# Reproducing-Kernel Hilbert Space Regression with Notes on the Wasserstein Distance

Stephen Page

Lancaster University

Submitted for the degree of Doctor of Philosophy at Lancaster University.

September 2019

STOR-i
excellence with impact

# Abstract

We study kernel least-squares estimators for the regression problem subject to a norm constraint. We bound the squared $L^2$ error of our estimators with respect to the covariate distribution. We also bound the worst-case squared $L^2$ error of our estimators with respect to a Wasserstein ball of probability measures centred at the covariate distribution. This leads us to investigate the extreme points of Wasserstein balls.

In Chapter 3, we provide bounds on our estimators both when the regression function is unbounded and when the regression function is bounded. When the regression function is bounded, we clip the estimators so that they are closer to the regression function. In this setting, we also use training and validation to adaptively select a size for our norm constraint based on the data.

In Chapter 4, we study a different adaptive estimation procedure called the Goldenshluger–Lepski method. Unlike training and validation, this method uses all of the data to create estimators for a range of sizes of norm constraint before using pairwise comparisons to select a final estimator. We are able to adaptively select both a size for our norm constraint and a kernel.

In Chapter 5, we examine the extreme points of Wasserstein balls. We show that the only extreme points which are not on the surface of the ball are the Dirac measures. This is followed by finding conditions under which points on the surface of the ball

are extreme points or not extreme points.

In Chapter 6, we provide bounds on the worst-case squared $L^2$ error of our estimators with respect to a Wasserstein ball of probability measures centred at the covariate distribution. We prove bounds both when the regression function is unbounded and when the regression function is bounded. We also investigate the analysis and computation of alternative estimators.

# Acknowledgements

There are many people who I need to thank for making this thesis possible. The most important person is my main supervisor, Steffen Grünewälder, for his help and guidance over the past four years. My other supervisors Nicos Pavlidis and David Leslie also provided useful feedback along the way.

My family have been incredibly supportive. My parents Liz and Anthony deserve particular recognition for looking after me whenever I am home, especially around Christmas. My brother Adam is always up for a game, whether physical or virtual. My sister Lyndsey is consistently hilarious, particularly when chasing around after her daughter Naomi, who has been alive for less time than I have been working on this thesis!

Katharine is wonderful. I'm sorry for occasionally working into the early hours of the morning. Thank you for taking me out on walks the day after. You are great.

Someone I very much need to thank is my long-term flatmate-then-housemate Chrissy. Thank you for being so accommodating, and for your insights on architecture and wedding dresses. My previous housemate Kasia is good for a long and often slightly weird chat. Sam is an excellent source of puns and introduced me to aubergine pickle. Gabi teaches me about Brazil and can thankfully sometimes make the puns stop!

I've met lots of other fantastic people in Lancaster. My friend Malika is excellent at combining household renovations with musical interludes. Craig can often be found defending the final frontier, as well as pensions. Dee is a fantastic cook, and might even give you a go on her Nintendo DS if you behave yourself.

I've also managed to keep in touch with a few people who aren't based in Lancaster. I need to thank my friend Sally for making Katharine and I feel so included in her wedding. Bel has good taste in tea, as well as most other things. Ren understands about chocolate, and sends me birthday cards by airmail.

Thank you to everyone at STOR-i for creating such a friendly atmosphere in which to work. My cohort have been incredibly kind and generous. When not studying, Maths Learning Development has been an excellent place to work. It has been extremely rewarding to support students and to see them improve. Thank you to Lancaster Labour Party for the interesting people and the lively discussions.

# Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Chapter 3 has been accepted for publication by the Journal of Machine Learning Research.

Chapter 4 was submitted to Bernoulli in November 2018.

The length of this thesis is 39651 words.

Stephen Page

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This thesis is primarily concerned with regression. The regression problem has a long history in statistics. It is the primary tool for understanding the relationship between different variables. We consider a random data set consisting of i.i.d. (independent and identically distributed) data points which come in pairs. Each pair consists of a covariate and a response variable. The response variables take real values, while the covariates may take values in any set. This set is known as the covariate set. In regression, we model response variables as noisy observations of a function of the covariates. This function is known as the regression function. A formal definition of our regression problem is given in Subsection 1.1.1.

Our aim is to estimate the regression function, and we are interested in showing that the error of our estimators with respect to the regression function is small. In particular, in this thesis, we consider the analysis of kernel estimators. Kernel estimators are defined as random elements of a reproducing-kernel Hilbert space (RKHS) with a small empirical error. These are different to, for example, kernel estimators in density estimation. An RKHS is a Hilbert space of functions with additional properties which are discussed in Subsection 1.1.2. The empirical error is a proxy for the actual error of the estimator which is based on the data.

Since RKHSs are usually selected so that they are infinite-dimensional, it is necessary to regularise the estimator in some way to prevent overfitting. Overfitting occurs when the estimator is too close to the response variables at the covariates. This results in the estimator being too dependent on the variability of the response variables compared to the regression function, resulting in the estimator having a large error.

Overfitting is often prevented by adding a regularisation function to the empirical error. We can then define an estimator as the minimiser of the resulting linear combination. This is known as Tikhonov regularisation, and is used in the definition of, for example, support vector machines (SVMs). However, in this thesis, we minimise the empirical error subject to a constraint on the regularisation function. This is known as Ivanov regularisation. In particular, we consider the case in which the regularisation function is equal to the norm of the RKHS. We then obtain estimators with a bounded RKHS norm. This is key to our analysis of these estimators.

We are mostly interested in bounding the squared $L^2(P)$ error of our estimators, where $P$ is the distribution of the covariates. We consider this error because it is the expected squared error of the estimator for an expectation over a new independent covariate, with the same distribution $P$. The empirical version of the squared $L^2(P)$ error is the sum of squares between the response variables and the estimator evaluated at the covariates. We refer to our estimators as Ivanov-regularised least-squares estimators when applying this empirical error.

We also consider ways of bounding the worst-case squared $L^2(Q)$ error of an estimator over all $Q$ in a ball of probability measures centred at $P$. This error is the worst-case expected squared error of the estimator for an expectation over a new independent covariate generated by a different distribution $Q$. The distribution $Q$ can be any perturbation of $P$ of any size up to the radius of the ball. We refer to the situation in which a new covariate is generated by a perturbation of $P$ as a covariate shift.

We define the ball of probability measures above using the Wasserstein distance from optimal transport. The optimal transport problem seeks to minimise the transport cost between two probability measures over the set of possible transport plans. Since the Wasserstein distance is determined by a cost function on the covariate set, information about the cost between two points is transferred to the distance between two probability measures. An important example is given by setting the cost function equal to some metric on the covariate set. The Wasserstein distance also arises naturally in the analysis of our Ivanov-regularised least-squares estimators.

When bounding the worst-case squared $L^2(Q)$ error, we also investigate the analysis and computation of estimators other than our Ivanov-regularised least-squares estimators. We define these alternative estimators using an empirical version of the worst-case squared $L^2(Q)$ error. In the empirical version, we centre the Wasserstein ball at $P_n$, the empirical distribution of the covariates. We show that, under suitable conditions, the empirical version of the worst-case squared $L^2(Q)$ error is attained at some $Q$ which is an extreme point of the Wasserstein ball centred at $P_n$. This motivates us to examine the extreme points of Wasserstein balls.

## 1.1 Key Concepts

We now define the key concepts which arise in this thesis. These are regression, RKHSs and their interpolation spaces, and optimal transport.

### 1.1.1 Regression

We give a formal definition of our regression problem. For a topological space $T$, let $\mathcal{B}(T)$ be the Borel $\sigma$-algebra of $T$. Let $(S, \mathcal{S})$ be a measurable space. Assume that $(X_i, Y_i)$ for $1 \le i \le n$ are $(S \times \mathbb{R}, \mathcal{S} \otimes \mathcal{B}(\mathbb{R}))$-valued random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which are i.i.d. with $X_i \sim P$ and $\mathbb{E}(Y_i^2) < \infty$. Here, $\mathbb{E}$ denotes integration with respect to $\mathbb{P}$. In this scenario, the $X_i$ are the covariates and the $Y_i$ are the response variables. We often refer to $S$ as the covariate set and to $P$, the law of the $X_i$, as the covariate distribution.

Recall the Kolmogorov definition of conditional expectation, defined using the Radon–Nikodym derivative. Since any version of $\mathbb{E}(Y_i|X_i)$ is $\sigma(X_i)$-measurable, where $\sigma(X_i)$ is the $\sigma$-algebra generated by $X_i$, we have that $\mathbb{E}(Y_i|X_i) = g(X_i)$ almost surely for some measurable function $g : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. This result can be found, for example, in Section A3.2 of Williams (1991). Since the $(X_i, Y_i)$ are identically distributed, we have, for any $A \in \mathcal{S}$, that $\mathbb{E}(Y_i \mathbb{1}(X_i \in A))$ is the same for all $1 \le i \le n$. Hence, from the definition of conditional expectation, we can choose $g$ to be the same for all $1 \le i \le n$. Since $\mathbb{E}(Y_i^2) < \infty$, it follows that $g \in L^2(P)$ by Jensen's inequality. The function $g$ is the regression function. Sometimes we assume that the regression function is bounded, so that $\|g\|_\infty \le C$ for $C > 0$.

In order to analyse estimators of the regression function $g$, we need to ensure that the response variables do not vary too much. With this in mind, we always assume that $\text{var}(Y_i|X_i) \le \sigma^2$ almost surely for $1 \le i \le n$ and $\sigma > 0$. However, sometimes this assumption does not give us enough control over the response variables. For example, this is the case when we require high-probability bounds on the error of an estimator. For an estimator $\hat{g}$ of the regression function $g$, we are usually interested in its squared

$L^2(P)$ error

$$\|\hat{g} - g\|_{L^2(P)}^2 = \int (\hat{g} - g)^2 \, dP.$$

We can assume reduced variability in the response variables using the concept of subgaussianity. A random variable is subgaussian if it is at least as concentrated around 0 as some normal distribution with mean 0. Let $U$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\sigma > 0$. Then $U$ is $\sigma^2$-subgaussian if $\mathbb{E}(\exp(tU)) \leq \exp(\sigma^2 t^2/2)$ for all $t \in \mathbb{R}$. The concentration of $U$ around 0 follows from Chernoff bounding. We can also define a conditional version of subgaussianity. Let $U$ and $V$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Then $U$ is $\sigma^2$-subgaussian given $V$ if $\mathbb{E}(\exp(tU)|V) \leq \exp(\sigma^2 t^2/2)$ almost surely for all $t \in \mathbb{R}$. We can then obtain reduced variability in the response variables by assuming $Y_i - g(X_i)$ is $\sigma^2$-subgaussian given $X_i$ for $1 \leq i \leq n$. This assumption implies our previous assumption that $\text{var}(Y_i|X_i) \leq \sigma^2$ almost surely for $1 \leq i \leq n$.

### 1.1.2 RKHSs and Their Interpolation Spaces

An RKHS is a space of function on a given set, with additional properties which we describe below. We consider RKHSs on the covariate set $S$ for our regression problem. We could assume that the regression function lies in a given RKHS. However, it is more realistic to assume that the regression function lies in some larger space between $L^2(P)$ and the RKHS. We define spaces between $L^2(P)$ and an RKHS using interpolation spaces.

Recall that a Hilbert space is a complete inner-product space. A Hilbert space $H$ of real-valued functions on $S$ is an RKHS if the evaluation functional $L_x : H \to \mathbb{R}$ by $L_x h = h(x)$ is bounded for all $x \in S$. This is equivalent to $L_x \in H^*$ the dual of $H$. By the Riesz representation theorem, $H^*$ is isomorphically isometric to $H$. Therefore, there is some $k_x \in H$ such that $h(x) = \langle h, k_x \rangle_H$ for all $h \in H$. Define $k : S \times S \to \mathbb{R}$

by $k(x_1, x_2) = \langle k_{x_1}, k_{x_2} \rangle_H$ for $x_1, x_2 \in S$. The function $k$ is the reproducing kernel of $H$. The kernel is symmetric and positive-definite.

For our regression problem, we assume that the regression function $g \in L^2(P)$. It is much more restrictive to assume $g \in H$, however it is reasonable to assume something between these two conditions. To do this, we use interpolation spaces. A detailed account of interpolation spaces is given by Bergh and Löfström (1976), however our definitions more closely follow Smale and Zhou (2003). Let $(Z, \|\cdot\|_Z)$ be a Banach space and $(V, \|\cdot\|_V)$ be a dense subspace of $Z$. The $K$-functional of $(Z, V)$ is

$$K(z,t) = \inf_{v \in V} (\|z - v\|_Z + t\|v\|_V)$$

for $z \in Z$ and $t > 0$. It follows quickly from this definition that $K(z, t)$ as a function of $t > 0$ is bounded by $\|z\|_Z$, non-decreasing and continuous. Furthermore, $K(z, t) \to 0$ as $t \to 0$ since $V$ is dense in $Z$. We can use the $K$-functional to define our interpolation spaces.

Let $\beta \in (0, 1)$ and $1 \leq q < \infty$. We first define the norms of the interpolation spaces by

$$\|z\|_{\beta,q} = \left( \int_0^\infty (t^{-\beta} K(z,t))^q t^{-1} dt \right)^{1/q} \text{ and } \|z\|_{\beta,\infty} = \sup_{t>0} (t^{-\beta} K(z,t))$$

for $z \in Z$. The interpolation space $[Z, V]_{\beta,q}$ is then defined to be the set of $z \in Z$ such that $\|z\|_{\beta,q} < \infty$. From the definition of the norms, we find that smaller values of $\beta$ give larger spaces. The space $[Z, V]_{\beta,q}$ is not much larger than $V$ when $\beta$ is close to 1, but we obtain spaces which get closer to $Z$ as $\beta$ decreases. For a fixed $\beta$, the largest interpolation space is given by $q = \infty$.

If $z \in [Z, V]_{\beta,\infty}$, then we know how well we can approximate $z$ using elements of $V$.

Theorem 3.1 of Smale and Zhou (2003) shows that

$$\inf\{\|v - z\|_Z : v \in V, \|v\|_V \leq r\} \leq \frac{\|z\|_{\beta,q}^{1/(1-\beta)}}{r^{\beta/(1-\beta)}}. \tag{1.1.1}$$

The authors only consider the case in which $\|v\|_Z \leq \|v\|_V$ for all $v \in V$, however the result holds by the same proof even without this condition. We can apply this approximation result to the interpolation spaces between $L^2(P)$ and $H$.

