|(Q_{i,j}(x))_s - (Q_{i,j}(y))_s|\right]}_{\text{term II}} \ .\end{aligned}$$

For term I if $h_j^{-1}(x_i - x), h_j^{-1}(x_i - y) \in \mathrm{Supp}(k)$, we have

$$\begin{aligned}
\text{term I} &= \sum_{i=1}^{j} \left| G^{(s)}\left(\frac{x_i - x}{h_j}\right) k\left(\frac{x_i - x}{h_j}\right) - G^{(s)}\left(\frac{x_i - y}{h_j}\right) k\left(\frac{x_i - y}{h_j}\right) \right| \\
&= \sum_{i=1}^{j} \Bigg| \left(G^{(s)}\left(\frac{x_i - x}{h_j}\right) - G^{(s)}\left(\frac{x_i - y}{h_j}\right)\right) k\left(\frac{x_i - x}{h_j}\right) \\
&\qquad + \left(k\left(\frac{x_i - x}{h_j}\right) - k\left(\frac{x_i - y}{h_j}\right)\right) G^{(s)}\left(\frac{x_i - y}{h_j}\right) \Bigg| \\
&\le j h_j^{-1} \mathtt{A}_3 \left\| x - y \right\| \ ,
\end{aligned}$$

where $\mathtt{A}_3 = \max_{u \in \mathrm{Supp}(k)} L_G k(u) + \max_{s \in [S], i \in [j], u \in \mathrm{Supp}(k)} L_k G^s(u)$. The scenarios when either one or both of the points $h_j^{-1}(x_i - x), h_j^{-1}(x_i - y)$ do not belong to $\mathrm{Supp}(k)$, can be treated similarly. For term II, with the exact same steps, we can write

$$\text{term II} \le j h_j^{-1} \mathtt{A}_3 \left\| x - y \right\| \ .$$

By combining all of these, we deduce that

$$\sum_{s=1}^{S} \left| j^{-1} h_j^{-d} \sum_{i=1}^{j} \left(\left(Q_{i,j}(x) - \mathbf{E}\left[Q_{i,j}(x)\right]\right)_s - \left(Q_{i,j}(y) + \mathbf{E}\left[Q_{i,j}(y)\right]\right)_s \right) \right| \le \mathtt{A}_{\mathrm{Lip}} h_j^{-d-1} \left\| x - y \right\| \ ,$$

where we introduced $\mathtt{A}_{\mathrm{Lip}} = 2S\mathtt{A}_3$.

Provide an upper bound for the probability For any $t \ge 0$, we can write

$$\begin{aligned}
\mathbf{P}\left[\sup_{x \in \Theta} \left\| B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right] \right\|_{\mathrm{op}} \ge t \right] &\le \mathbf{P}\left[\sum_{s=1}^{S} \sup_{x \in \Theta} \left| j^{-1} h_j^{-d} \sum_{i=1}^{j} F_i^{(s)}(x) \right| \ge t \right] \\
&\le \underbrace{\sum_{s=1}^{S} \mathbf{P}\left[\sup_{x \in \Theta} \left| j^{-1} h_j^{-d} \sum_{i=1}^{j} F_i^{(s)}(x) \right| \ge \frac{t}{S} \right]}_{\text{term III}} \ ,
\end{aligned} \tag{5.22}$$

where we defined $F_i^{(s)}(x) = \left(Q_{i,j}(x) - \mathbf{E}\left[Q_{i,j}(x)\right]\right)_s$. From now on, we focus on providing an upper bound for term III. For $\epsilon > 0$, consider an ϵ-net of Θ, namely $\mathcal{N}$, with cardinality $\mathcal{N}(\Theta, \epsilon)$. Therefore, for any $x \in \Theta$, there exists $y \in \mathcal{N}$, such that $\|x - y\| < \epsilon$, and we can write

$$\begin{aligned}
\text{term III} &\le \sum_{s=1}^{S} \mathcal{N}(\Theta, \epsilon) \sup_{x \in \mathcal{N}} \mathbf{P}\left[\left| j^{-1} h_j^{-d} \sum_{i=1}^{j} F_i^{(s)}(x) \right| \ge \frac{t}{S} - \mathtt{A}_{\mathrm{Lip}} h_j^{-d-1} \epsilon \right] \\
&\le \sum_{s=1}^{S} \left(\frac{\mathrm{diam}(\Theta)}{\epsilon} + 1 \right)^d \sup_{x \in \mathcal{N}} \mathbf{P}\left[\left| j^{-1} h_j^{-d} \sum_{i=1}^{j} F_i^{(s)}(x) \right| \ge \frac{t}{S} - \mathtt{A}_{\mathrm{Lip}} h_j^{-d-1} \epsilon \right] \ .
\end{aligned}$$

where $\mathrm{diam}(\Theta) = \max_{x,y\in\Theta} \|x-y\|$, and we used the fact that $\mathcal{N}(\Theta,\epsilon) \leq \left(\frac{\mathrm{diam}(\Theta)}{\epsilon}+1\right)^d$. By assigning $\epsilon = \frac{t}{2\mathtt{A}_{\mathrm{Lip}}S} h_j^{d+1}$, we get

$$\text{term III} \leq \sum_{s=1}^{S} \left(\frac{2\mathtt{A}_{\mathrm{Lip}}S\mathrm{diam}(\Theta)}{t} \cdot h_j^{-d-1}+1\right)^d \sup_{x\in\mathcal{N}} \mathbf{P}\left[\left|j^{-1}h_j^{-d}\sum_{i=1}^{j} F_i^{(s)}(x)\right| \geq \frac{t}{2S}\right] \ .$$

By invoking Bernstein inequality, we deduce that

$$\text{term III} \leq \sum_{s=1}^{S} \left(\frac{2\mathtt{A}_{\mathrm{Lip}}S\mathrm{diam}(\Theta)}{t} \cdot h_j^{-d-1}+1\right)^d \mathbf{P}\left[-\frac{1}{2}\cdot\min\left(\frac{jt^2}{S^2\upsilon^2}, \frac{jt}{SK'}\right)\right] \ , \tag{5.23}$$

where we introduced

$$\upsilon^2 = \sup_{x\in\mathcal{N}, s\in[S]} \mathbf{E}\left[\left|h_j^{-d}F_1^{(s)}(x)\right|^2\right] \ , \quad \text{and} \quad K' = \sup_{x\in\mathcal{N}, s\in[S], i\in[j]} h_j^{-d}\left|\left(Q_{i,j}(x) - \mathbf{E}\left[Q_{i,j}(x)\right]\right)_s\right| \ .$$