When the RKHS $H$ is dense in $L^2(P)$, we can define the interpolation spaces $[L^2(P), H]_{\beta,q}$ for $\beta \in (0,1)$ and $1 \leq q \leq \infty$. In particular, we consider $[L^2(P), H]_{\beta,\infty}$. In order to understand how well the regression function $g$ can be approximated by elements of $H$, we define

$$I_2(g, r) = \inf\left\{\|h_r - g\|_{L^2(P)}^2 : h_r \in rB_H\right\}$$

for $r > 0$. If we assume $g \in [L^2(P), H]_{\beta,\infty}$ with norm at most $B$ for $\beta \in (0,1)$ and $B > 0$, then we find

$$I_2(g, r) \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}}$$

from the approximation result (1.1.1) above.

Based on our approximation result for the regression function $g$ with respect to the RKHS $H$, we define estimators of $g$ which lie in $H$. In order to analyse these estimators, we make further assumptions on $H$. For an RKHS $H$ with kernel $k$, we assume that $H$ is separable and that $k$ is a bounded measurable function on $(S \times S, \mathcal{S} \otimes \mathcal{S})$. The assumptions on $k$ have implications for all functions in $H$. In particular, since $k$ is measurable on $(S \times S, \mathcal{S} \otimes \mathcal{S})$, we find that all functions in $H$ are measurable on $(S, \mathcal{S})$ by Lemma 4.24 of Steinwart and Christmann (2008). We can ensure that $H$ is separable by, for example, assuming that $k$ is continuous and $S$ is a separable topological space. This is shown by Lemma 4.33 of Steinwart and Christmann (2008).

### 1.1.3 Optimal Transport

We now consider the separate topic of optimal transport. The optimal transport problem aims to find the optimal transportation of one probability measure to another with respect to a given cost function. This is done by finding a transport map between the two probability measures which minimises the transport cost.

Let $(X, d_X)$ and $(Y, d_Y)$ be complete separable metric spaces. Furthermore, let $\mathcal{B}(X)$, $\mathcal{B}(Y)$ and $\mathcal{B}(X \times Y)$ be the set of Borel sets on $X$, $Y$ and $X \times Y$, and let $\mathcal{P}(X)$, $\mathcal{P}(Y)$ and $\mathcal{P}(X \times Y)$ be the set of Borel probability measures on $X$, $Y$ and $X \times Y$. We consider the problem of optimally transporting a probability measure $P \in \mathcal{P}(X)$ to $Q \in \mathcal{P}(Y)$ with respect to some Borel cost function $c : X \times Y \to [0, \infty)$. We must specify how the transportation from $P$ to $Q$ can occur.

We define the marginals of $\gamma \in \mathcal{P}(X \times Y)$. Let $\pi_1 : \mathcal{P}(X \times Y) \to \mathcal{P}(X)$ by $(\pi_1\gamma)(A) = \gamma(A \times Y)$ for all $A \in \mathcal{B}(X)$ and let $\pi_2 : \mathcal{P}(X \times Y) \to \mathcal{P}(Y)$ by $(\pi_2\gamma)(B) = \gamma(X \times B)$ for all $B \in \mathcal{B}(Y)$. The marginals of $\gamma \in \mathcal{P}(X \times Y)$ are $\pi_1\gamma \in \mathcal{P}(X)$ and $\pi_2\gamma \in \mathcal{P}(Y)$. We now define

$$\Pi(P, Q) = \{\gamma \in \mathcal{P}(X \times Y) : \pi_1\gamma = P \text{ and } \pi_2\gamma = Q\}.$$

The set $\Pi(P, Q)$ is precisely the set of transportations from $P$ to $Q$. This is because $\gamma \in \Pi(P, Q)$ determines how much probability should be transported from $A \in \mathcal{B}(X)$ to $B \in \mathcal{B}(Y)$ by $\gamma(A \times B)$. We refer to $\gamma$ as a transport plan.

Now that we have defined the set of transport plans, we can define the optimal transport problem itself. We seek

$$\inf_{\gamma \in \Pi(P,Q)} \int c \, d\gamma.$$

The integral determines the transport cost of the transport plan $\gamma \in \Pi(P,Q)$. We are interested in making this as small as possible. If the infimum is attained by some $\gamma \in \Pi(P,Q)$, then $\gamma$ is referred to as an optimal transport plan. By Theorem 4.1 of Villani (2009), an optimal transport plan exists if we assume that $c$ is lower semicontinuous.

We can use the optimal transport problem to measure the difference between $P \in \mathcal{P}(X)$ and $Q \in \mathcal{P}(Y)$. For each $P$ and $Q$, the problem has a minimum transport cost which we refer to as the Wasserstein distance

$$W_c(P,Q) = \inf \left\{ \int c \, d\gamma : \gamma \in \Pi(P,Q) \right\}.$$

This infimum is attained if we assume that $c$ is lower semicontinuous. Note that our definition differs slightly from, for example, Definition 6.1 of Villani (2009). We can use $W_c$ to define balls in $\mathcal{P}(Y)$. The closed Wasserstein ball

$$B_c[P,r] = \{Q \in \mathcal{P}(Y) : W_c(P,Q) \leq r\}$$

for $P \in \mathcal{P}(X)$ and $r \geq 0$. It is straightforward to verify that $B[P,r]$ is convex.

Some transport plans transport probability measures by mapping each point $x \in X$ to a point $y \in Y$. A transport map $T : X \to Y$ is a Borel function such that $P(T^{-1}(B)) = Q(B)$ for all $B \in \mathcal{B}(Y)$. There is a unique transport plan induced by the transport map $T$. This transport plan is $\gamma \in \Pi(P,Q)$ with $\gamma(C) = P(\{x \in X : (x,T(x)) \in C\})$ for $C \in \mathcal{B}(X \times Y)$. Note that $\{x \in X : (x,T(x)) \in C\} \in \mathcal{B}(X)$ because the function $f : X \to X \times Y$ by $f(x) = (x,T(x))$ is Borel.

From the definition, there are some useful results about a transport plan $\gamma$ induced by a transport map $T$. Firstly, $\gamma(A \times B) = P(A \cap T^{-1}(B))$ for $A \in \mathcal{B}(X)$ and $B \in \mathcal{B}(Y)$.

Furthermore, the graph $G = \{(x, y) \in X \times Y : T(x) = y\}$ of $T$ has $\gamma(G) = 1$. Note that $G \in \mathcal{B}(X \times Y)$ because $G = \{(x, y) \in X \times Y : d_Y(T(x), y) = 0\}$ and $f : X \times Y \to [0, \infty)$ by $f(x, y) = d(T(x), y)$ is Borel. In particular, if $f : X \times Y \to \mathbb{R}$ is Borel and either $\gamma$-integrable or non-negative, then

$$\int f \, d\gamma = \int f(x, T(x)) \, dP(x).$$

We consider a final important property of the optimal transport problem, which is the dual problem. In this problem, we seek

$$\sup_{\psi \in L^1(P), \phi \in L^1(Q)} \left\{ \int \phi \, dQ - \int \psi \, dP : \phi(y) - \psi(x) \leq c(x, y) \text{ for all } (x, y) \in X \times Y \right\}.$$

For $\psi \in L^1(P)$ and $\phi \in L^1(Q)$ such that $\phi(y) - \psi(x) \leq c(x, y)$ for all $(x, y) \in X \times Y$, we have

$$\begin{aligned} \int \phi \, dQ - \int \psi \, dP &= \int (\phi(y) - \psi(x)) \, d\gamma(x, y) \\ &\leq \int c(x, y) \, d\gamma(x, y) \end{aligned}$$

for all $\gamma \in \Pi(P, Q)$. By taking an infimum over $\gamma \in \Pi(P, Q)$, we find that

$$\int \phi \, dQ - \int \psi \, dP \leq W_c(P, Q).$$

Hence, the maximum value of the dual problem is always at most the minimum transport cost. We refer to $\psi$ and $\phi$ as dual functions. If we assume that $c$ is lower semicontinuous, then the two problems have the same optimum values by Theorem 5.10 of Villani (2009). If we also assume $c(x, y) \leq c_X(x) + c_Y(y)$ for all $(x, y) \in X \times Y$ and some $c_X \in L^1(P)$ and $c_Y \in L^1(Q)$, then the supremum in the dual problem is attained by some dual functions $\psi$ and $\phi$, again by Theorem 5.10 of Villani (2009).

Such $\psi$ and $\phi$ are referred to as optimal dual functions.

## 1.2 Overview

We now give an overview of the main content of the thesis. After the literature review in Chapter 2, we start by considering our regression problem in Chapter 3. We study least-squares estimators in an RKHS under a norm constraint. This form of regularisation is known as Ivanov regularisation, and it provides better control of the norm of the estimator than the well-established Tikhonov regularisation. Tikhonov regularisation in this context is regularised least-squares estimation in the RKHS, which is used to define SVMs, for example. We assume only that the RKHS is separable with a bounded and measurable kernel.

We provide rates of convergence for the expected squared $L^2(P)$ error of our estimator under the weak assumption that the variance of the response variables is bounded and the unknown regression function lies in an interpolation space between $L^2(P)$ and the RKHS. We then obtain faster rates of convergence when the regression function is bounded by clipping the estimator. Clipping the estimator restricts the values that the estimator can take so that they are not less than or greater than the possible values of the regression function. In this setting, we attain the optimal rate of convergence. Furthermore, we provide a high-probability bound under the stronger assumption that the response variables have subgaussian errors and that the regression function lies in an interpolation space between $L^\infty$ and the RKHS.

We then derive adaptive results for the settings in which the regression function is bounded. We do this by splitting the data into a training set and a validation set. We use the training set to produce our estimators for a range of sizes of norm constraint before using the validation set to select a final estimator. We obtain the same rates of convergence as when we use all of the data to produce the estimator with the best possible size of norm constraint. This training and validation procedure is adaptive because the size of the norm constraint is determined by the data while the best rates of convergence are still attained. The estimators produced from the training set are non-adaptive as they have fixed sizes of norm constraint which do not depend on the data.

In Chapter 4, we study a different adaptive estimation procedure for our clipped Ivanov-regularised least-squares estimators called the Goldenshluger–Lepski method. In contrast to procedures such as training and validation, the Goldenshluger–Lepski method uses all of the data to produce non-adaptive estimators for a range of sizes of norm constraint. We then select an adaptive estimator by performing pairwise comparisons between these estimators. Applying the Goldenshluger–Lepski method is non-trivial as it requires a simultaneous high-probability bound on all of the pairwise comparisons. This bound is known as the majorant.

For our regression problem, use of the Goldenshluger–Lepski method is made more complicated by the fact that we cannot use the $L^2(P)$ norm to perform the pairwise comparisons. This is because the covariate distribution $P$, and hence the $L^2(P)$ norm, are unknown. For this method, the $L^2(P)$ norm would normally be used for making the comparisons as it is the norm in which we seek guarantees on our estimator. However, we are able to adapt the method so that we can perform the comparisons using the $L^2(P_n)$ norm instead, while still obtaining guarantees on our estimator in the $L^2(P)$ norm. Here, $P_n$ is the empirical distribution of the covariates.

We use the Goldenshluger–Lepski method to create two estimation procedures. In the first procedure, the RKHS is fixed and we adapt over a range of sizes of norm constraint. This is similar to the training and validation procedure discussed above, as we adapt over the same parameter. In the second procedure, we adapt over both a collection of RKHSs with Gaussian kernels and a range of sizes of norm constraint in the RKHSs. In this case, we must first produce the non-adaptive estimators for not only the range of sizes of norm constraint but also for each RKHS in the collection.

In Chapter 5, we move away from regression. We study the extreme points of Wasserstein balls of probability measures. We first show that the only extreme points which are not on the surface of the ball are the Dirac measures. By the surface of the ball, we mean the points in the ball whose distance from the centre of the ball is equal to the radius. We then consider points which are on the surface of the ball. We show that if the Wasserstein distance is uniquely attained by a transport plan induced by a transport map, then we have an extreme point. Conversely, under conditions on the centre of the ball and the cost function, we show that if the Wasserstein distance is attained by two distinct transport plans induced by continuous transport maps, then we do not have an extreme point. We then consider the special case in which the probability measures have finite support.

We return to our regression problem in Chapter 6. We seek to control the worst-case squared $L^2(Q)$ error of an estimator over all $Q$ in a Wasserstein ball of probability measures centred at the covariate distribution $P$. We first analyse the worst-case squared $L^2(Q)$ error of our Ivanov-regularised least-squares estimators. We produce expectation bounds both when the regression function is unbounded and when the regression is bounded. We clip our estimators when the regression function is bounded. Furthermore, we produce a high-probability bound when the regression function is bounded and the errors of the response variables are subgaussian.

We then consider alternative estimators defined using an empirical version of the worst-case squared $L^2(Q)$ error. In the empirical version, we centre the Wasserstein ball at $P_n$, the empirical distribution of the covariates. We discuss the problems with the analysis and computation of such estimators. We show that, under suitable conditions, the empirical version of the worst-case squared $L^2(Q)$ error is attained at some $Q$ which is an extreme point of the Wasserstein ball centred at $P_n$. We then briefly consider the approximation properties of the regression function with respect to the supremum of the $L^2(Q)$ norms. Under similar conditions, this supremum is attained at some $Q$ which is an extreme point of the Wasserstein ball centred at $P$. These results are the motivation for the study of the extreme points of Wasserstein balls in Chapter 5. We conclude in Chapter 7 by reviewing the main content of the thesis and discussing some directions for further research.

# Chapter 2

# Literature Review

We now review the literature of the areas most closely related to this thesis. These are reproducing-kernel Hilbert space (RKHS) regression, the Goldenshluger–Lepski method, optimal transport and covariate shift.

## 2.1 RKHS Regression

Estimators in RKHS regression are usually analysed using the spectral decomposition of the kernel operator $T : L^2(P) \to L^2(P)$ by

$$(Tf)(x_1) = \int k(x_1, x_2) f(x_2) \, dP(x_2).$$

If

$$\int k(x, x) dP(x) < \infty,$$

then

$$Tf = \sum_{i=1}^{\infty} \lambda_i \langle f, e_i \rangle_{L^2(P)} e_i$$

for $f \in L^2(P)$, where the $e_i$ for $i \geq 1$ are orthonormal eigenfunctions of $T$ and the $\lambda_i$ are the corresponding eigenvalues (Lemma 2.3 of Steinwart and Scovel, 2012). The $\lambda_i$ are strictly positive and may be selected so that they are non-increasing. Furthermore, $\lambda \in \ell_1$. We can define $T^\alpha : L^2(P) \to L^2(P)$ by

$$T^\alpha f = \sum_{i=1}^{\infty} \lambda_i^\alpha \langle f, e_i \rangle_{L^2(P)} e_i$$

for $\alpha \geq 0$.

### 2.1.1 Early Research

Early research on RKHS regression does not make assumptions about the decay of the $\lambda_i$ for $i \geq 1$. Smale and Zhou (2007) estimate the regression function $g$ using support vector machines (SVMs). These are defined by

$$\hat{f}_\lambda = \operatorname*{arg\,min}_{f \in H} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 + \lambda \|f\|_H^2 \right\}$$

for $\lambda > 0$. It is assumed that the response variables are bounded, so that $|Y_i| \leq M$ for $M > 0$.

The first bound presented by Smale and Zhou (2007) is on the squared RKHS error of an SVM when the regression function is at least as smooth as a general element of $H$. Assume that $g \in T^{\beta/2}(L^2(P))$ for $\beta \in (1, 2]$. Furthermore, let $t > 0$ and

$$\lambda = (3\|k\|_\infty M)^{2/(1+\beta)} \|T^{-\beta/2} g\|_{L^2(P)}^{-2/(1+\beta)} n^{-1/(1+\beta)}.$$

Theorem 2 of Smale and Zhou (2007) shows that

$$\|\hat{f}_\lambda - g\|_H^2 \leq 16 \log(2)^2 (3\|k\|_\infty M)^{2(\beta-1)/(1+\beta)} \|T^{-\beta/2} g\|_{L^2(P)}^{4/(1+\beta)} t^2 n^{-(\beta-1)/(1+\beta)}$$

with probability at least $1 - e^{-t}$. The complexity of the regression function $g$ is measured by $\|T^{-\beta/2}g\|_{L^2(P)}$. The authors also provide a bound on the squared $L^2(P)$ error of an SVM in the same setting. Let

$$\lambda = \log(4)(12\|k\|_\infty M)^{2/(1+\beta)}\|T^{-\beta/2}g\|_{L^2(P)}^{-2/(1+\beta)}tn^{-1/(1+\beta)}.$$

Corollary 5 of Smale and Zhou (2007) shows that

$$\|\hat{f}_\lambda - g\|_{L^2(P)}^2 \leq 4\log(4)^2(12\|k\|_\infty M)^{2\beta/(1+\beta)}\|T^{-\beta/2}g\|_{L^2(P)}^{2/(1+\beta)}tn^{-\beta/(1+\beta)}$$

with probability at least $1 - e^{-t}$ for sufficiently large $n$. Note that this bound is of order $n^{-\beta/(1+\beta)}$.

The final bound of Smale and Zhou (2007) is on the squared $L^2(P)$ error of an SVM when the regression function is less smooth than a general element of $H$. Assume that $g \in T^{\beta/2}(L^2(P))$ for $\beta \in (0, 1]$. Let $t > 0$ and

$$\lambda = 8\log(4)\|k\|_\infty^2 tn^{-1/2}.$$

Corollary 5 of Smale and Zhou (2007) also shows that

$$\|\hat{f}_\lambda - g\|_{L^2(P)}^2 \leq \log(4)^2(8M + 8^{\beta/2}\|k\|_\infty^\beta\|T^{-\beta/2}g\|_{L^2(P)})^2 n^{-\beta/2}$$

with probability at least $1 - e^{-t}$ for $n \geq 1$. This bound is only of order $n^{-\beta/2}$, which is larger than order $n^{-\beta/(1+\beta)}$.

### 2.1.2 Eigenvalue Decay and Smooth Regression Functions

Initial research on RKHS regression which makes use of the decay of the $\lambda_i$ for $i \geq 1$ only applies when the regression function is at least as smooth as a general element of $H$. This is done by Caponnetto and de Vito (2007). We do not consider response variables which are not real-valued, RKHSs which are finite-dimensional or squared $L^2(P)$ errors of estimators with respect to functions other than the regression function, all of which are also covered in the paper.

The estimators considered by Caponnetto and de Vito (2007) are again SVMs. The authors assume that for $M, \sigma > 0$ we have

$$\mathbb{E}\left(\exp\left(\frac{|Y_i - g(X_i)|}{M}\right) - 1 - \frac{|Y_i - g(X_i)|}{M} \,\middle|\, X_i\right) \leq \frac{\sigma^2}{2M^2}$$

for $1 \leq i \leq n$. This condition ensures that the response variables do not vary too much around the regression function evaluated at the covariates. It is often referred to as subexponentiality. The decay of the $\lambda_i$ for $i \geq 1$ is captured by assuming that $\lambda_i \in [ui^{-1/p}, vi^{-1/p}]$ for $v \geq u > 0$ and $p \in (0, 1)$.

We first consider the case in which $g \in T^{\beta/2}(L^2(P))$ for $\beta \in (1, 2]$, so that the regression function is strictly smoother than a general element of $H$. Let $\lambda = n^{-1/(\beta+p)}$. Theorem 1 of Caponnetto and de Vito (2007) shows that any fixed quantile of the squared $L^2(P)$ error of $\hat{f}_\lambda$ is of order at most $n^{-\beta/(\beta+p)}$. Note that this is always smaller than the order $n^{-\beta/(1+\beta)}$ of Smale and Zhou (2007). Furthermore, Theorem 2 of Caponnetto and de Vito (2007) shows that no estimator can attain a smaller order than $n^{-\beta/(\beta+p)}$, so the bound for $\hat{f}_\lambda$ is optimal.

We now consider the case in which $g \in T^{1/2}(L^2(P))$, so that the regression function is as smooth as a general element of $H$. Let $\lambda = \log(n)^{1/(1+p)} n^{-1/(1+p)}$. Theorem 1 of Caponnetto and de Vito (2007) also shows that any fixed quantile of the squared $L^2(P)$ error of $\hat{f}_\lambda$ is of order at most $\log(n)^{1/(1+p)}n^{-1/(1+p)}$. This is always smaller than the order $n^{-1/2}$ of Smale and Zhou (2007). Theorem 2 of Caponnetto and de Vito (2007) shows that no estimator can attain a smaller order than $n^{-1/(1+p)}$, so the bound for $\hat{f}_\lambda$ is close to optimal.

### 2.1.3 Eigenvalue Decay and Non-Smooth Regression Functions

Later research focuses on the case in which the regression function is at most as smooth as a general element of $H$. Mendelson and Neeman (2010) assume that the response variables are bounded, so that $|Y_i| \leq M$ for $1 \leq i \leq n$ and some $M > 0$. The authors also assume that $\lambda_i \leq vi^{-1/p}$ for $v > 0$ and $p \in (0, 1)$. Various Tikhonov-regularised estimators are considered in the paper. The authors assume that $\|k\|_\infty \leq 1$, although this is only to simplify the notation.

The first bound of Mendelson and Neeman (2010) is given for an estimator of the form

$$\hat{g}_\lambda = \operatorname*{arg\,min}_{f \in H} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 + C_1 a(\|f\|_H + 1, \lambda) \right\}$$

for some constant $C_1 > 0$ and

$$a(r, \lambda) = b(2r, \lambda + \log(\pi^2/6) + 2\log(1 + C_2 n + \log r))$$

for $r \geq 1$ and $\lambda > 0$ for some constant $C_2 > 0$. Here,

$$b(r, \lambda) = C_3 r^2 v^{1/(1+p)} n^{-1/(1+p)} + C_4(1 + r^2)\lambda n^{-1}$$

for some constants $C_3, C_4 > 0$. The regularisation function is still of order $\|f\|_H^2$ up to logarithmic terms, which is the same as for SVMs. The authors first consider $g \in T^{1/2}(L^2(P))$, so that the regression function is as smooth as a general element of $H$. The discussion after Theorem 3.8 of Mendelson and Neeman (2010) shows that a fixed quantile depending on $\lambda$ of the squared $L^2(P)$ error of $\hat{g}_\lambda$ is of order at most $n^{-1/(1+p)}$ for all $\lambda > 0$. Note that this is smaller than the order $\log(n)^{1/(1+p)} n^{-1/(1+p)}$ of Caponnetto and de Vito (2007), who show that the rate $n^{-1/(1+p)}$ is in fact optimal.

The authors then consider $g \in T^{\beta/2}(L^2(P))$ for $\beta \in (0, 1)$, so that the regression function is less smooth than a general element of $H$. The discussion after Theorem 3.8 of Mendelson and Neeman (2010) also shows that a fixed quantile depending on $\lambda$ of the squared $L^2(P)$ error of $\hat{g}_\lambda$ is of order at most $n^{-\beta/(1+p)}$ for all $\lambda > 0$. This is always smaller than the order $n^{-\beta/2}$ of Smale and Zhou (2007).

In order to improve the order of the second bound, Mendelson and Neeman (2010) continue by assuming that the eigenfunctions of $T$ are uniformly bounded, so that $\sup_i \|e_i\|_\infty < \infty$. This is a very strong condition which need not hold even when the kernel of the RKHS is very smooth, as discussed after Theorem A of Mendelson and Neeman (2010). The authors then consider new estimators of the form

$$\hat{g}_\lambda = \operatorname*{arg\,min}_{f \in H} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 + C_1 a(f, \lambda) \right\}$$

for some constant $C_1 > 0$ and

$$a(f, \lambda) = C_2(1 + \lambda + C_3 \log n + \log^2(\|f\|_H + e))(\|f\|_H + 1)^{2p/(1+p)} \log(n)^{2/(1+p)} n^{-p/(1+p)}$$

for $f \in H$ and $\lambda > 0$ and some constants $C_2, C_3 > 0$. Note that the regularisation function is now of order $\|f\|_H^{2p/(1+p)}$ up to logarithmic terms, which is always smaller than $\|f\|_H^2$.

The authors again consider $g \in T^{\beta/2}(L^2(P))$ for $\beta \in (0,1)$. The discussion after Corollary 5.5 of Mendelson and Neeman (2010) shows that a fixed quantile depending on $\lambda$ of the squared $L^2(P)$ error of $\hat{g}_\lambda$ is of order at most $n^{-\beta/(\beta+p)}$ for some interval of $\lambda > 0$. This is always smaller than the authors' earlier order of $n^{-\beta/(1+p)}$.

### 2.1.4 Recent Research

The uniform boundedness condition on the eigenfunctions can be relaxed. Steinwart, Hush, and Scovel (2009) instead assume that

$$\|h\|_\infty \le C_1 \|h\|_H^p \|h\|_{L^2(P)}^{1-p}$$

for all $h \in H$ and some constant $C_1 \ge 1$. Here, $p \in (0,1)$ is such that $\lambda_i \le v i^{-1/p}$ for $v > 0$. This assumption is shown to be weaker than uniform boundedness of the eigenfunctions in Theorem 2 of Steinwart et al. (2009). Again, it is assumed that the response variables are bounded, so that $|Y_i| \le M$ for $1 \le i \le n$ and some $M > 0$, and that $\|k\|_\infty \le 1$. Various Tikhonov-regularised estimators are considered, including SVMs. The authors also assume that the regression function $g \in [L^2(P), H]_{\beta,\infty}$ for $\beta \in (0,1)$. This assumption is shown to be weaker than $g \in T^{\beta/2}(L^2(P))$ by Corollary 4.7 of Steinwart and Scovel (2012).

The estimators considered by Steinwart et al. (2009) are of the form

$$\hat{g}_{q,\lambda} = \operatorname*{arg\,min}_{f \in H} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 + \lambda \|f\|_H^q \right\}$$

for $q \ge 1$ and $\lambda > 0$. The authors investigate regularisation functions of various orders $\|f\|_H^q$. Note that $\hat{g}_{2,\lambda} = \hat{f}_\lambda$ an SVM. Since $|Y_i| \le M$ for $1 \le i \le n$, we have $\|g\|_\infty \le M$. Hence, the estimators $\hat{g}_{q,\lambda}$ can be made closer to the regression function

$g$ by clipping them. The authors obtain $V\hat{g}_{q,\lambda}$, where $V : \mathbb{R} \to [-M, M]$ by

$$V(t) = \begin{cases} -M & \text{if} \quad t < -M \\ t & \text{if} \quad |t| \leq M \\ M & \text{if} \quad t > M \end{cases}$$

for $t \in \mathbb{R}$.

Corollary 6 of Steinwart et al. (2009) shows that there is some constant $C_2 \geq 1$ such that, for

$$\lambda = n^{-\frac{2\beta+q(1-\beta)}{2\beta+2p}}$$

and all $t \geq 1$, we have

$$\|V\hat{g}_{q,\lambda} - g\|^2_{L^2(P)} \leq C_2 t n^{-\beta/(\beta+p)}$$

with probability at least $1 - 3\exp(-tn^{\beta p/(\beta+p)})$. Note that this bound is of order $n^{-\beta/(\beta+p)}$, which matches that of Mendelson and Neeman (2010). However, Steinwart et al. (2009) make weaker assumptions. Furthermore, the bound is attained for any $q \geq 1$. Generally, $q = 2$ is preferred for computational reasons. Steinwart et al. (2009) show in Theorem 9 that if we also assume $\lambda_i \geq ui^{-1/p}$ for $u \in (0, v]$, then the rate $n^{-\beta/(\beta+p)}$ is the optimal power of $n$.

## 2.2 The Goldenshluger–Lepski Method

The Goldenshluger–Lepski method is based on Lepski's method, which can perform adaptation over a single parameter. Lepski's method uses all of the data to produce a collection of non-adaptive estimators. It then selects the smoothest non-adaptive estimator, subject to a bound on a series of pairwise comparisons involving all estimators at most as smooth as the resulting estimator. Lepski's method can only adapt to one parameter because of the need for an ordering of the collection of non-adaptive estimators. We discuss this method first.

### 2.2.1 Lepski's Method

Lepski's method is introduced by Lepski (1991b). The author considers the stochastic process $X_\varepsilon$ on $[0,1]$ defined by

$$dX_\varepsilon(t) = S(t)\,dt + \varepsilon\,dW(t),$$

where $S : [0,1] \to \mathbb{R}$, $W$ is a standard Weiner process on $[0,1]$ and $\varepsilon > 0$ determines the variability of $X_\varepsilon$. It is assumed that $S$ is contained by some set of smooth functions $S \in \Sigma(\beta, L)$ for $\beta > 0$ and $L > 0$. Let $\beta = m + \alpha$ for $m$ a non-negative integer and $\alpha \in (0,1]$. Then $\Sigma(\beta, L)$ is defined to be the set of $S : [0,1] \to \mathbb{R}$ such that $S$ is $m$-times continuously differentiable and $|S^{(m)}(t_1) - S^{(m)}(t_1)| \leq L|t_1 - t_2|^\alpha$ for all $t_1, t_2 \in [0,1]$. The aim is to estimate $S(t_0)$ for some fixed $t_0 \in [0,1]$ under the assumption that $\beta$ is unknown. The author obtains expectation bounds on the $q$th power of the error of an adaptive estimator with respect to the Euclidean norm for $q > 0$.