We proceed the proof by providing upper bounds for the terms υ and K'. For υ, we can write

$$\begin{aligned}
\upsilon^2 &= \sup_{x\in\mathcal{N}, s\in[S]} h_j^{-2d}\mathbf{E}\left[\left|\left(Q_{1,j}(x) - \mathbf{E}\left[Q_{1,j}(x)\right]\right)_s\right|^2\right] \\
&\leq \sup_{x\in\mathcal{N}, s\in[S]} h_j^{-2d}\mathbf{E}\left[\left|\left(Q_{1,j}(x)\right)_s\right|^2\right] \\
&\leq \sup_{x\in\mathcal{N}, s\in[S]} h_j^{-d}\int\left|\left(U(u)U^\top(u)k(u)\right)_s\right|^2 p(x+h_ju) \leq h_j^{-d}\mathtt{A}_4 \ ,
\end{aligned}$$

where $\mathtt{A}_4 = p_{\max}\sup_{s\in[S]}\int\left|\left(U(u)U^\top(u)k(u)\right)_s\right|^2 \mathrm{d}u$. Similarly, for K', we have

$$K' \leq \sup_{x\in\mathcal{N}, s\in[S], i\in[j]} h_j^{-d}\left(\left|\left(Q_{i,j}(x)\right)_s\right| + \mathbf{E}\left[\left|\left(Q_{i,j}(x)\right)_s\right|\right]\right) \leq h_j^{-d}\mathtt{A}_5 \ ,$$

where we introduced $\mathtt{A}_5 = 2\sup_{u\in\mathrm{Supp}(k), s\in[S]}\kappa_{\max}\left|\left(U(u)U^\top(u)\right)_s\right|$. By substituting these bounds in (5.23), we get

$$\begin{aligned}
\text{term III} &\leq \sum_{s=1}^{S}\left(\frac{\mathtt{A}_7}{t}h_j^{-d-1}+1\right)^d \exp\left(-\mathtt{A}_6\cdot\min\left(jt^2h_j^d, jth_j^d\right)\right) \\
&= S\exp\left(-\mathtt{A}_6\cdot\min\left(jt^2h_j^d, jth_j^d\right) + d\log\left(\frac{\mathtt{A}_7}{t}h_j^{-d-1}+1\right)\right) \ ,
\end{aligned} \tag{5.24}$$

where $\mathtt{A}_6 = \min\left(\frac{1}{2S^2\mathtt{A}_2}, \frac{1}{2S\mathtt{A}_3}\right)$, and $\mathtt{A}_7 = 2\mathtt{A}_{\mathrm{Lip}}S\mathrm{diam}(\Theta)$. Finlay, by replacing (5.24) in (5.22), we get

$$\mathbf{P}\left[\sup_{x\in\Theta}\left\|B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right]\right\|_{\mathrm{op}} \geq t\right] \leq S\exp\left(-\mathtt{A}_6\cdot\min\left(jt^2h_j^d, jth_j^d\right) + d\log\left(\frac{\mathtt{A}_7}{t}h_j^{-d-1}+1\right)\right) \ .$$

The upper bound: For any $a \geq 0$, we can write

$$\mathbf{E}\left[\sup_{x\in\Theta} \|B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right]\|_{\text{op}}^4\right] = \int_{t=0}^{\infty} 4t^3 \mathbf{P}\left[\sup_{x\in\Theta} \|B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right]\|_{\text{op}} \geq t\right] \mathrm{d}t$$

$$\leq a^4 + S\int_{t=a}^{\infty} 4t^3 \exp\left(\underbrace{-\mathtt{A}_6 \cdot \min\left(jt^2h_j^d, jth_j^d\right)}_{\text{term IV}} + \underbrace{d\log\left(\frac{\mathtt{A}_8}{t}h_j^{-d-1}+1\right)}_{\text{V}}\right) \mathrm{d}t \ ,$$

where $\mathtt{A}_8 = \max(\mathtt{A}_7, 1, \mathtt{A}_6^{-2})$. Now we wish to assign a large enough to ensure that term IV dominates term V. Let $a = 2d{\cdot}\max(1, \frac{d+3}{4\beta+2d})^{\frac{2\beta+d}{\beta}} \max\left(\sqrt[3]{\frac{\mathtt{A}_8}{\mathtt{A}_6}}, \sqrt{\frac{\mathtt{A}_8}{\mathtt{A}_6}}\right) \sqrt{\log(j)}h_j^{-\frac{d}{2}}j^{-\frac{1}{2}}$. So we have two possibilities. First assume that $a < 1$, then we have

$$\text{term IV} = \mathtt{A}_6 j a^2 h_j^d \geq 4d^2 \max\left(1, \left(\frac{d+3}{4\beta+2d}\right)^2\right) \cdot \mathtt{A}_6^{\frac{1}{3}}\mathtt{A}_8^{\frac{2}{3}} \log(j) \ .$$

Since $\mathtt{A}_8 \geq 1$, we have

$$\text{term V} \leq d\log\left(\frac{2\mathtt{A}_8}{a}h_j^{-d-1}\right) \leq d\log\left(\frac{\mathtt{A}_6^{\frac{1}{3}}\mathtt{A}_8^{\frac{2}{3}}}{d}\left(\frac{j}{\log(j)}\right)^{\frac{d+3}{4\beta+2d}}\right)$$

$$\leq d\max\left(\frac{d+3}{4\beta+2d}, 1\right)\log\left(\frac{\mathtt{A}_6^{\frac{1}{3}}\mathtt{A}_8^{\frac{2}{3}}}{d}j\right) \ ,$$

where the last inequality is obtained from the fact that $\log(j) \geq 1$. Therefore, we can deduce that

$$2 \cdot \text{term V} \leq \text{term IV} \ .$$

Now, assume that $a \geq 1$, then we have

$$\text{term IV} = \mathtt{A}_6 j a h_j^d \geq 2d\max\left(1, \frac{d+3}{4\beta+2d}\right) \cdot \frac{2\beta+d}{\beta} \cdot \mathtt{A}_6^{\frac{2}{3}}\mathtt{A}_8^{\frac{1}{3}} \sqrt{\log(j)}h_j^{\frac{d}{2}}j^{\frac{1}{2}}$$

$$\geq 2d\max\left(1, \frac{d+3}{4\beta+2d}\right) \cdot \frac{2\beta+d}{\beta} \cdot \mathtt{A}_6^{\frac{2}{3}}\mathtt{A}_8^{\frac{1}{3}} j^{\frac{\beta}{2\beta+d}} \ ,$$

and

$$
\begin{aligned}
\text{term V} &\le d \log\left(\frac{\mathtt{A}_6^{\frac{2}{3}}\mathtt{A}_8^{\frac{1}{3}}}{2d}\left(\frac{j}{\log(j)}\right)^{\frac{d+3}{4\beta+2d}} + 1\right)\\
&\le d \max\left(1, \frac{d+3}{4\beta+d}\right) \log\left(\frac{\mathtt{A}_6^{\frac{2}{3}}\mathtt{A}_8^{\frac{1}{3}}}{2d} j + 1\right)\\
&\le d \max\left(1, \frac{d+3}{4\beta+d}\right) \cdot \frac{2\beta+d}{\beta} \cdot \log\left(\left(\mathtt{A}_6^{\frac{2}{3}}\mathtt{A}_8^{\frac{1}{3}}\right)^{\frac{\beta}{2\beta+d}} j^{\frac{\beta}{2\beta+d}} + 1\right)\\
&\le d \max\left(1, \frac{d+3}{4\beta+d}\right) \cdot \frac{2\beta+d}{\beta} \cdot \mathtt{A}_6^{\frac{2}{3}}\mathtt{A}_8^{\frac{1}{3}} j^{\frac{\beta}{2\beta+d}} \quad ,
\end{aligned}
$$

where the last holds, since $\mathtt{A}_6^{\frac{2}{3}}\mathtt{A}_8^{\frac{1}{3}} \ge 1$, and $\frac{\beta}{2\beta+d} \le 1$. Also, we used the inequality $\log(x+1) \le x$, for all $x > 0$. Then we have

$$
2 \cdot \text{term V} \le \text{term IV} \quad .
$$

Therefore, we deduce that

$$
\begin{aligned}
\mathbf{E}\left[\sup_{x\in\Theta} \|B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right]\|_{\text{op}}^4\right] &\le \mathtt{A}_9 j^{-2} h_j^{-2d} \log(j)^2\\
&\quad + \underbrace{S\int_{t=0}^{\infty} 4t^3 \exp\left(-\mathtt{A}_{10} \min\left(jt^2 h_j^d, jth_j^d\right)\right) \mathrm{d}t}_{\text{term VI}} \quad ,
\end{aligned}
$$

where $\mathtt{A}_9 = \left(2d \cdot \max(1, \frac{d+3}{4\beta+2d})\frac{2\beta+d}{\beta} \max\left(\sqrt[3]{\frac{\mathtt{A}_8}{\mathtt{A}_6}}, \sqrt{\frac{\mathtt{A}_8}{\mathtt{A}_6}}\right)\right)^4$ and $\mathtt{A}_{10} = \frac{\mathtt{A}_6}{2}$. To conclude the proof it is enough to provide an upper bound for term VI. In order to calculate the integral appeared in term VI, we proceed with he similar steps as in the proof of Lemma 5.16 and we obtain

$$
\text{term VI} \le \mathtt{A}_{11} j^{-2} h_j^{-2d} \quad ,
$$

where $\mathtt{A}_{11} > 0$ only depends on $\mathtt{A}_{10}$ and S. We conclude the first part of the proof by letting $\mathtt{A}_1 = \mathtt{A}_9 + \mathtt{A}_{11}$. For the second part of the proof, similar to the proof of Lemma 5.16, we can write

$$
\begin{aligned}
\mathbf{E}\left[\sup_{x\in\Theta} \|B_{j,\lambda}(x)^{-1}\|_{\text{op}}^4\right] &\le 4\mathbf{E}\left[\sup_{x\in\Theta} \left\|B_{j,\lambda}(x)^{-1} - \left(\mathbf{E}\left[B_{j,\lambda}(x)\right]\right)^{-1}\right\|_{\text{op}}^4\right] + 4\sup_{x\in\Theta} \left\|\left(\mathbf{E}\left[B_{j,\lambda}(x)\right]\right)^{-1}\right\|_{\text{op}}^4\\
&\le 4\mu_{\text{min}}^{-4}\lambda^{-4}\mathbf{E}\left[\sup_{x\in\Theta} \|B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right]\|_{\text{op}}^4\right] + 4\lambda_{\text{min}}^{-4} \quad .
\end{aligned}
$$

By the first part of the proof we have $\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)-\mathbf{E}\left[B_{j,\lambda}(x)\right]\|_{\mathrm{op}}^4\right] \le \mathtt{A}_1 j^{-2} h_j^{-2}\log(j)^2$, which gives

$$\mathbf{E}\left[\sup_{x\in\Theta}\left\|B_{j,\lambda}(x)^{-1}\right\|_{\mathrm{op}}^4\right] \le 4\mathtt{A}_1\mu_{\min}^{-4} j^{-2} h_j^{-2d}\log(j)^2\lambda_j^{-4} + 4\lambda_{\min}^{-4} \enspace .$$

Since $j \ge \lambda^2 h_j^{-d}\log(j)$, we deduce that

$$\mathbf{E}\left[\sup_{x\in\Theta}\left\|B_{j,\lambda}(x)^{-1}\right\|_{\mathrm{op}}^4\right] \le 4(\mathtt{A}_1+1)\lambda_{\min}^{-4} \enspace .$$

We finish the proof by assigning $\mathtt{A}_2 = 4(\mathtt{A}_1+1)$. □

Lemma 5.16. *Let $j \in [n]$, with $1 \le j h_j^d$, and assume that Assumption 10 holds. Then, for any $x \in \Theta$, we have*

$$\mathbf{E}\left\|B_{j,\lambda}(x)-\mathbf{E}\left[B_{j,\lambda}(x)\right]\right\|_{\mathrm{op}}^4 \le \mathtt{A}_1 h_j^{-2d} j^{-2} \enspace .$$

Furthermore, for $j \ge \lambda_j^{-2} h_j^{-d}$, we have $\mathbf{E}\left\|B_{j,\lambda}(x)\right\|_{\mathrm{op}}^{-4} \le \mathtt{A}_2\mu_{\min}^{-4}$, where $\mathtt{A}_1, \mathtt{A}_2 > 0$ are numerical constants.