Lepski (1991b) considers a range of non-adaptive kernel estimators indexed by the closed bounded set $I \subseteq (0,\infty)$. Let $a = \inf I$ and $b = \sup I$. The kernel function is defined as $g : \mathbb{R} \to \mathbb{R}$ with support $[0,1]$ such that

$$\int_0^1 g(t)\,dt = 1 \text{ and } \int_0^1 t^j g(t)\,dt = 0$$

for $1 \leq j \leq \lfloor b \rfloor + 1$. For each $\beta \in I$, the author defines the estimator

$$T_\varepsilon(\beta) = \delta(\beta)^{-1} \int_0^1 g(\delta(\beta)^{-1}(t - t_0))\, dX_\varepsilon(t)$$

for the function $\delta : I \to (0, \infty)$ by

$$\delta(\beta) = (b - \beta)^{1/(2\beta+1)} \log(1/\varepsilon)^{1/(2\beta+1)} \varepsilon^{2/(2\beta+1)}$$

for $\beta \neq b$ and $\delta(b) = \varepsilon^{2/(2b+1)}$. The function $\delta$ determines the order of the bound on the adaptive estimator. A finite collection of these estimators is considered. Let $h_\varepsilon = (\log(1/\varepsilon))^{-1}$ and let $\beta_k = a + kh_\varepsilon$ for $1 \leq k \leq h_\varepsilon^{-1}(b - a)$, where $h_\varepsilon^{-1}(b - a)$ is assumed to be a strictly positive integer. The estimators considered are $T_{\varepsilon,k} = T_\varepsilon(\beta_k)$.

Having defined the non-adaptive estimators, Lepski (1991b) defines the adaptive estimator as follows. Let $v_0 = (2a + 1)^{-1/2}((\lfloor a \rfloor \vee 0)!)^{-1}L$, $\sigma = \|g\|_{L^2(0,1)}$ and $d = 4\sigma(2q + 1)^{1/2} + 2v_0\sigma$. Then the adaptive estimator is $T_{\varepsilon,\hat{k}}$, where

$$\hat{k} = \sup\{1 \leq k \leq h_\varepsilon^{-1}(b - a) : |T_{\varepsilon,k} - T_{\varepsilon,l}| \leq d\delta(\beta_l) \text{ for all } l < k\}.$$

Note that the calculation of $\hat{k}$ requires $d$ to be known. Theorem 3 of Lepski (1991b) shows that

$$\sup_{\beta \in I} \lim_{\varepsilon \to 0} \sup_{S \in \Sigma(\beta,L)} \mathbb{E}(\delta(\beta)^{-q}|T_{\varepsilon,\hat{k}} - S(t_0)|^q) < \infty.$$

Therefore, $T_{\varepsilon,\hat{k}} - S(t_0)$ is of order $\log(1/\varepsilon)^{1/(2\beta+1)}\varepsilon^{2/(2\beta+1)}$ as $\varepsilon \to 0$ for all $S \in \Sigma(\beta, L)$. This holds for all $\beta \in I$, even though the adaptive estimator $T_{\varepsilon,\hat{k}}$ does not depend on $\beta$.

### 2.2.2 Lepski's Method for RKHS Regression

Lepski's method has been applied to RKHS regression under the name of the balancing principle. However, as far as we are aware, Lepski's method has not been used to target the true regression function. Instead, it has only been used to target an RKHS element which approximates the regression function.

De Vito, Pereverzyev, and Rosasco (2010) assume that there is some collection of estimators $f^\lambda$ such that

$$\|f^\lambda - f_H\|_{L^2(P)} \leq \alpha(\eta)\lambda^{1/2}(\omega(\lambda)^{-1}n^{-1/2} + A(\lambda))$$

and

$$\|f^\lambda - f_H\|_H \leq \alpha(\eta)(\omega(\lambda)^{-1}n^{-1/2} + A(\lambda))$$

for some $f_H \in H$ simultaneously for all $\lambda \in [n^{-1/2}, 1]$ with probability at least $1 - \eta$ for $\eta \in (0, 1]$. The functions $A : [0, 1] \to [0, \infty)$ and $\omega : (0, 1] \to (0, \infty)$ are continuous, $A(0) = 0$, $\alpha(\eta) \geq \log(2/\eta)^{1/4} \vee 1$ and $\omega(\lambda)A(\lambda) \leq C_1\lambda$ for some constant $C_1 > 0$. Furthermore, $\lambda^{1/2}A(\lambda)$ and $\lambda^{1/2}\omega(\lambda)$ are increasing in $\lambda$. This assumption is very strong. Assumptions of this form are discussed in Section 3.1 of De Vito et al. (2010).

We briefly discuss the implications of the above assumption. The function $f_H$ is the function to be targetted in place of the regression function. The term $\omega(\lambda)^{-1}n^{-1/2}$ corresponds to the sample error of $f^\lambda$ with respect to $f_H$, while $A(\lambda)$ corresponds to the approximation error. A good choice of $\lambda$ balances the sample error and the approximation error. Let $\lambda^* > 0$ satisfy

$$\omega(\lambda^*)^{-1}n^{-1/2} = A(\lambda^*).$$

It is assumed $\lambda^* \leq 1$. Then

$$\|f^{\lambda^*} - f_H\|_{L^2(P)} \leq 2\alpha(\eta)(\lambda^*)^{1/2}A(\lambda^*)$$

and

$$\|f^{\lambda^*} - f_H\|_H \leq 2\alpha(\eta)A(\lambda^*).$$

There is a difficulty in using Lespki's method to control the squared $L^2(P)$ norm of $f^{\hat{\lambda}} - f_H$ for some estimator $\hat{\lambda}$ of $\lambda^*$. Lepski's method requires the norm we are interested in controlling to be known in order to perform the pairwise comparisons. However, the covariate distribution $P$ is unknown in this situation.

De Vito et al. (2010) continue by performing Lepski's method for two different norms and combining the results. Let $\lambda_i \in [n^{-1/2}, 1]$ for $0 \leq i \leq I$ such that $\lambda_{i-1} < \lambda_i$ for $1 \leq i \leq I$. Here, $I$ is some some strictly positive integer. Let $P_n$ be the empirical distribution of the covariates. The authors define

$$\hat{\lambda}_1 = \max\{\lambda_i : \|f^{\lambda_i} - f^{\lambda_j}\|_{L^2(P_n)} \leq 4C_2\alpha(\eta)\lambda_j^{1/2}\omega(\lambda_j)^{-1}n^{-1/2} \text{ for all } 0 \leq j \leq i-1\}$$

for some constant $C_2 > 0$ and

$$\hat{\lambda}_2 = \max\{\lambda_i : \|f^{\lambda_i} - f^{\lambda_j}\|_H \leq 4\alpha(\eta)\omega(\lambda_j)^{-1}n^{-1/2} \text{ for all } 0 \leq j \leq i-1\}.$$

These estimators of $\lambda^*$ are combined to form $\hat{\lambda} = \hat{\lambda}_1 \wedge \hat{\lambda}_2$. Assume that $\lambda_0 \leq C_1^{-1}n^{-1/2}$ and $\omega(\lambda_i) \leq q\omega(\lambda_{i-1})$ for $1 \leq i \leq I$ and some $q > 1$. Theorem 3 of De Vito et al. (2010) shows that

$$\|f^{\hat{\lambda}} - f_H\|_{L^2(P)} \leq C_3 q\alpha(\eta)\lambda^* A(\lambda^*)$$

for some constant $C_3 > 0$. Note that this bound for $f^{\hat{\lambda}}$ is of order $(\lambda^*)^{1/2}$ bigger than the bound for $f^{\lambda^*}$.

### 2.2.3 Multiple Parameters

We now discuss the Goldenshluger–Lepski method itself. The method is introduced by Goldenshluger and Lepski (2008). Let $D$ be an open interval of $\mathbb{R}^d$ containing $D_0 = [-1/2, 1/2]^d$. The authors consider the stochastic process $Y$ on $D$ defined by

$$dY(t) = F(t)\,dt + \varepsilon\,dW(t),$$

where $F : D \to \mathbb{R}$, $W$ is a standard Weiner process on $\mathbb{R}^d$ and $\varepsilon \in (0, 1)$ determines the variability of $Y$. It is assumed that $F$ is continuous and bounded. The aim is to estimate $F(x_0)$ for some fixed $x_0 \in D_0$. The authors obtain expectation bounds on the $r$th power of the error of an adaptive estimator with respect to the Euclidean norm for $r > 0$.

Goldenshluger and Lepski (2008) consider a range of non-adaptive kernel estimators indexed by the compact set $\Theta \subseteq \mathbb{R}^m$ equipped with the Euclidean norm $|\cdot|_2$. Let $\mathcal{K}_\Theta$ be the set of kernels $K_\mu : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for $\mu \in \Theta$. The authors assume that there is some open interval $D_1$ of $\mathbb{R}^d$ with $D_0 \subseteq D_1 \subseteq D$ such that, for all $\mu \in \Theta$, $\operatorname{supp}(K_\mu(\cdot, y)) \subseteq D_1$ for all $y \in D_0$ and

$$\int_D K_\mu(t, y)\,dt = 1$$

for all $y \in D_1$. This ensures that $K_\mu$ satisfies the usual definition of a kernel, along with some regularity conditions.

The authors also assume some boundedness properties of the collection of kernels $\mathcal{K}_\Theta$. In order to express these properties, we define the norms $\|\cdot\|_p$ as the $L^p(D_0)$ norm for

$p \in [1, \infty]$ and

$$\|f\|_{p,\infty} = \sup_{x \in D_0} \left( \int_{\mathbb{R}^d} |f(t,x)|^p \, dt \right)^{1/p} \quad \text{and} \quad \|f\|_{\infty,\infty} = \sup_{x \in D_0} \sup_{t \in \mathbb{R}^d} |f(t,x)|$$

for $f : \mathbb{R}^d \times D_0 \to \mathbb{R}$ and $p \in [1, \infty)$. The assumption is that $\sup_{\mu \in \Theta} \|K_\mu\|_{1,\infty} < \infty$ and $\sup_{\mu \in \Theta} \|K_\mu\|_{2,\infty} < \infty$. Continuity of the kernels is also required. It is assumed that there exist $\gamma \in (0, 1]$ and $L > 0$ such that

$$\sup_{\mu,\mu' \in \Theta} \frac{\|\tilde{K}_\mu - \tilde{K}_{\mu'}\|_{2,\infty}}{|\mu - \mu|_2^\gamma} \leq L \quad \text{and} \quad \sup_{\mu,\mu' \in \Theta} \sup_{x \in \mathbb{R}^d} \frac{|1 - \|K_\mu(\cdot, x)\|_2 / \|K_{\mu'}(\cdot, x)\|_2|}{|\mu - \mu|_2^\gamma} \leq L,$$

where $\tilde{K}_\mu(\cdot, x) = K_\mu(\cdot, x)/\|K_\mu(\cdot, x)\|_2$ for $\mu \in \Theta$.

Goldenshluger and Lepski (2008) define the set $\mathcal{K}_{\Theta \times \Theta}$ of auxiliary kernels $K_{\mu,\nu}$ : $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by

$$K_{\mu,\nu}(t,x) = \int_{D_1} K_\mu(t,y) K_\nu(y,x) \, dy$$

for $t, x \in \mathbb{R}^d$. The kernel $K_{\mu,\nu}$ is in some sense smoother than both $K_\mu$ and $K_\nu$. The authors also demand that $K_{\mu,\nu}(t,x) = K_{\nu,\mu}(t,x)$ for all $t, x \in \mathbb{R}^d$ and all $\mu, \nu \in \Theta$. This occurs if, for example, $K_\mu(t,x) = K_\mu(t - x, 0)$ for all $t, x \in \mathbb{R}^d$ and all $\mu \in \Theta$.

Having defined all of the kernels, the authors define the non-adaptive estimators. Let

$$\hat{F}_\mu(x) = \int_D K_\mu(t,x) \, dY(t) \quad \text{and} \quad \hat{F}_{\mu,\nu}(x) = \int_D K_{\mu,\nu}(t,x) \, dY(t)$$

for $x \in D_0$ and $\mu, \nu \in \Theta$. We first consider the bias of these estimators. Let

$$B_\mu(x) = \int_D K_\mu(t,x) F(t) \, dt - F(x) \quad \text{and} \quad B_{\mu,\nu}(x) = \int_D K_{\mu,\nu}(t,x) F(t) \, dt - F(x)$$

for $x \in D_0$ and $\mu, \nu \in \Theta$. We also require

$$\tilde{B}_\mu(x) = \left( \sup_{\nu \in \Theta} |B_{\mu,\nu}(x) - B_\nu(x)| \right) \vee |B_\mu(x)|$$

for $x \in D_0$ and $\mu \in \Theta$. In order to define the adaptive estimator, the variability of the non-adaptive estimators must be taken into account. Let

$$\xi_\mu(x) = \int_D K_\mu(t, x)\, dW(t) \;\text{ and }\; \xi_{\mu,\nu}(x) = \int_D K_{\mu,\nu}(t, x)\, dW(t)$$

for $x \in D_0$ and $\mu, \nu \in \Theta$, and let $\sigma_\mu(x) = \|K_\mu(\cdot, x)\|_2$ and $\sigma_{\mu,\nu}(x) = \|K_{\mu,\nu}(\cdot, x) - K_\nu(\cdot, x)\|_2$. We also require

$$\tilde{\sigma}_\mu(x) = \left( \sup_{\nu \in \Theta} \int_{\mathbb{R}^d} |K_\nu(y, x)| \sigma_\mu(y)\, dy \right) \vee \sigma_\mu(x)$$

for $x \in D_0$ and $\mu \in \Theta$.

Goldenshluger and Lepski (2008) continue by defining the majorant, which is necessary for defining the adaptive estimator. Recall that $x_0 \in D$ is the point at which we are interested in estimating $F$. Let $\Sigma_\Theta = \{\tilde{\sigma}_\mu(x_0) : \mu \in \Theta\}$ and $\sigma_{\min} = \inf \Sigma_\Theta$. In order to define the majorant, the authors must control $g : \Sigma_\Theta \to [0, \infty)$ by

$$g(\sigma) = \sup_{\mu \in \Theta} \mathbb{E} \left( \sup_{\nu \in \Theta : \tilde{\sigma}_\nu(x_0) \leq \sigma} |\xi_{\mu,\nu} - \xi_\nu| \right).$$

The function $g$ measures the variability of the pairwise comparisons involved in the definition of the adaptive estimator.

The authors assume that there is a known function $e : \Sigma_\Theta \to [0, \infty)$, which is continuous and non-decreasing, such that $e(\sigma) \geq g(\sigma)$ for all $\sigma \in \Sigma_\Theta$. It is also assumed that $e(2\sigma)/e(\sigma) \in [c_e, C_e]$ for all $\sigma \in \Sigma_\Theta$ and some $C_e \geq c_e > 1$. This is similar to the function $e$ being slowly varying. In general, finding $e$ is a very difficult problem. The majorant can then be defined as $Q : \Sigma_\Theta \to [0, \infty)$ by

$$Q(\sigma) = \varkappa_0 e(\sigma) + \sigma(1 + \varkappa_1 \log(\sigma/\sigma_{\min}))^{1/2},$$

where $\varkappa_0 = 2C_e$ and $\varkappa_1 = 128r(1 \vee (\log(C_e)/\log(2)))$. This is an inflated version of $e(\sigma)$, with the increase based on the variability $\sigma \in \Sigma_\Theta$.