Proof. Let $Q_{i,j}(x) = h_j^{-d}M_{i,j}(x) - h_j^{-d}\mathbf{E}\left[B_{j,\lambda}(x)\right]$. We introduce

$$K' = \sup_{x\in\Theta}\max_{1\le i\le n}\|Q_{i,j}(x)\|_{\mathrm{op}}\,, \quad \text{and} \quad \upsilon^2 = \sup_{x\in\Theta}\left\|\sum_{i=1}^n \mathbf{E}Q_{i,j}^\top(x)Q_{i,j}(x)\right\|_{\mathrm{op}} \enspace . \tag{5.25}$$

Note that for any $i \in [j]$ and $x \in \mathbb{R}^d$, $Q_{i,j}(x) \in \mathbb{R}^{S\times S}$. Then by (Vershynin, 2019, Theorem 5.4.1), for any $t \ge 0$, we have

$$\mathbf{P}\left[\left\|\sum_{i=1}^j Q_{i,j}(x)\right\|_{\mathrm{op}} \ge t\right] \le 2S\exp\left(-c\min\left(\frac{t^2}{\upsilon^2},\frac{t}{K'}\right)\right),$$

where $c > 0$ is a numerical constant.

$$\begin{aligned}\mathbf{E}\left\|\sum_{i=1}^j Q_{i,j}(x)\right\|_{\mathrm{op}}^4 &= \int_0^\infty 4t^3\mathbf{P}\left[\left\|\sum_{i=1}^j \boldsymbol{Q}_{i,j}(x)\right\|_{\mathrm{op}} \ge t\right]\mathrm{d}t \\ &= \underbrace{4S\int_0^{\frac{\upsilon^2}{K'}} t^3\exp\left(-c\frac{t^2}{\upsilon^2}\right)\mathrm{d}t}_{\text{term I}} + \underbrace{4S\int_{\frac{\upsilon^2}{K'}}^\infty t^3\exp\left(-c\frac{t}{K'}\right)\mathrm{d}t}_{\text{term II}} \enspace .\end{aligned}$$

We provide upper bounds for the terms I and II, separately.

$$\begin{aligned}
\text{term I} &= 2S\frac{\upsilon^2}{c}\left(-t^2\exp\left(-c\frac{t^2}{\upsilon^2}\right)\right)\Bigg|_{t=0}^{\frac{\upsilon^2}{K'}} + 4S\frac{\upsilon^2}{c}\int_0^{\frac{\upsilon^2}{K'}} t\exp\left(-c\frac{t^2}{\upsilon^2}\right)\,\mathrm{d}t\\
&\leq 4S\frac{\upsilon^2}{c}\int_0^{\frac{\upsilon^2}{K'}} t\exp\left(-c\frac{t^2}{\upsilon^2}\right)\,\mathrm{d}t\\
&= 2S\frac{\upsilon^4}{c^2}\int_0^{\frac{\upsilon^2}{K'}} \frac{2ct}{\upsilon^2}\exp\left(-c\frac{t^2}{\upsilon^2}\right)\,\mathrm{d}t\\
&= -2S\frac{\upsilon^4}{c^2}\exp(-c\frac{t^2}{\upsilon^2})\Bigg|_{t=0}^{\frac{\upsilon^2}{K'}} \leq 2s\frac{\upsilon^4}{c^2} \ .
\end{aligned}$$

Similarly, for term II we can write

$$\begin{aligned}
\text{term II} &= -4S\frac{K'}{c}t^3\exp\left(-c\frac{t}{K'}\right)\Bigg|_{\frac{\upsilon^2}{K'}}^{\infty} + 12S\frac{K'}{c}\int_{\frac{\upsilon^2}{K'}}^{\infty} t^2\exp\left(-c\frac{t}{K'}\right)\,\mathrm{d}t\\
&= 4S\frac{\upsilon^6}{cK'^2}\exp\left(-c\frac{\upsilon^2}{K'^2}\right) + 12S\frac{K'}{c}\int_{\frac{\upsilon^2}{K'}}^{\infty} t^2\exp\left(-c\frac{t}{K'}\right)\,\mathrm{d}t\\
&\leq 4S\frac{\upsilon^4}{c^2} + 12S\frac{K'}{c}\int_{\frac{\upsilon^2}{K'}}^{\infty} t^2\exp\left(-c\frac{t}{K'}\right)\,\mathrm{d}t\\
&\leq 4S\frac{\upsilon^4}{c^2} - 12S\frac{K'^2}{c^2}t^2\exp\left(-c\frac{t}{K'}\right)\Bigg|_{\frac{\upsilon^2}{K'}}^{\infty} + 24S\frac{K'^2}{c^2}\int_{\frac{\upsilon^2}{K'}}^{\infty} t\exp(-c\frac{t}{K'})\,\mathrm{d}t\\
&\leq 4S\frac{\upsilon^4}{c^2} + 12S\frac{\upsilon^4}{c^2}\exp\left(-c\frac{\upsilon^2}{K'^2}\right) + 24S\frac{K'^2}{c^2}\int_{\frac{\upsilon^2}{K'}}^{\infty} t\exp(-c\frac{t}{K'})\,\mathrm{d}t\\
&= 4S\frac{\upsilon^4}{c^2} + 12S\frac{\upsilon^2K'^2}{c^3} - 24S\frac{K'^3}{c^3}t\exp\left(-c\frac{t}{K'}\right)\Bigg|_{\frac{\upsilon^2}{K'}}^{\infty} + 24S\frac{K'^3}{c^3}\int_{\frac{\upsilon^2}{K'}}^{\infty}\exp\left(-c\frac{t}{K'}\right)\,\mathrm{d}t\\
&\leq 4S\frac{\upsilon^4}{c^2} + 12s\frac{\upsilon^2K'^2}{c^3} + 24S\frac{K'^4}{c^4} + 24s\frac{K'^3}{c^3}\int_{\frac{\upsilon^2}{K'}}^{\infty}\exp\left(-c\frac{t}{K'}\right)\,\mathrm{d}t\\
&\leq 4S\frac{\upsilon^4}{c^2} + 12S\frac{\upsilon^2K'^2}{c^3} + 24S\frac{K'^4}{c^4} - 24S\frac{K'^4}{c^4}\exp\left(-c\frac{t}{K'}\right)\Bigg|_{\frac{\upsilon^2}{K'}}^{\infty}\\
&\leq 4S\frac{\upsilon^4}{c^2} + 12S\frac{\upsilon^2K'^2}{c^3} + 24S\frac{K'^4}{c^4} + 24S\frac{K'^6}{c^5\upsilon^2} \ .
\end{aligned}$$

By combining the provided bounds for the terms I and II, we deduce that

$$\mathbf{E}\left\|\sum_{i=1}^{j} Q_{i,j}(x)\right\|_{\mathrm{op}}^4 \leq 6S\frac{\upsilon^4}{c^2} + 12S\frac{\upsilon^2K'^2}{c^3} + 24S\frac{K'^4}{c^4} + 24S\frac{K'^6}{c^5\upsilon^2} \ . \tag{5.26}$$

To conclude the first part of the proof, it is enough to upper bound the terms K' and σ, which are defined in (5.25). For K', we can write

$$K' \leq \sup_{x\in\Theta} \max_{i\in[j]} 2h_j^{-d} \left\| U\left(\frac{x_i - x}{h_j}\right) U\left(\frac{x_i - x}{h_j}\right)^\top k\left(\frac{x_i - x}{h_j}\right) \right\|_{\text{op}} \leq \mathtt{A}_4 h_n^{-d} \ ,$$

where $\mathtt{A}_4 = \max_{u\in\text{Supp}(k)} \left\| U\left(u\right) U\left(u\right)^\top k\left(u\right) \right\|_{\text{op}}$. For υ^2, by Lemma 5.14(ii), we have

$$\begin{aligned}\upsilon^2 = \sup_{x\in\Theta} \sum_{i=1}^{j} \left\| \mathbf{E} Q_{i,j}^\top(x) Q_{i,j}(x) \right\|_{\text{op}} &\leq j h_j^{-2d} \sup_{x\in\Theta} \mathbf{E} \left\| U\left(\frac{x_1 - x}{h_j}\right) U\left(\frac{x_1 - x}{h_j}\right)^\top k\left(\frac{x_1 - x}{h_j}\right) \right\|_{\text{op}}^2 \\ &\leq p_{\max} \nu_{2,2} j h_j^{-d} \ .\end{aligned}$$

By substituting the above bounds in (5.26), we get

$$\mathbf{E} \left\| \sum_{i=1}^{j} Q_{i,j}(x) \right\|_{\text{op}}^4 \leq \mathtt{A}_4 \left[j^2 h_j^{-2d} + j h_n^{-3d} + h_j^{-4d} \right] \ ,$$

where $\mathtt{A}_4 > 0$ is a numerical constant. Since $\frac{1}{j}\sum_{i=1}^{j} Q_{i,j}(x) = B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right]$, we deduce that

$$\mathbf{E} \left\| B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right] \right\|_{\text{op}}^4 \leq \mathtt{A}_3 \left[j^{-2} h_j^{-2d} + j^{-3} h_j^{-3d} + j^{-4} h_j^{-4d} \right] \ .$$

Since $1 \leq j h_j^d$, we get

$$\mathbf{E} \left\| B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right] \right\|_{\text{op}}^4 \leq \mathtt{A}_1 h_j^{-2d} j^{-2} \ ,$$

with $\mathtt{A}_1 = 3\mathtt{A}_3$. For the second part the proof, we can write

$$\begin{aligned}\mathbf{E} \left\| B_{j,\lambda}(x)^{-1} \right\|_{\text{op}}^4 &\leq 4\mathbf{E} \left\| B_{j,\lambda}(x)^{-1} - \left(\mathbf{E}[B_{j,\lambda}(x)]\right)^{-1} \right\|_{\text{op}}^4 + \frac{4}{\mu_{\min}^4} \\ &\leq \frac{4\mathtt{A}_1}{\lambda^4 \mu_{\min}^4} h_j^{-4d} j^{-2} + \frac{4}{\mu_{\min}^4} \ .\end{aligned}$$

So for any $j \geq \lambda_j^{-2} h_j^{-d}$, we have

$$\mathbf{E} \left\| B_{j,\lambda}(x)^{-2} \right\|_{\text{op}} \leq \frac{\mathtt{A}_2}{\mu_{\min}^4},$$

where we introduced $\mathtt{A}_2 = 4\mathtt{A}_1 + 4$. □

Lemma 5.17. *Let $j \in [n]$, with $1 \leq jh_j^d$, and assume that Assumptions 9(iv) and 10 hold. Then, for any $x \in \Theta$, we have*

$$\mathbf{E}\left[\|C_j(x) - \mathbf{E}\left[C_j(x)\right]\|^4\right] \leq \mathtt{A}h_j^{-2d}j^{-2} \ .$$

Proof. The proof can be deduced by similar steps as the proof of Lemma 5.16. □

Lemma 5.18. *Let $j \in [n]$, with $1 \leq jh_j^d$, and assume that Assumptions 9(iv) and 10 hold. Then, for any $x \in \Theta$, we have*

$$\mathbf{E}\left[\sup_{x\in\Theta}\|C_j(x) - \mathbf{E}\left[C_j(x)\right]\|^4\right] \leq \mathtt{A}h_j^{-2d}j^{-2}\log(j)^2 \ .$$

Proof. The proof can be deduced by similar steps as the proof of Lemma 5.15. □

Lemma 5.19. *Let Assumption 10 holds. If $1 \leq jh_j^d$, and $j \geq \lambda_j^{-2}h_j^{-d}$, then for any $x \in \Theta$, we have*

$$\mathbf{E}\left[\sup_{x\in\Theta}\left\|B_{j,\lambda}(x)^{-1} - \left(\mathbf{E}[B_{j,\lambda}(x)]\right)^{-1}\right\|_{\mathrm{op}}\sup_{x\in\Theta}\|C_j(x) - \mathbf{E}\left[C_j(x)\right]\|\right] \leq \mathtt{A}j^{-1}h_j^{-d} \ ,$$

where $\mathtt{A} > 0$ is a numerical constant.