Goldenshluger and Lepski (2008) then define the adaptive estimator. Let

$$\hat{R}_\mu = \sup_{\nu \in \Theta : \tilde{\sigma}_\nu(x_0) \geq \tilde{\sigma}_\mu(x_0)} \left( |\hat{F}_{\mu,\nu} - \hat{F}_\nu| - \varepsilon Q(\tilde{\sigma}_\nu(x_0))/2 \right).$$

For $\delta = \varepsilon Q(\sigma_{\min})/4$, let $\hat{\mu}$ be a random element of $\Theta$ such that

$$\hat{R}_{\hat{\mu}} + \varepsilon Q(\tilde{\sigma}_{\hat{\mu}}(x_0)) \leq \inf_{\mu \in \Theta} \left( \hat{R}_\mu + \varepsilon Q(\tilde{\sigma}_\mu(x_0)) \right) + \delta.$$

The adaptive estimator is given by $\hat{F}_{\hat{\mu}}$. The authors provide a bound on the error of $\hat{F}_{\hat{\mu}}(x_0)$ under a final assumption.

Let $\Theta_F$ be the set of $\mu \in \Theta$ such that for all $\sigma \in \Sigma_\Theta$ for which $\sigma \geq \tilde{\sigma}_\mu(x_0)$, there exists $\theta \in \Theta$ such that $\tilde{\sigma}_\theta(x_0) = \sigma$ and $\tilde{B}_\theta(x_0) \leq \varepsilon Q(\tilde{\sigma}_\theta(x_0))/4$. It is assumed that $\Theta_F \neq \varnothing$, which gives a condition on $F$ with respect to the kernels. Define $\mu^* = \arg\min_{\mu \in \Theta_F} \tilde{\sigma}_\mu(x_0)$. Theorem 1 of Goldenshluger and Lepski (2008) shows that

$$(\mathbb{E}(|\hat{F}_{\hat{\mu}}(x_0) - F(x_0)|^r))^{1/r} \leq C\varepsilon Q(\tilde{\sigma}_{\mu^*}(x_0))$$

for $\varepsilon \in (0, 1)$ sufficiently small and some $C > 0$. Therefore, the error of $\hat{F}_{\hat{\mu}}(x_0)$ is bounded by an inflated version of the variability of the pairwise comparisons with respect to $\mu^*$, where $\mu^*$ in some sense produces an estimator with small variability.

## 2.3 Optimal Transport

We now discuss a sample of the relevant literature from optimal transport. We start with the first modern treatment.

### 2.3.1 Early Research

The optimal transport problem in its modern form is introduced by Kantorovitch (1958). The author allows transport between any finite measures with the same total mass. However, here we assume that the measures are probability measures. The author demands that $X = Y$ is compact and the cost function $c$ is continuous. The quantity

$$W_c(P, Q) = \inf_{\gamma \in \Pi(P,Q)} \int c \, d\gamma$$

for $P, Q \in \mathcal{P}(X)$ is defined, and considered as a distance on $\mathcal{P}(X)$. Kantorovitch (1958) states that $W_c(P, Q)$ is attained by some $\gamma \in \Pi(P, Q)$ because $\Pi(P, Q)$ is compact.

The author then considers an early form of the dual problem. Define $U : X \to \mathbb{R}$ to be a potential for $\gamma \in \Pi(P, Q)$ if for all $x, y \in X$ we have $|U(y) - U(x)| \leq c(x, y)$, and furthermore $U(y) - U(x) = c(x, y)$ if $\gamma(A \times B) > 0$ for all open sets $A, B$ such that $x \in A$ and $y \in B$. The theorem of Kantorovitch (1958) shows that $\gamma \in \Pi(P, Q)$ attains $W_c(P, Q)$ if and only if it has a potential.

### 2.3.2 Quadratic Cost Function

Study of the dual problem can lead to the discovery of properties of the solutions to the optimal transport problem itself. Such results often depend on ideas from convex analysis. Rüschendorf and Rachev (1990) consider the case in which $X = Y = \mathbb{R}^k$ equipped with the Euclidean norm $|\cdot|$, and the cost function $c(x, y) = |x - y|^2$ for $x, y \in \mathbb{R}^k$. It is assumed that $P, Q \in \mathcal{P}(\mathbb{R}^k)$ with

$$\int |x|^2 \, dP < \infty \quad \text{and} \quad \int |y|^2 \, dQ < \infty,$$

which ensures that $W_c(P, Q) < \infty$.

Let $f : \mathbb{R}^k \to [-\infty, \infty]$ be a convex function. The subdifferential of a convex function $f : \mathbb{R}^k \to [-\infty, \infty]$ at $x \in \mathbb{R}^k$ is

$$\partial f(x) = \{y \in \mathbb{R}^k : f(z) \geq f(x) + \langle y, z - x \rangle \text{ for all } z \in \mathbb{R}^k\}.$$

This set consists of the gradients of all possible tangents of $f$ at $x$. Theorem 1 of Rüschendorf and Rachev (1990) shows that there exists $\gamma \in \Pi(P, Q)$ which attains $W_c(P, Q)$. Furthermore, the theorem shows that $\gamma \in \Pi(P, Q)$ attains $W_c(P, Q)$ if and only if $\gamma(\{(x, y) : y \in \partial f(x)\}) = 1$ for some lower semicontinuous convex function $f$.

### 2.3.3 General Cost Functions

The previous result can be generalised to other cost functions $c : X \times Y \to [0, \infty)$. Rüschendorf (1995) provides such a result. We call $f : X \to [-\infty, \infty]$ a $c$-convex function if there exists a function $\zeta : Y \to [-\infty, \infty]$ such that

$$f(x) = \sup_{y \in Y} (\zeta(y) - c(x, y)).$$

The $c$-subdifferential of $f$ at $x \in X$ is

$$\partial_c f(x) = \{y \in Y : f(z) \geq f(x) + c(x, y) - c(z, y) \text{ for all } z \in X\}.$$

For $y \in \partial_c f(x)$, the function $f(x) + c(x, y) - c(z, y)$ of $z \in X$ is the equivalent of a tangent of $f$ at $x$. These are slight generalisations of the definitions of Rüschendorf (1995). Furthermore, we replace $c$ with $-c$ in the author's definitions as we are interested in minimising the transport cost as opposed to maximising it.

The author assumes that $X = Y = \mathbb{R}^k$ and $P, Q \in \mathcal{P}(\mathbb{R}^k)$. Furthermore, it is assumed that $c(x, y) \leq c_X(x) + c_Y(y)$ for all $x, y \in \mathbb{R}^k$ and some $c_X \in L^1(P)$ and $c_Y \in L^1(Q)$. Theorem 2 of Rüschendorf (1995) shows that $\gamma \in \Pi(P, Q)$ attains $W_c(P, Q)$ if and only if $\gamma(\{(x, y) : y \in \partial_c f(x)\}) = 1$ for some $c$-convex function $f$. Furthermore, if $c$ is lower semicontinuous, then there exists $\gamma \in \Pi(P, Q)$ which attains $W_c(P, Q)$.

### 2.3.4 Recent Research

A recent book on the subject of optimal transport has been written by Villani (2009). The book is expansive, so we only discuss continuations of the above literature. We allow general complete metric spaces $X$ and $Y$, but restrict the cost function $c : X \times Y \to [0, \infty)$ to be lower semicontinuous. Let $P \in \mathcal{P}(X)$ and $Q \in \mathcal{P}(Y)$. Section 4 of Villani (2009) covers some basic properties of the optimal transport problem. In particular, Theorem 4.1 shows that there exists $\gamma \in \Pi(P, Q)$ which attains $W_c(P, Q)$. This extends the second part of Theorem 2 of Rüschendorf (1995) to more general $X$ and $Y$.

Duality is covered in Section 5 of Villani (2009). Theorem 5.10 is a detailed version of the duality theorem. In particular, it shows that $W_c(P, Q)$ is equal to

$$\sup_{\psi \in L^1(P), \phi \in L^1(Q)} \left\{ \int \phi \, dQ - \int \psi \, dP : \phi(y) - \psi(x) \leq c(x, y) \text{ for all } (x, y) \in X \times Y \right\}.$$

Furthermore, we may restrict $\psi$ to be $c$-convex. The theorem also shows that the supremum is attained if $c(x,y) \leq c_X(x) + c_Y(y)$ for all $(x,y) \in X \times Y$ and some $c_X \in L^1(P)$ and $c_Y \in L^1(Q)$.

Theorem 5.30 of Villani (2009) gives conditions under which the optimal transport problem is solved by a unique transport plan which is induced by a transport map. Suppose that $W_c(P,Q) < \infty$ and, for all $c$-convex functions $f : X \to [-\infty, \infty]$, the set of $x \in X$ such that $\partial_c f(x)$ contains more than one element is $P$-null. Then the theorem shows that $W_c(P,Q)$ is attained by a unique $\gamma \in \Pi(P,Q)$ induced by a transport map $T : X \to Y$. Recall that this means $\gamma(C) = P(\{x \in X : (x, T(x)) \in C\})$ for $C \in \mathcal{B}(X \times Y)$ and that $T$ is Borel. Furthermore, we can select $T$ so that there exists a $c$-convex function $\psi$ such that $T(x) \in \partial_c \psi(x)$ for all $x \in X$.

Consider the above condition that for all $c$-convex functions $f : X \to [-\infty, \infty]$, the set of $x \in X$ such that $\partial_c f(x)$ contains more than one element is $P$-null. It is essentially this condition and an extension of Theorem 2 of Rüschendorf (1995) to more general $X$ and $Y$ which prove Theorem 5.30 of Villani (2009). However, there are only very special circumstances in which the condition is satisfied.

One circumstance in which the condition on $c$-convex functions is satisfied is as follows. Let $X = Y = \mathbb{R}^k$ equipped with the Euclidean norm $|\cdot|$, and let the cost function $c(x,y) = |x-y|^2$ for $x, y \in \mathbb{R}^k$. Furthermore, let $P, Q \in \mathcal{P}(\mathbb{R}^k)$ with

$$\int |x|^2 \, dP < \infty \quad \text{and} \quad \int |y|^2 \, dQ < \infty.$$

Suppose that $P(A) = 0$ for any $A \in \mathcal{B}(\mathbb{R}^k)$ with dimension at most $k-1$. In this case, Theorem 9.4 of Villani (2009) shows that the condition is satisfied and that the result of Theorem 5.30 of Villani (2009) applies. For this cost function, $\psi$ is simply a lower semicontinuous convex function and $T(x) \in \partial_c \psi(x) = \partial \psi(x)$ for all $x \in \mathbb{R}^k$.

## 2.4 Covariate Shift

We present a brief overview of the literature on covariate shift in regression. We start in the parametric setting.

### 2.4.1 Parametric Regression

Covariate shift is first considered for parametric problems. Shimodaira (2000) assumes that the covariates and response variables $(x_t, y_t)$ for $1 \leq t \leq n$ are i.i.d. with density $q(y|x)q_0(x)$ with respect to the Lebesgue measure. However, after the data has been collected, the covariate distribution shifts to $q_1(x)$. The response variables $y_t$ are not required to be one-dimensional.

The author aims to estimate $q(y|x)$ by using a density from the collection $p(y|x, \theta)$ for $\theta \in \Theta \subseteq \mathbb{R}^m$. This restricts the problem so that only some $\theta \in \Theta$ needs to be estimated. Due to the covariate shift, the author considers the loss function

$$\text{loss}_1(\theta) = -\int q_1(x) \int q(y|x) \log p(y|x, \theta) \, dy \, dx.$$

This is the Kullback–Leibler divergence between $q(y|x)q_1(x)$ and $p(y|x, \theta)q_1(x)$, up to an additive constant.

In order to estimate the $\theta \in \Theta$ of interest, Shimodaira (2000) considers a maximum weighted log-likelihood estimation procedure. Let $w$ be some non-negative weight function on the covariate set and define $l_w(x, y|\theta) = -w(x) \log p(y|x, \theta)$ for $\theta \in \Theta$. The weighted log-likelihood function is then

$$L_w(\theta) = -\sum_{t=1}^{n} l_w(x_t, y_t|\theta).$$

The maximum weighted log-likelihood estimator $\hat{\theta}_w = \arg\max_{\theta\in\Theta} L_w(\theta)$. However, only certain weight functions are allowed by the author.

The proper weight functions $w$ considered by Shimodaira (2000) must satisfy the following properties. Let $E_0$ denote integration with respect to $q(y|x)q_0(x)$, the density generating the data. It is required that $E_0(l_w(x,y|\theta))$ exists for all $\theta \in \Theta$. Furthermore, $E_0(l_w(x,y|\theta))$ must have a unique minimiser $\theta_w^*$ which lies in $\Theta^\circ$ the interior of $\Theta$. Finally, $E_0(l_w(x,y|\theta))$ must have a non-singular Hessian at $\theta_w^*$.

Shimodaira (2000) uses these definitions to describe the $\theta \in \Theta$ that we are interested in estimating. Note that if $w = q_1/q_0$, then we have $E_0(l_w(x,y|\theta)) = \text{loss}_1(\theta)$. Hence, in this case, $\theta_w^* = \arg\min_{\theta\in\Theta} \text{loss}_1(\theta)$. This $\theta_w^*$ is referred to as $\theta_1^*$. It is this value of $\theta \in \Theta$ that we are interested in estimating. Furthermore, in this case, the estimator $\hat{\theta}_w$ is referred to as $\hat{\theta}_1$.

Lemma 1 of Shimodaira (2000) tells us how well $\hat{\theta}_w$ estimates $\theta_w^*$ for any proper weight function $w$. Suppose that the model is sufficiently smooth and that $p(y|x,\theta)$ has the same support as $q(y|x)$ for all $\theta \in \Theta$. Furthermore, assume that the $m \times m$ matrices $G_w$ and $H_w$ defined by

$$G_{w,i,j} = E_0\left(\frac{\partial l_w(x,y|\theta_w^*)}{\partial\theta_i}\frac{\partial l_w(x,y|\theta_w^*)}{\partial\theta_j}\right) \quad \text{and} \quad H_{w,i,j} = E_0\left(\frac{\partial^2 l_w(x,y|\theta_w^*)}{\partial\theta_i\,\partial\theta_j}\right)$$

are nonsingular. Then $n^{1/2}(\hat{\theta}_w - \theta_w^*)$ converges in distribution to $\text{N}(0, H_w^{-1}G_wH_w^{-1})$. In general, a weight function $w$ not proportional to $q_1/q_0$ has $\theta_w^* \neq \theta_1^*$ and $\text{loss}_1(\theta_w^*) > \text{loss}_1(\theta_1^*)$. In this case, Lemma 1 of Shimodaira (2000) shows that $\theta_1^* \in \Theta$ should be estimated by $\hat{\theta}_1$. Note that this requires both covariate distributions $q_0$ and $q_1$ to be known.

### 2.4.2 Nonparametric Regression

Sugiyama, Suzuki, Nakajima, Kashima, von Bünau, and Kawanabe (2008) consider the problem of estimating the weight function above when the two covariate distributions are unknown. The authors allow more general nonparametric models. Let $Q$ be the distribution generating the original covariates and $P$ be the distribution of the covariates after a covariate shift, both defined on $D \subseteq \mathbb{R}^d$. It is assumed that $P$ and $Q$ are equivalent. The aim is to estimate $g_0 = dP/dQ$. The authors assume that $\inf_{x \in D} g_0(x) > 0$ and $\sup_{x \in D} g_0(x) < \infty$.

In order to estimate the weight function $g_0$, the authors assume that we have i.i.d. samples from $P$ and $Q$. It is assumed for convenience that there are the same number $n$ of samples from both distributions. The empirical distributions of these samples are referred to as $P_n$ and $Q_n$. For any measure $\mu$ and any $\mu$-integrable function $f$, the authors use the notation $\mu f$ to refer to the integral of $f$ with respect to $\mu$.

Given the samples above, Sugiyama et al. (2008) estimate $g_0$ using a linear combination of basis functions. Let $F$ be some set of non-negative basis functions on $D$. $F$ may be infinite, however it is assumed that $\inf_{\phi \in F} Q\phi > 0$ and $\sup_{\phi \in F} \|\phi\|_\infty < \infty$. Furthermore, the authors demand that the subset of basis functions $F_n \subseteq F$ considered when estimating $g_0$ from the pair of $n$ samples is finite. However, $F_n$ is allowed to depend on the samples and therefore be random. For an example of this scenario, consider a kernel $k : D \times D \to \mathbb{R}$. We can let $F = \{k(x, \cdot) : x \in D\}$ and $F_n$ consist of the $k(x, \cdot)$ such that $x$ is a sample generated by $P$.

Having defined the basis functions, the authors then define their linear combinations. Let

$$G = \left\{ \sum_{l=1}^{L} \alpha_l \phi_l : \alpha_l \geq 0 \text{ and } \phi_l \in F \text{ for } 1 \leq l \leq L \text{ and all } L \geq 1 \right\}$$

be all finite positive linear combinations of elements of $F$ and let $G^M = \{g \in G : \|g\|_\infty \leq M\}$ for all $M \geq 0$. Furthermore, let

$$G_n = \left\{ \sum_{l=1}^{|F_n|} \alpha_l \phi_l : \alpha_l \geq 0 \text{ and } \phi_l \in F \text{ for } 1 \leq l \leq |F_n| \right\}$$

be all finite positive linear combinations of elements of $F_n$. The authors then define their estimator

$$\hat{g}_n = \operatorname*{arg\,max}_{g \in G_n} \{P_n \log(g) : Q_n g = 1\}.$$

Here, $P_n \log(g)$ is an empirical version of the negative Kullback–Leibler divergence between $P$ and the measure $\tilde{P}$ such that $d\tilde{P}/dQ = g$, up to an additive constant. We must have $Qg = 1$ for $\tilde{P}$ to be a probability measure. The empirical version of this constraint is $Q_n g = 1$. The authors assume that $\hat{g}_n$ is unique.

In order to present bounds on $\hat{g}_n$, Sugiyama et al. (2008) bound the size of $G^M$ for all $M \geq 0$. Let $N_{[]}(\varepsilon, G^M, L^2(Q))$ be the $\varepsilon > 0$ bracketing number of $G^M$ with respect to the $L^2(Q)$ norm. This number is defined to be the smallest integer $n \geq 1$ such that there exist functions $l_i : D \to \mathbb{R}$ and $u_i : D \to \mathbb{R}$ for $1 \leq i \leq n$ for which $\|l_i - u_i\|_{L^2(Q)} < \varepsilon$ and, for all $f \in G^M$, there exists $i$ such that $l_i \leq f \leq u_i$. The authors assume that there exist $\gamma \in (0, 2)$ and $K \geq 0$ such that $\log N_{[]}(\varepsilon, G^M, L^2(Q)) \leq K(M/\varepsilon)^\gamma$ for all $M \geq 0$. The value of $\gamma$ is larger when the $G^M$ are bigger. The bounds on $\hat{g}_n$ are given with respect to the generalised Hellinger distance $h_Q(g_1, g_2) = \|\sqrt{g_1} - \sqrt{g_2}\|_{L^2(Q)}$ for non-negative functions $g_1$ and $g_2$.

Let $a_0^n = (Q_n g_0)^{-1}$ and $\delta_n = (P_n \log(a_0^n g_0/\hat{g}_n)) \vee 0$. Remark 2 of Sugiyama et al. (2008) shows that $h_Q(\hat{g}_n, g_0)$ is of order $n^{-1/(2+\gamma)} + \sqrt{\delta_n}$ in probability. Note that this result depends on the size of the $G^M$ for $M \geq 0$ through $\gamma \in (0, 2)$. However, the dependence on $\delta_n$ is not desirable. Assume that there exists $g_n^* \in G_n$ such that $Q_n g_n^* = 1$ and $\|g_0/g_n^*\|_\infty < \infty$. Then Theorem 2 of Sugiyama et al. (2008) shows that

$h_Q(\hat{g}_n, g_0)$ is of order $n^{-1/(2+\gamma)} + h_Q(g_n^*, g_0)$ in probability. This replaces $\sqrt{\delta_n}$ with $h_Q(g_n^*, g_0)$, which is easier to interpret.

# Chapter 3

# Ivanov-Regularised Least-Squares Estimators over Large RKHSs and Their Interpolation Spaces

One of the key problems to overcome in nonparametric regression is overfitting, due to estimators coming from large hypothesis classes. To avoid this phenomenon, it is common to ensure that both the empirical risk and some regularisation function are small when defining an estimator. There are three natural ways to achieve this goal. We can minimise the empirical risk subject to a constraint on the regularisation function, minimise the regularisation function subject to a constraint on the empirical risk or minimise a linear combination of the two. These techniques are known as Ivanov regularisation, Morozov regularisation and Tikhonov regularisation respectively (Oneto, Ridella, and Anguita, 2016). Ivanov and Morozov regularisation can be viewed as dual problems, while Tikhonov regularisation can be viewed as the Lagrangian relaxation of either.

Tikhonov regularisation has gained popularity as it provides a closed-form estimator in many situations. In particular, Tikhonov regularisation in which the estimator is selected from a reproducing-kernel Hilbert space (RKHS) has been extensively studied (Smale and Zhou, 2007; Caponnetto and de Vito, 2007; Steinwart and Christmann, 2008; Mendelson and Neeman, 2010; Steinwart et al., 2009). Although Tikhonov regularisation produces an estimator in closed form, it is Ivanov regularisation which provides the greatest control over the hypothesis class, and hence over the estimator it produces. For example, if the regularisation function is the norm of the RKHS, then the bound on this function forces the estimator to lie in a ball of predefined radius inside the RKHS. An RKHS norm measures the smoothness of a function, so the norm constraint bounds the smoothness of the estimator. By contrast, Tikhonov regularisation provides no direct control over the smoothness of the estimator.

The control we have over the Ivanov-regularised estimator is useful in many settings. The most obvious use of Ivanov regularisation is when the regression function lies in a ball of known radius inside the RKHS. In this case, Ivanov regularisation can be used to constrain the estimator to lie in the same ball. Suppose, for example, that we are interested in estimating the trajectory of a particle from noisy observations over time. Assume that the velocity or acceleration of the particle is constrained by certain physical conditions. Constraints of this nature can be imposed by bounding the norm of the trajectory in a Sobolev space. Certain Sobolev spaces are RKHSs, so it is possible to use Ivanov regularisation to enforce physical conditions on an estimator of the trajectory which match those of the trajectory itself. Ivanov regularisation can also be used within larger inference methods. It is compatible with validation, allowing us to control an estimator selected from an uncountable collection. This is because the Ivanov-regularised estimator is continuous in the size of the ball containing it (see Lemma 3.15.2), so the estimators parametrised by an interval of ball sizes can be controlled simultaneously using chaining.

In addition to the other useful properties of the Ivanov-regularised estimator, Ivanov regularisation can be performed almost as quickly as Tikhonov regularisation. The Ivanov-regularised estimator is a support vector machine (SVM) with regularisation parameter selected to match the norm constraint (see Lemma 3.6.1). This parameter can be selected to within a tolerance $\varepsilon$ using interval bisection with order $\log(1/\varepsilon)$ iterations. In general, Ivanov regularisation requires the calculation of order $\log(1/\varepsilon)$ SVMs.

In this chapter, we study the behaviour of the Ivanov-regularised least-squares estimator with regularisation function equal to the norm of the RKHS. We derive a number of novel results concerning the rate of convergence of the estimator in various settings and under various assumptions. Our analysis is performed by controlling empirical processes over balls in the RKHS. By contrast, the analysis of Tikhonov-regularised estimators usually relies on the spectral decomposition of the kernel operator $T$ on $L^2(P)$. Here, $P$ is the covariate distribution.

We first prove an expectation bound on the squared $L^2(P)$ error of our estimator of order $n^{-\beta/2}$, under the weak assumption that the response variables have bounded variance. Here, $n$ is the number of data points, and $\beta$ parametrises the interpolation space between $L^2(P)$ and $H$ containing the regression function. As far as we are aware, the analysis of an estimator in this setting has not previously been considered. The definition of an interpolation space is given in Section 3.1. The expected squared $L^2(P)$ error can be viewed as the expected squared error of our estimator at a new independent covariate, with the same distribution $P$. If we also assume that the regression function is bounded, then it makes sense to clip our estimator so that it takes values in the same interval as the regression function. This further assumption allows us to achieve an expectation bound on the squared $L^2(P)$ error of the clipped estimator of order $n^{-\beta/(1+\beta)}$.

We then move away from the average behaviour of the error towards its behaviour in the worst case. We obtain high-probability bounds of the same order, under the stronger assumption that the response variables have subgaussian errors and the interpolation space is between $L^\infty$ and $H$. The second assumption is quite natural as we already assume that the regression function is bounded, and $H$ can be continuously embedded in $L^\infty$ since it has a bounded kernel $k$. Note that this assumption means that the set of possible regression functions is independent of the covariate distribution.

When the regression function is bounded, we also analyse an adaptive version of our estimator, which does not require us to know which interpolation space contains the regression function. This adaptive estimator obtains bounds of the same order as the non-adaptive one.

Our expectation bound of order $n^{-\beta/(1+\beta)}$, when the regression function is bounded, improves on the high-probability bound of Smale and Zhou (2007) of order $n^{-\beta/2}$. Their bound is attained under the stronger assumption that the regression function lies in the image of a power of the kernel operator, instead of an interpolation space (see Steinwart and Scovel, 2012). The authors also assume that the response variables are bounded. Furthermore, for a fixed $\beta \in (0, 1)$, Steinwart et al. (2009) show that there is an instance of our problem with a bounded regression function such that the following holds. For all estimators $\hat{f}$ of $g$, for some $\varepsilon > 0$, we have

$$\|\hat{f} - g\|_{L^2(P)}^2 \geq C_{\alpha,\varepsilon} n^{-\alpha}$$

with probability at least $\varepsilon$ for all $n \geq 1$, for some constant $C_{\alpha,\varepsilon} > 0$, for all $\alpha > \beta/(1+\beta)$. Hence, for all estimators $\hat{f}$ of $g$, we have

$$\mathbb{E}\left(\|\hat{f} - g\|_{L^2(P)}^2\right) \geq C_{\alpha,\varepsilon} \varepsilon n^{-\alpha}$$

for all $n \geq 1$, for all $\alpha > \beta/(1+\beta)$. In this sense, our expectation bound in this setting is optimal because it attains the order $n^{-\beta/(1+\beta)}$, the smallest possible power of $n$. Our expectation bound on the adaptive version of our estimator is also optimal, because the bound is of the same order as in the easier non-adaptive setting.

The high-probability bound of Steinwart et al. (2009) is optimal in a similar sense, although the authors achieve faster rates by assuming a fixed rate of decay of the eigenvalues of the kernel operator $T$, as discussed in Section 3.2. Since there is an additional parameter for the decay of the eigenvalues, the collection of problem instances for a fixed set of parameters is smaller in their paper. This means that our optimal rates are the slowest of the optimal rates in Steinwart et al. (2009).

## 3.1 RKHSs and Their Interpolation Spaces

A Hilbert space $H$ of real-valued functions on $S$ is an RKHS if the evaluation functional $L_x : H \to \mathbb{R}$, $L_x h = h(x)$, is bounded for all $x \in S$. In this case, $L_x \in H^*$ the dual of $H$ and the Riesz representation theorem tells us that there is some $k_x \in H$ such that $h(x) = \langle h, k_x \rangle_H$ for all $h \in H$. The kernel is then given by $k(x_1, x_2) = \langle k_{x_1}, k_{x_2} \rangle_H$ for $x_1, x_2 \in S$, and is symmetric and positive-definite.

Now suppose that $(S, \mathcal{S})$ is a measurable space on which $P$ is a probability measure. We can define a range of interpolation spaces between $L^2(P)$ and $H$ (Bergh and Löfström, 1976). Let $(Z, \|\cdot\|_Z)$ be a Banach space and $(V, \|\cdot\|_V)$ be a dense subspace of $Z$. The $K$-functional of $(Z, V)$ is

$$K(z, t) = \inf_{v \in V} (\|z - v\|_Z + t\|v\|_V)$$

for $z \in Z$ and $t > 0$. For $\beta \in (0, 1)$ and $1 \leq q < \infty$, we define

$$\|z\|_{\beta,q} = \left( \int_0^\infty (t^{-\beta} K(z,t))^q t^{-1} dt \right)^{1/q} \text{ and } \|z\|_{\beta,\infty} = \sup_{t>0} (t^{-\beta} K(z,t))$$

for $z \in Z$. The interpolation space $[Z, V]_{\beta,q}$ is defined to be the set of $z \in Z$ such that $\|z\|_{\beta,q} < \infty$. Smaller values of $\beta$ give larger spaces. The space $[Z, V]_{\beta,q}$ is not much larger than $V$ when $\beta$ is close to 1, but we obtain spaces which get closer to $Z$ as $\beta$ decreases. The following result is essentially Theorem 3.1 of Smale and Zhou (2003). The authors only consider the case in which $\|v\|_Z \leq \|v\|_V$ for all $v \in V$, however the result holds by the same proof even without this condition.