Proof. By using Cauchy-Schwarz inequality, it is enough to provide an upper bound for the following terms:

$$\Bigg(\underbrace{\mathbf{E}\left[\sup_{x\in\Theta}\left\|B_{j,\lambda}(x)^{-1} - \left(\mathbf{E}[B_{j,\lambda}(x)]\right)^{-1}\right\|^2_{\mathrm{op}}\right]}_{\text{term I}}\underbrace{\mathbf{E}\left[\sup_{x\in\Theta}\|C_j(z_n) - \mathbf{E}\left[C_j(x)\right]\|^2\right]}_{\text{term II}}\Bigg)^{\frac{1}{2}} \ .$$

For term I, we utilize Cauchy-Schwarz inequality, once more, and we get

$$\text{term I} \leq \mu_{\min}^{-2}\left(\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x)\|^{-4}_{\mathrm{op}}\right]\mathbf{E}\left[\sup_{x\in\Theta}\|B_{j,\lambda}(x) - \mathbf{E}\left[B_{j,\lambda}(x)\right]\|^4_{\mathrm{op}}\right]\right)^{\frac{1}{2}} \leq \mathtt{A}_1 j^{-1}h_j^{-d}\log(j) \ ,$$

where the last inequality is obtained by Lemma 5.15, , and $\mathtt{A}_1 > 0$ is a numerical constant. For term II, Lemma 5.18, yields

$$\text{term II} \leq \left(\mathbf{E}\left[\|C_j(x) - \mathbf{E}\left[C_j(x)\right]\|^2\right]\right) \leq \mathtt{A}_2 j^{-1}h_j^{-d}\log(j) \ ,$$

where we introduced $\mathtt{A}_2 > 0$ as the numerical constant that appears in Lemma 5.18. By combining all of these, we conclude the proof. □

Chapter 6

Conclusions and Future Work

Chapter 2 provided a thorough analyses of the generalization properties of weighted kernel ridge regression under the covariate shift assumption. Our main objectives were to:

(i) Investigate the properties of IW correction applied to high capacity models.

(ii) Analyze the relative merits of high-capacity models over low-capacity models under covariate shifts.

(iii) Derive alternative re-weighting procedures allowing optimally tackling hard shift scenarios.

We gave a fairly satisfactory explanation for each of the above-mentioned points.

For the moderate covariate shift scenarios (when IW is bounded by W) we show that the kernel least squares corrected by the importance weights is optimal and matches the learning rates of KRR without covariate shift. Moreover, the lower bound on the minimax risk over the Sobolev class of regression functions in Ma et al. (2022) suggests that the appearance of the constant $W = \|w\|_\infty$ in our upper rate of IW-KRR is also optimal (at least asymptotically). The generalization result from Ma et al. (2022) shows that the same optimal rates can be achieved by unweighted KRR whenever the regression function belongs to the RKHS i.e. the model is wellspecified. In this case, we can safely forget about the covariate shift and construct the KRR with a suitably chosen kernel. In practice, however, full KRR is rarely an option due to scalability issues. The main contributions that enabled scalability are based on the Nyström approximation (Williams and Seeger, 2000) and the random feature approximation (Rahimi and Recht, 2008). Both approximations are essentially low-rank approximations of the nonparametric kernel-based model. Under covariate shift we have two competing issues. We would like to fit exact (full-rank) kernel ridge regression in order to avoid IW correction. On the other hand, for the low-rank approximations, used to speed up the matrix inversion, we need IW correction to avoid the bias related to the mismatch between the projections of the regression function under training and testing measures. Giving the precise trade-off between the model capacity and the deviation from the importance weighting strategy is a prominent future direction.

For the more severe covariate shifts (when IW is integrable, but not bounded) we show the minimax optimality of KRR corrected by the clipped importance weights. To the best of my knowledge, the question of optimality of unclipped IW is open. The main technical difficulty to achieve the generalization bounds is that unclipped IW requires considering unbounded random variables where the exponential tail inequalities (like Bernstein or Hoeffding inequalities) are no longer applicable. This problem was observed earlier in Cortes et al. (2010) where the learning bound was obtained for finite VC classes. In this work, however, the optimality of obtained rates was not discussed.

Beyond KRR, it would be interesting to study other methods of regression function estimation.

Chapter 3 addressed the learning problem under the target shift. In a way, the situation here is simpler than for the covariate shift. The main takeaway message was that the IW correction is the only reasonable approach to handle the distributional shift in the output space. Deviation from the IW correction strategy leads to the irreducible bias term related to the mismatch between training and testing regression functions.

Chapter 4 developed a novel framework to localize Gaussian processes. The method was inspired by the local methods of regression where we fit a different but simple models separately at each query point x_0. This is done by applying a localisation operation used to down-weight contributions from input points that are far from the given test point x_0. This localization is achived via a kernel k_h, which assigns a weights to the training points based on its distance from x_0. A parameter h dictates the width of the neighbourhood. The form of the introduced localized GP maintains positive definiteness of the covariance, and it allows for considerable speedups compared to standard global GPR due to the sparsification effect of the Gram matrix. Let us mention some advantages and disadvantages of this approach.

- **Sparse Gram matrix.** Compared to global GPs, in the local formulation, in order to compute the posterior mean and variance, we need to invert s_0 by s_0 Gram matrix, where s_0 is the cardinality of active training set. For each target point x_0 the complexity of performing matrix inversion is $\mathcal{O}(s_0)$.

- **Spatial Adaptivity.** Besides scalability issues, Gaussian process regression as well as other fixed spatial scale methods (e.g.fixed-bandwidth local kernel methods, linear spline smoothers) known to have pure performance on the datasets where the smoothness of the underling regression function varies over the input space. We have shown that the local GPs are able to infer the underlying latent functions and improve regression performance when the datasets are non-stationary; this is achieved by choosing the localization parameter in a location dependent way.

- **Difficulty of using gradient based methods.** The proposed method requires cross-validation to tune the scale parameter h of localizing kernel, while other GP-based techniques use a less expensive marginal log-likelihood gradient optimization to tune these types of parameters. We found MLL gradient optimization problematic because of the non-smoothness of local kernel with respect to the scale parameter, which in turn makes MLL function non-differentiable with respect to this parameter.

Chapter 5 introduced two novel algorithms to estimate the location and size of the strongly convex and smooth regression function. These algorithms are constructed for the passive design framework, i.e. for the case when the points x_i are random and independent. The main point of the work is to show that estimating the location and size of the optimum are as difficult as estimating the gradient and function value at a *fixed* point. More precisely, we provide tight upper and lower bounds for the performance of proposed estimators.

An important future work is to make our algorithms adaptive to the unknown smoothness β. In other words, designing theoretically sound and data driven procedures to determine the optimal smoothing parameter h and regularization parameter λ. Such adaptive estimates can be obtained using the so-called Lepski method. When considering adaptation to the unknown smoothness of the function, the optimal rates for the estimation are slower than the minimax rates by a logarithmic factor.