**Lemma 3.1.1** *Let $(Z, \|\cdot\|_Z)$ be a Banach space, $(V, \|\cdot\|_V)$ be a dense subspace of $Z$ and $z \in [Z, V]_{\beta,\infty}$. We have*

$$\inf\{\|v - z\|_Z : v \in V, \|v\|_V \leq r\} \leq \frac{\|z\|_{\beta,\infty}^{1/(1-\beta)}}{r^{\beta/(1-\beta)}}.$$

When $H$ is dense in $L^2(P)$, we can define the interpolation spaces $[L^2(P), H]_{\beta,q}$, where $L^2(P)$ is the space of measurable functions $f$ on $(S, \mathcal{S})$ such that $f^2$ is integrable with respect to $P$. We work with $q = \infty$, which gives the largest space of functions for a fixed $\beta \in (0, 1)$. We can then use the approximation result in Lemma 3.1.1. When $H$ is dense in $L^\infty$, we also require $[L^\infty, H]_{\beta,q}$, where $L^\infty$ is the space of bounded measurable functions on $(S, \mathcal{S})$.

## 3.2 Literature Review

Early research on RKHS regression does not make assumptions on the rate of decay of the eigenvalues of the kernel operator. For example, Smale and Zhou (2007) assume that the response variables are bounded and the regression function is of the form $g = T^{\beta/2} f$ for $\beta \in (0, 1]$ and $f \in L^2(P)$. Here, $T : L^2(P) \to L^2(P)$ is the kernel operator and $P$ is the covariate distribution. The authors achieve a squared $L^2(P)$ error of order $n^{-\beta/2}$ with high probability by using SVMs.

Initial research which does make assumptions on the rate of decay of the eigenvalues of the kernel operator, such as that of Caponnetto and de Vito (2007), assumes that the regression function is at least as smooth as an element of $H$. However, their paper still allows for regression functions of varying smoothness by letting $g \in T^{(\beta-1)/2}(H)$ for $\beta \in [1, 2]$. By assuming that the $i$th eigenvalue of $T$ is of order $i^{-1/p}$ for $p \in (0, 1]$, the authors achieve a squared $L^2(P)$ error of order $n^{-\beta/(\beta+p)}$ with high probability by using SVMs. This squared $L^2(P)$ error is shown to be of optimal order for $\beta \in (1, 2]$.

Later research focuses on the case in which the regression function is at most as smooth as an element of $H$. Often, this research demands that the response variables are bounded. For example, Mendelson and Neeman (2010) assume that $g \in T^{\beta/2}(L^2(P))$ for $\beta \in (0, 1)$ to obtain a squared $L^2(P)$ error of order $n^{-\beta/(1+p)}$ with high probability by using Tikhonov-regularised least-squares estimators. The authors also show that if the eigenfunctions of the kernel operator $T$ are uniformly bounded in $L^\infty$, then the order can be improved to $n^{-\beta/(\beta+p)}$. Steinwart et al. (2009) relax the condition on the eigenfunctions to the condition

$$\|h\|_\infty \leq C_p \|h\|_H^p \|h\|_{L^2(P)}^{1-p}$$

for all $h \in H$ and some constant $C_p > 0$. The same rate is attained by using clipped Tikhonov-regularised least-squares estimators, including clipped SVMs, and is shown to be optimal. The authors assume that $g$ is in an interpolation space between $L^2(P)$ and $H$, which is slightly more general than the assumption of Mendelson and Neeman (2010). A detailed discussion about the image of $L^2(P)$ under powers of $T$ and interpolation spaces between $L^2(P)$ and $H$ is given by Steinwart and Scovel (2012).

Lately, the assumption that the response variables must be bounded has been relaxed to allow for subexponential errors. However, the assumption that the regression function is bounded has been maintained. For example, Fischer and Steinwart (2017) assume that $g \in T^{\beta/2}(L^2(P))$ for $\beta \in (0, 2]$ and that $g$ is bounded. The authors also assume that $T^{\alpha/2}(L^2(P))$ is continuously embedded in $L^\infty$, with respect to an appropriate norm on $T^{\alpha/2}(L^2(P))$, for some $\alpha < \beta$. This gives the same squared $L^2(P)$ error of order $n^{-\beta/(\beta+p)}$ with high probability by using SVMs.

## 3.3 Contribution

In this chapter, we provide bounds on the squared $L^2(P)$ error of our Ivanov-regularised least-squares estimator when the regression function comes from an interpolation space between $L^2(P)$ and an RKHS $H$, which is separable with a bounded and measurable kernel $k$. We use the norm of the RKHS as our regularisation function. Under the weak assumption that the response variables have bounded variance, we prove a bound on the expected squared $L^2(P)$ error of order $n^{-\beta/2}$ (Theorem 3.7.2 on page 57). As far as we are aware, the analysis of an estimator in this setting has not previously been considered. If we assume that the regression function is bounded, then we can clip the estimator and achieve an expected squared $L^2(P)$ error of order $n^{-\beta/(1+\beta)}$ (Theorem 3.7.4 on page 59).

Under the stronger assumption that the response variables have subgaussian errors and the regression function comes from an interpolation space between $L^\infty$ and $H$, we show that the squared $L^2(P)$ error is of order $n^{-\beta/(1+\beta)}$ with high probability (Theorem 3.8.2 on page 65). For the settings in which the regression function is bounded, we use training and validation on the data in order to select the size of the constraint on the norm of our estimator. This gives us an adaptive estimation result which does not require us to know which interpolation space contains the regression function. We obtain a squared $L^2(P)$ error of order $n^{-\beta/(1+\beta)}$ in expectation and with high probability, depending on the setting (Theorems 3.7.6 and 3.8.4 on pages 62 and 67). In order to perform training and validation, the response variables in the validation set must have subgaussian errors. The expectation results for bounded regression functions are of optimal order in the sense discussed at the end of the introduction. The results not involving validation are summarised in Table 3.1. The columns for which there is an $L^\infty$ bound on the regression function also make the $L^2(P)$ interpolation assumption. Orders of bounds marked with $(*)$ are known to be optimal.

| Regression Function | $L^2(P)$ Interpolation | $L^\infty$ Bound | $L^\infty$ Interpolation |
|---|---|---|---|
| Response Variables | Bounded Variance | Bounded Variance | Subgaussian Errors |
| Bound Type | Expectation | Expectation | High Probability |
| Bound Order | $n^{-\beta/2}$ | $n^{-\beta/(1+\beta)}$ $(*)$ | $n^{-\beta/(1+\beta)}$ |

Table 3.1: Orders of bounds on squared $L^2(P)$ error

The validation results are summarised in Table 3.2. Again, the columns for which there is an $L^\infty$ bound on the regression function also make the $L^2(P)$ interpolation assumption. The assumptions on the response variables relate to those in the validation set, which has $\tilde{n}$ data points. We assume that $\tilde{n}$ is equal to some multiple of $n$. Again, orders of bounds marked with $(*)$ are known to be optimal.

| Regression Function | $L^\infty$ Bound | $L^\infty$ Interpolation |
|---|---|---|
| Response Variables | Subgaussian Errors | Subgaussian Errors |
| Bound Type | Expectation | High Probability |
| Bound Order | $n^{-\beta/(1+\beta)}$ $(*)$ | $n^{-\beta/(1+\beta)}$ |

Table 3.2: Orders of validation bounds on squared $L^2(P)$ error

## 3.4 Problem Definition

We now formally define our regression problem. For a topological space $T$, let $\mathcal{B}(T)$ be the Borel $\sigma$-algebra of $T$. Let $(S, \mathcal{S})$ be a measurable space. Assume that $(X_i, Y_i)$ for $1 \leq i \leq n$ are $(S \times \mathbb{R}, \mathcal{S} \otimes \mathcal{B}(\mathbb{R}))$-valued random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which are i.i.d. with $X_i \sim P$ and $\mathbb{E}(Y_i^2) < \infty$, where $\mathbb{E}$ denotes integration with respect to $\mathbb{P}$. Since any version of $\mathbb{E}(Y_i | X_i)$ is $\sigma(X_i)$-measurable, where $\sigma(X_i)$ is the $\sigma$-algebra generated by $X_i$, we have that $\mathbb{E}(Y_i | X_i) = g(X_i)$ almost surely for some function $g$ which is measurable on $(S, \mathcal{S})$ (Section A3.2 of Williams, 1991). From the definition of conditional expectation and the identical distribution of the $(X_i, Y_i)$, it is clear that we can choose $g$ to be the same for all $1 \leq i \leq n$. The conditional expectation used is that of Kolmogorov, defined using the Radon–Nikodym derivative. Its definition is unique almost surely. Since $\mathbb{E}(Y_i^2) < \infty$, it follows that $g \in L^2(P)$ by Jensen's inequality. To summarise, $\mathbb{E}(Y_i | X_i) = g(X_i)$ almost surely for $1 \leq i \leq n$ with $g \in L^2(P)$. We assume throughout that

$$(Y1) \qquad \operatorname{var}(Y_i | X_i) \leq \sigma^2 \text{ almost surely for } 1 \leq i \leq n.$$

Our results depend on how well $g$ can be approximated by elements of an RKHS $H$ with kernel $k$. We make the following assumptions.

$(H)$ The RKHS $H$ with kernel $k$ has the following properties:

- The RKHS $H$ is separable.
- The kernel $k$ is bounded.
- The kernel $k$ is a measurable function on $(S \times S, \mathcal{S} \otimes \mathcal{S})$.

We define

$$\|k\|_\infty = \sup_{x \in S} k(x,x)^{1/2} < \infty.$$

We can guarantee that $H$ is separable by, for example, assuming that $k$ is continuous and $S$ is a separable topological space (Lemma 4.33 of Steinwart and Christmann, 2008). The fact that $H$ has a kernel $k$ which is measurable on $(S \times S, \mathcal{S} \otimes \mathcal{S})$ guarantees that all functions in $H$ are measurable on $(S, \mathcal{S})$ (Lemma 4.24 of Steinwart and Christmann, 2008).

## 3.5 Ivanov Regularisation

We now consider Ivanov regularisation for least-squares estimators. Let $P_n$ be the empirical distribution of the $X_i$ for $1 \leq i \leq n$. The definition of Ivanov regularisation provides us with the following result.

**Lemma 3.5.1** *Let $A \subseteq L^2(P)$. It may be that $A$ is a function of $\omega \in \Omega$ and does not contain $g$. Let*

$$\hat{f} \in \arg\min_{f \in A} \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2.$$

*Then, for all $f \in A$ and all $\omega \in \Omega$, we have*

$$\|\hat{f} - f\|_{L^2(P_n)}^2 \leq \frac{4}{n} \sum_{i=1}^n (Y_i - g(X_i))(\hat{f}(X_i) - f(X_i)) + 4\|f - g\|_{L^2(P_n)}^2.$$

In general, the first term of the right-hand side of the inequality must be controlled by bounding it with

$$\sup_{f_1, f_2 \in A} \frac{4}{n} \sum_{i=1}^n (Y_i - g(X_i))(f_1(X_i) - f_2(X_i)). \tag{3.5.1}$$

This is usually not measurable. However, if $A$ is a fixed subset of a separable RKHS, then $A$ is separable and the function which evaluates $f \in A$ at $X_i$ is continuous for $1 \leq i \leq n$. This means that the supremum can be replaced with a countable supremum, so the quantity is a random variable on $(\Omega, \mathcal{F})$. Clearly, this term increases as $A$ gets larger. However, if $A$ gets larger, then we may select $f \in A$ closer to $g$. Hence, we can make the second term of the right-hand side of the inequality in Lemma 3.5.1 smaller. This demonstrates the trade-off in selecting the size of $A$ for the Ivanov-regularised least-squares estimator constrained to lie in $A$.

The next step in analysing $\hat{f}$ is to move to a bound on

$$\|\hat{f} - f\|_{L^2(P)}^2 \leq \|\hat{f} - f\|_{L^2(P_n)}^2 + \sup_{f_1, f_2 \in A} \left| \|f_1 - f_2\|_{L^2(P_n)}^2 - \|f_1 - f_2\|_{L^2(P)}^2 \right| . \quad (3.5.2)$$

The second term on the right-hand side of this inequality is measurable when $A$ is a fixed subset of a separable RKHS. It also increases with $A$. Finally, we obtain a bound on

$$\|\hat{f} - g\|_{L^2(P)}^2 \leq 2\|\hat{f} - f\|_{L^2(P)}^2 + 2\|f - g\|_{L^2(P)}^2.$$

This again demonstrates why $f \in A$ should be close to $g$.

## 3.6 Estimator Definition

Let $B_H$ be the closed unit ball of $H$ and $r > 0$. The Ivanov-regularised least-squares estimator constrained to lie in $rB_H$ is

$$\hat{h}_r = \operatorname*{arg\,min}_{f \in rB_H} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2.$$

We also define $\hat{h}_0 = 0$.

**Lemma 3.6.1** *Assume (H). Let $K$ be the $n \times n$ symmetric matrix with $K_{i,j} = k(X_i, X_j)$. Then $K$ is an $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$ and there exist an orthogonal matrix $A$ and a diagonal matrix $D$ which are both $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrices on $(\Omega, \mathcal{F})$ such that $K = ADA^\mathsf{T}$. Furthermore, the diagonal entries of $D$ are non-negative and non-increasing. Let $m = \operatorname{rk} K$, which is a random variable on $(\Omega, \mathcal{F})$. For $r > 0$, if*

$$r^2 < \sum_{i=1}^{m} D_{i,i}^{-1} (A^\mathsf{T} Y)_i^2,$$

*then define $\mu(r) > 0$ by*

$$\sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu(r))^2} (A^\mathsf{T} Y)_i^2 = r^2. \tag{3.6.1}$$

*Otherwise, let $\mu(r) = 0$. We have that $\mu(r)$ is strictly decreasing when $\mu(r) > 0$, and $\mu(r)$ is measurable on $(\Omega \times (0, \infty), \mathcal{F} \otimes \mathcal{B}((0, \infty)))$, where $r$ varies in $(0, \infty)$. Let $a \in \mathbb{R}^n$ be defined by*

$$(A^\mathsf{T} a)_i = (D_{i,i} + n\mu(r))^{-1} (A^\mathsf{T} Y)_i$$

*for $1 \leq i \leq m$ and $(A^\mathsf{T} a)_i = 0$ for $m + 1 \leq i \leq n$, noting that $A^\mathsf{T}$ has the inverse $A$ since it is an orthogonal matrix. For $r \geq 0$, we can uniquely define $\hat{h}_r$ by demanding that $\hat{h}_r \in \operatorname{sp}\{k_{X_i} : 1 \leq i \leq n\}$. This gives*

$$\hat{h}_r = \sum_{i=1}^{n} a_i k_{X_i}$$

*for $r > 0$ and $\hat{h}_0 = 0$. We have that $\hat{h}_r$ is a $(H, \mathcal{B}(H))$-valued measurable function on $(\Omega \times [0, \infty), \mathcal{F} \otimes \mathcal{B}([0, \infty)))$, where $r$ varies in $[0, \infty)$.*

Let $r > 0$. There are multiple methods for calculating $\mu(r)$ to within a given tolerance $\varepsilon > 0$. We call this value $\nu(r)$.