Bibliography

A. Akhavan, M. Pontil, and A. B. Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

A. Akhavan, M. Pontil, and A. B. Tsybakov. Distributed zero-order optimization under adversarial noise. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

E. Arias-Castro, W. Qiao, and L. Zheng. Estimation of the global mode of a density: Minimaxity, adaptation, and computational complexity. *Electronic Journal of Statistics*, 16(1):2774–2795, 2022.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

A. Asuncion and D. J. Newman. UCI machine learning repository, 2007.

K. Azizzadenesheli, A. Liu, F. Yang, and A. Anandkumar. Regularized learning for domain adaptation under label shifts. *International Conference on Learning Representations*, 2019.

F. Bach and V. Perchet. Highly-smooth zero-th order online optimization. In *Proc. 29th Annual Conference on Learning Theory*, pages 1–27, 2016.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

E. Belitser, S. Ghosal, and H. van Zanten. Optimal two-stage procedures for estimating location and size of the maximum of a multivariate regression function. *Ann. Statist.*, 40(6):2850–2876, 2012.

E. Belitser, S. Ghosal, and H. van Zanten. Correction note: ”Optimal two-stage procedures for estimating location and size of the maximum of a multivariate regression function”. *Ann. Statist.*, 49(1):612–613, 2021.

M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

E. Blanzieri and F. Melgani. An adaptive SVM nearest neighbor classifier for remotely sensed imagery. In *2006 IEEE International Symposium on Geoscience and Remote Sensing*, pages 3931–3934. IEEE, 2006.

E. Blanzieri and F. Melgani. Nearest neighbor classification of remote sensing images with the maximal margin principle. *IEEE Transactions on geoscience and remote sensing*, 46(6):1804–1811, 2008.

J. R. Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

L. Bottou and V. Vapnik. Local learning algorithms. *Neural computation*, 4(6):888–900, 1992.

J. Byrd and Z. Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR, 2019.

A. Caponnetto. Optimal rates for regularization operators in learning theory. Technical report, Massachusetts inst of tech Cambridge computer science and artificial intelligence lab, 2006.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

A. Carpentier, O. Collier, L. Comminges, A. Tsybakov, and Y. Wang. Minimax rate of testing in sparse linear regression. *Automation and Remote Control*, 80:1817–1834, 2019.

L. Carratino, S. Vigogna, D. Calandriello, and L. Rosasco. Park: Sound and efficient kernel ridge regression by feature space partitions. *Advances in Neural Information Processing Systems*, 34, 2021.

H. Chen. Lower rate of convergence for locating a maximum of a function. *The Annals of Statistics*, pages 1330–1334, 1988.

H. Cheng, P.-N. Tan, and R. Jin. Localized support vector machine and its efficient algorithm. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 461–466. SIAM, 2007.

H. Chernoff. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16(1):31–41, 1964.

C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.

C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computationals Mathematics*, page 413–428, 2002a.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002b.

F. Cucker and D. X. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

K. Cutajar, M. Osborne, J. Cunningham, and M. Filippone. Preconditioning Kernel Matrices. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2529–2538. JMLR.org, 2016.

K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. Random feature expansions for deep Gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 884–893, International Convention Centre, Sydney, Australia, Aug. 2017a. PMLR.

K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. Random feature expansions for deep Gaussian processes. In *International Conference on Machine Learning*, pages 884–893. PMLR, 2017b.

T. Dalenius. The mode–a neglected statistical parameter. *Journal of the Royal Statistical Society. Series A (General)*, pages 110–117, 1965.

S. Dasgupta and S. Kpotufe. Optimal rates for k-nn density and mode estimation. *Advances in Neural Information Processing Systems*, 27, 2014.

E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5(1):59–85, 2005a.

E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, F. Odone, and P. Bartlett. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(5), 2005b.

R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. Mathematical methods for supervised learning. *IMI Preprints*, 22:1–51, 2004.

R. A. DeVore. Nonlinear approximation. *Acta numerica*, 7:51–150, 1998.

L. Devroye, L. Györfi, and A. Krzyżak. The Hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227, 1998.

J. Dippon. Accelerated randomized stochastic optimization. *The Annals of Statistics*, 31(4):1260–1281, 2003.

V. Dupač. O kiefer-wolfowitzově aproximační methodě. *Časopis pro pěstování matematiky*, 82(1):47–75, 1957.

V. Fabian. Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics*, pages 191–200, 1967.

M. R. Facer and H.-G. Müller. Nonparametric estimation of the location of a maximum in a response surface. *Journal of Multivariate Analysis*, 87(1):191–217, 2003.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Routledge, 2018.

M. Filippone and R. Engler. Enabling scalable stochastic gradient-based inference for Gaussian processes by employing the Unbiased LInear System SolvEr (ULISSE). In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, July 6-11, 2015*, pages 1015–1024, 2015.

S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *J. Mach. Learn. Res.*, 21:205–1, 2020.

S. Garg, Y. Wu, S. Balakrishnan, and Z. Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020.

R. B. Gramacy and D. W. Apley. Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2): 561–578, 2015.

U. Grenander. Some direct estimates of the mode. *The Annals of Mathematical Statistics*, pages 131–138, 1965.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

L. Györfi, M. Kohler, A. Krzyzak, H. Walk, et al. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

W. Härdle and R. Nixdorf. Nonparametric sequential estimation of zeros and extrema of regression functions. *IEEE transactions on information theory*, 33(3):367–372, 1987.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 282–290. AUAI Press, 2013.

I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation, Asymptotic Theory.* Springer, New York, 1981.

I. A. Ibragimov and R. Z. Khas'minskii. Estimation of the maximum value of a signal in gaussian white noise. *Mat. Zametki*, 32(4):746–750, 1982.

J. Jiang and C. Zhai. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 264–271, 2007.

M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

C. G. Kaufman, M. J. Schervish, and D. W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, 2008.

R. Z. Khas'minskii. Lower bound for the risks of nonparametric estimates of the mode. *Contributions to statistics*, pages 91–97, 1979.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

J. Klemelä. Adaptive estimation of the mode of a multivariate density. *Journal of Nonparametric Statistics*, 17(1):83–105, 2005.

S. Kpotufe. Lipschitz density-ratios, structured data, and data-driven tuning. In *Artificial Intelligence and Statistics*, pages 1320–1328. PMLR, 2017.

S. Kpotufe and G. Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

K. Krauth, E. V. Bonilla, K. Cutajar, and M. Filippone. AutoGP: Exploring the capabilities and limitations of Gaussian process models. In *Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, August 11-15, 2017, Sydney, Australia*, 2017.

V. Krishnamurthy and G. Yin. Multikernel passive stochastic gradient algorithms and transfer learning. *IEEE Trans. Automat. Control*, 67:1792–1805, 2022.

M. Lázaro-Gredilla, J. Quinonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal. Sparse Spectrum Gaussian Process Regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.

O. V. Lepski. Estimation of the maximum of a nonparametric signal up to a constant. *Theory Probab. Appl.*, 38:152–158, 1993.

O. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

C. Ma, R. Pathak, and M. J. Wainwright. Optimally tackling covariate shift in rkhs-based nonparametric regression. *arXiv preprint arXiv:2205.02986*, 2022.

D. J. MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the rényi divergence. UAI '09, page 367–374. AUAI Press, 2009a.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009b.

F. Meier, P. Hennig, and S. Schaal. Incremental local Gaussian regression. *Advances in Neural Information Processing Systems*, 27, 2014a.

F. Meier, P. Hennig, and S. Schaal. Local Gaussian regression. *arXiv preprint arXiv:1402.0645*, 2014b.

M. Meister and I. Steinwart. Optimal learning rates for localized SVMs. *The Journal of Machine Learning Research*, 17(1):6722–6765, 2016.

S. Mendelson and J. Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

A. Mokkadem and M. Pelletier. A companion for the kiefer–wolfowitz–blum stochastic approximation algorithm. *The Annals of Statistics*, 35(4):1749–1772, 2007.

N. Mücke. Reducing training time by efficient localized kernel regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2603–2610. PMLR, 2019.

H.-G. Müller. Kernel estimators of zeros and of location and size of extrema of regression functions. *Scandinavian journal of statistics*, pages 221–232, 1985.

H.-G. Müller. Adaptive nonparametric peak estimation. *The Annals of Statistics*, pages 1053–1069, 1989.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9 (1):141–142, 1964.

A. Nazin, B. Polyak, and A. Tsybakov. Passive stochastic approximation. *Automat. Remote Control*, 50:1563–1569, 1989.

A. Nazin, B. Polyak, and A. Tsybakov. Optimal and robust algorithms of passive stochastic approximation. *IEEE Transactions on Information Theory*, 38:1577–1583, 1992.

R. M. Neal. *Bayesian Learning for Neural Networks (Lecture Notes in Statistics)*. Springer, 1 edition, Aug. 1996.

V. Novitskii and A. Gasnikov. Improved exploiting higher order smoothness in derivative-free optimization and continuous bandit. *arXiv preprint arXiv:2101.03821*, 2021.

E. Parzen. An approach to time series analysis. *The Annals of Mathematical Statistics*, 32(4):951–989, 1961.

E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962a.

E. Parzen. Extraction and detection problems and reproducing kernel hilbert spaces. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1 (1):35–62, 1962b.

E. Parzen. Probability density functionals and reproducing kernel hilbert spaces. In *Proceedings of the Symposium on Time Series Analysis*, volume 196, pages 155–169. Wiley, New York, 1963.

R. Pathak, C. Ma, and M. Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pages 17517–17530. PMLR, 2022.

J. Pennington, F. X. X. Yu, and S. Kumar. Spherical random features for polynomial kernels. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

B. T. Polyak and A. B. Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problems of Information Transmission*, 26(2):45–53, 1990.

D. Precup, R. S. Sutton, and S. P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766, 2000.

F. Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

R. Pytlak. *Conjugate gradient algorithms in nonconvex optimization*, volume 89. Springer Science & Business Media, 2008.

J. Quinonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

A. Rudi and L. Rosasco. Generalization properties of learning with random features. In *NIPS*, pages 3215–3225, 2017.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. ICML'12, page 459–466. Omnipress, 2012.

N. Segata and E. Blanzieri. Fast and scalable local kernel machines. *Journal of Machine Learning Research*, 11(6), 2010.

N. Segata, E. Pasolli, F. Melgani, and E. Blanzieri. Local SVM approaches for fast and accurate classification of remote-sensing images. *International Journal of Remote Sensing*, 33(19):6186–6201, 2012.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, New York, NY, USA, 2004.

D. Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524, 1968.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

S. Smale and D.-X. Zhou. Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, 41(3):279–305, 2004.

S. Smale and D.-X. Zhou. Shannon sampling ii: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

E. Snelson and Z. Ghahramani. Local and global sparse Gaussian process approximations. In *Artificial Intelligence and Statistics*, pages 524–531. PMLR, 2007.

I. Steinwart and A. Christmann. *Support vector machines.* Springer Science & Business Media, 2008.

I. Steinwart, D. R. Hush, C. Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

A. Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009.

R. Tandon, S. Si, P. Ravikumar, and I. Dhillon. Kernel ridge regression via partitioning. *arXiv preprint arXiv:1608.01976*, 2016.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, volume 5 of *JMLR Proceedings*, pages 567–574. JMLR.org, 2009.

N. Tripuraneni, B. Adlam, and J. Pennington. Overparameterization improves robustness to covariate shift in high dimensions. *Advances in Neural Information Processing Systems*, 34, 2021.

A. B. Tsybakov. Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, pages 69–84, 1986.

A. B. Tsybakov. Recursive estimation of the mode of a multivariate distribution. *Problems of Information Transmission*, pages 31–37, 1990a.

A. B. Tsybakov. Locally-polynomial algorithms of passive stochastic approximation. *Problems of Control and Information Theory*, pages 181–195, 1990b.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, New York, 2009.

V. N. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1998.

J. Venter. On estimation of the mode. *The Annals of Mathematical Statistics*, pages 1446–1455, 1967.

R. Vershynin. High-dimensional probability. *Cambridge University Press*, 2019.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

J. Wen, C.-N. Yu, and R. Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, pages 631–639. PMLR, 2014.

F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6): 80–83, 1945. ISSN 00994987.

C. K. I. Williams and M. Seeger. The Effect of the Input Density Distribution on Kernel-based Classifiers. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1159–1166, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep Kernel Learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 370–378, Cadiz, Spain, May 2016. PMLR.

D. Xu, Y. Ye, and C. Ruan. Understanding the role of importance weighting for deep learning. *arXiv preprint arXiv:2103.15209*, 2021.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

W. W. Yoo and S. Ghosal. Bayesian mode and maximum estimation and accelerated rates of contraction. *Bernoulli*, 25(3):2330–2358, 2019.

C. Zhang, L. Zhang, and J. Ye. Generalization bounds for domain adaptation. *Advances in neural information processing systems*, 4:3320, 2012.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. PMLR, 2013.

T. Zhang. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems*, volume 15, pages 454–461, 2002.

T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3): 739–767, 2002.