### 3.6.1 Diagonalising $K$

Firstly, $\mu(r) = 0$ if and only if

$$r \geq \left( \sum_{i=1}^{m} D_{i,i}^{-1} (A^\mathsf{T} Y)_i^2 \right)^{1/2},$$

so in this case we set $\nu(r) = 0$. Otherwise, $\mu(r) > 0$ and

$$\begin{aligned} r^2 &= \sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu(r))^2} (A^\mathsf{T} Y)_i^2 \\ &\leq n^{-2} \left( \sum_{i=1}^{m} D_{i,i} (A^\mathsf{T} Y)_i^2 \right) \mu(r)^{-2}. \end{aligned}$$

Hence,

$$\mu(r) \leq n^{-1} \left( \sum_{i=1}^{m} D_{i,i} (A^\mathsf{T} Y)_i^2 \right)^{1/2} r^{-1}. \tag{3.6.2}$$

The function

$$\sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu)^2} (A^\mathsf{T} Y)_i^2$$

of $\mu \geq 0$ is continuous. Hence, we can calculate $\nu(r)$ using interval bisection on the interval with lower end point 0 and upper end point equal to the right-hand side of (3.6.2). We can then approximate $a$ by replacing $\mu(r)$ with $\nu(r)$ in the calculation of $a$ in Lemma 3.6.1.

### 3.6.2 Not Diagonalising $K$

We can calculate an alternative $\nu(r)$ without diagonalising $K$. Note that if $\mu(r) > 0$, then (3.6.1) can be written as

$$Y^\mathsf{T} (K + n\mu(r) I)^{-1} K (K + n\mu(r) I)^{-1} Y = r^2.$$

Since $\mu(r)$ is strictly decreasing for $\mu(r) > 0$, we have

$$r \geq \left(Y^\mathsf{T}(K + n\varepsilon I)^{-1}K(K + n\varepsilon I)^{-1}Y\right)^{1/2}$$

if and only if $\mu(r) \in [0, \varepsilon]$, so in this case we set $\nu(r) = \varepsilon$. Otherwise, $\mu(r) > \varepsilon$ and (3.6.2) can be written as

$$\mu(r) \leq n^{-1}(Y^\mathsf{T}KY)^{1/2}r^{-1}. \tag{3.6.3}$$

The function

$$Y^\mathsf{T}(K + n\mu I)^{-1}K(K + n\mu I)^{-1}Y$$

of $\mu > 0$ is continuous. Hence, we can calculate $\nu(r)$ using interval bisection on the interval with lower end point $\varepsilon$ and upper end point equal to the right-hand side of (3.6.3). When $\mu(r) > 0$ or $K$ is invertible, we can also calculate $a$ in Lemma 3.6.1 using $a = (K + n\mu(r)I)^{-1}Y$. Since $\nu(r) > 0$, we can approximate $a$ by $(K + n\nu(r)I)^{-1}Y$.

If we have that $K$ is invertible, then we can calculate the $\nu(r)$ in Subsection 3.6.1 while still not diagonalising $K$. We have $\mu(r) = 0$ if and only if $r \geq (Y^\mathsf{T}K^{-1}Y)^{1/2}$, so in this case we set $\nu(r) = 0$. Otherwise, $\mu(r) > 0$ and (3.6.2) can be written as

$$\mu(r) \leq n^{-1}(Y^\mathsf{T}KY)^{1/2}r^{-1},$$

so we can again use interval bisection to calculate $\nu(r)$. We can still approximate $a$ by $(K + n\nu(r)I)^{-1}Y$.

### 3.6.3 Approximating $\hat{h}_r$

Having discussed how to approximate $\mu(r)$ by $\nu(r)$ to within a given tolerance $\varepsilon > 0$, we now consider the estimator produced by this approximation. We find that this estimator is equal to $\hat{h}_s$ for some $s > 0$. We only have $\nu(r) = 0$ for the methods considered above when $\mu(r) = 0$, in which case we can let $s = r$ to obtain the approximate estimator $\hat{h}_s = \hat{h}_r$. Otherwise, $\nu(r) > 0$. Let

$$s = \left( \sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\nu(r))^2} (A^{\mathsf{T}} Y)_i^2 \right)^{1/2} .$$

By (3.6.1), we have $\mu(s) = \nu(r)$ and the approximate estimator is equal to $\hat{h}_s$. Assume that $r$ is bounded away from 0 as $n \to \infty$ and let $C > 0$ be some constant not depending on $n$. We can ensure that $s$ is of the same order as $r$ as $n \to \infty$ by demanding that $s$ is within $C$ of $r$. This is enough to ensure that the orders of convergence for $\hat{h}_r$ apply to $\hat{h}_s$. In order to attain this value of $\nu(r)$, interval bisection should terminate at $x \in \mathbb{R}$ such that

$$\left( \sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + nx)^2} (A^{\mathsf{T}} Y)_i^2 \right)^{1/2}$$

is within $C$ of $r$. Note that this guarantees $\|\hat{h}_s - \hat{h}_r\|_H \leq C^{1/2}(r + s)^{1/2}$ by Lemma 3.15.2.

## 3.7 Expectation Bounds

To capture how well $g$ can be approximated by elements of $H$, we define

$$I_2(g, r) = \inf \left\{ \|h_r - g\|_{L^2(P)}^2 : h_r \in rB_H \right\}$$

for $r > 0$. We consider the distance of $g$ from $rB_H$ because we constrain our estimator $\hat{h}_r$ to lie in this set. The supremum in (3.5.1) with $A = rB_H$ can be controlled using the reproducing kernel property and the Cauchy–Schwarz inequality to obtain

$$8r\left(\frac{1}{n^2}\sum_{i,j=1}^{n}(Y_i-g(X_i))(Y_j-g(X_j))k(X_i,X_j)\right)^{1/2}.$$

The expectation of this quantity can be bounded using Jensen's inequality. Something very similar to this argument gives the first term of the bound in Theorem 3.7.1 below. The expectation of the supremum in (3.5.2) with $A = rB_H$ can be controlled using symmetrisation (Lemma 2.3.1 of van der Vaart and Wellner, 1996) to obtain

$$2\,\mathbb{E}\left(\sup_{f\in 2rB_H}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)^2\right|\right),$$

where the $\varepsilon_i$ for $1 \le i \le n$ are i.i.d. Rademacher random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, independent of the $X_i$. Since $\|f\|_\infty \le 2\|k\|_\infty r$ for all $f \in 2rB_H$, we can remove the squares on the $f(X_i)$ by using the contraction principle for Rademacher processes (Theorem 3.2.1 of Giné and Nickl, 2016). This quantity can then be bounded in a similar way to the supremum in (3.5.1), giving the second term of the bound in Theorem 3.7.1 below.

**Theorem 3.7.1** *Assume (Y1) and (H). Let $r > 0$. We have*

$$\mathbb{E}\left(\|\hat{h}_r - g\|_{L^2(P)}^2\right) \le \frac{8\|k\|_\infty \sigma r}{n^{1/2}} + \frac{64\|k\|_\infty^2 r^2}{n^{1/2}} + 10I_2(g,r).$$

We can obtain rates of convergence for our estimator $\hat{h}_r$ if we make an assumption about how well $g$ can be approximated by elements of $H$. Let us assume

$(g1)$ $\quad g \in [L^2(P), H]_{\beta,\infty}$ with norm at most $B$ for $\beta \in (0,1)$ and $B > 0$.

The assumption $(g1)$, together with Lemma 3.1.1, give

$$I_2(g, r) \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} \tag{3.7.1}$$

for $r > 0$. We obtain an expectation bound on the squared $L^2(P)$ error of our estimator $\hat{h}_r$ of order $n^{-\beta/2}$.

**Theorem 3.7.2** *Assume (Y1), (H) and (g1). Let $r > 0$. We have*

$$\mathbb{E}\left(\|\hat{h}_r - g\|_{L^2(P)}^2\right) \leq \frac{8\|k\|_\infty \sigma r}{n^{1/2}} + \frac{64\|k\|_\infty^2 r^2}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}}.$$

*Let $D_1 > 0$. Setting*

$$r = D_1 \|k\|_\infty^{-(1-\beta)} B n^{(1-\beta)/4}$$

*gives*

$$\mathbb{E}\left(\|\hat{h}_r - g\|_{L^2(P)}^2\right) \leq D_2 \|k\|_\infty^{2\beta} B^2 n^{-\beta/2} + D_3 \|k\|_\infty^{\beta} B \sigma n^{-(1+\beta)/4}$$

*for constants $D_2, D_3 > 0$ depending only on $D_1$ and $\beta$.*

Since we must let $r \to \infty$ for the initial bound in Theorem 3.7.2 to tend to 0, the second term of the initial bound is asymptotically larger than the first. If we ignore the first term and minimise the second and third terms over $r > 0$, we get

$$r = \left(\frac{5\beta}{32(1-\beta)}\right)^{(1-\beta)/2} \|k\|_\infty^{-(1-\beta)} B n^{(1-\beta)/4}.$$

In particular, $r$ is of the form in Theorem 3.7.2. This choice of $r$ gives

$$D_2 = 64\left(\frac{5\beta}{32(1-\beta)}\right)^{1-\beta} + 10\left(\frac{32(1-\beta)}{5\beta}\right)^{\beta} \text{ and } D_3 = 8\left(\frac{5\beta}{32(1-\beta)}\right)^{(1-\beta)/2}.$$

The fact that the second term of the initial bound is larger than the first produces some interesting observations. Firstly, the choice of $r$ above does not depend on $\sigma^2$. Secondly, we can decrease the bound if we can find a way to reduce the second term, without having to alter the other terms. The increased size of the second term is due to the fact that the bound on $f \in 2rB_H$ is given by $\|f\|_\infty \leq 2\|k\|_\infty r$ when applying the contraction principle for Rademacher processes. If we can use a bound which does not depend on $r$, then we can reduce the size of the second term.

We now also assume

$$(g2) \qquad \|g\|_\infty \leq C \text{ for } C > 0$$

and clip our estimator. Let $r > 0$. Since $g$ is bounded in $[-C, C]$, we can make $\hat{h}_r$ closer to $g$ by constraining it to lie in the same interval. Similarly to Chapter 7 of Steinwart and Christmann (2008) and Steinwart et al. (2009), we define the projection $V : \mathbb{R} \to [-C, C]$ by

$$V(t) = \begin{cases} -C & \text{if} \quad t < -C \\ t & \text{if} \quad |t| \leq C \\ C & \text{if} \quad t > C \end{cases}$$

for $t \in \mathbb{R}$. We can apply the inequality

$$\|V\hat{h}_r - Vh_r\|^2_{L^2(P_n)} \leq \|\hat{h}_r - h_r\|^2_{L^2(P_n)}$$

for all $h_r \in rB_H$. We continue analysing $V\hat{h}_r$ by bounding

$$\sup_{f_1, f_2 \in rB_H} \left| \|Vf_1 - Vf_2\|^2_{L^2(P_n)} - \|Vf_1 - Vf_2\|^2_{L^2(P)} \right|.$$

The expectation of the supremum can be bounded in the same way as before, with some adjustments. After symmetrisation, we can remove the squares on the $Vf_1(X_i) - Vf_2(X_i)$ for $f_1, f_2 \in rB_H$ and $1 \leq i \leq n$ by using the contraction principle for

Rademacher processes with $\|Vf_1 - Vf_2\|_\infty \leq 2C$. We can then use the triangle inequality to remove $Vf_2(X_i)$, before applying the contraction principle again to remove $V$. The expectation bound on the squared $L^2(P)$ error of our estimator $V\hat{h}_r$ follows in the same way as before.

**Theorem 3.7.3** *Assume (Y1), (H) and (g2). Let $r > 0$. We have*

$$\mathbb{E}\left(\|V\hat{h}_r - g\|_{L^2(P)}^2\right) \leq \frac{8\|k\|_\infty(16C + \sigma)r}{n^{1/2}} + 10I_2(g, r).$$

We can obtain rates of convergence for our estimator $V\hat{h}_r$ by again assuming $(g1)$. We obtain an expectation bound on the squared $L^2(P)$ error of $V\hat{h}_r$ of order $n^{-\beta/(1+\beta)}$.

**Theorem 3.7.4** *Assume (Y1), (H), (g1) and (g2). Let $r > 0$. We have*

$$\mathbb{E}\left(\|V\hat{h}_r - g\|_{L^2(P)}^2\right) \leq \frac{8\|k\|_\infty(16C + \sigma)r}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}}.$$

*Let $D_1 > 0$. Setting*

$$r = D_1\|k\|_\infty^{-(1-\beta)/(1+\beta)}B^{2/(1+\beta)}(16C + \sigma)^{-(1-\beta)/(1+\beta)}n^{(1-\beta)/(2(1+\beta))}$$

*gives*

$$\mathbb{E}\left(\|V\hat{h}_r - g\|_{L^2(P)}^2\right) \leq D_2\|k\|_\infty^{2\beta/(1+\beta)}B^{2/(1+\beta)}(16C + \sigma)^{2\beta/(1+\beta)}n^{-\beta/(1+\beta)}$$

*for a constant $D_2 > 0$ depending only on $D_1$ and $\beta$.*

If we minimise the initial bound in Theorem 3.7.4 over $r > 0$, we get

$$r = \left(\frac{5\beta}{2(1-\beta)}\right)^{(1-\beta)/(1+\beta)}\|k\|_\infty^{-(1-\beta)/(1+\beta)}B^{2/(1+\beta)}(16C + \sigma)^{-(1-\beta)/(1+\beta)}n^{(1-\beta)/(2(1+\beta))}.$$

In particular, $r$ is of the form in Theorem 3.7.4. This choice of $r$ gives

$$D_2 = 2 \cdot 5^{(1-\beta)/(1+\beta)} \cdot 4^{2\beta/(1+\beta)} \left( \left( \frac{2\beta}{1-\beta} \right)^{(1-\beta)/(1+\beta)} + \left( \frac{1-\beta}{2\beta} \right)^{2\beta/(1+\beta)} \right).$$

Although the second bound in Theorem 3.7.4 is of theoretical interest, it is in practice impossible to select $r$ of the correct order in $n$ for the bound to hold without knowing $\beta$. Since assuming that we know $\beta$ is not realistic, we must use some other method for determining a good choice of $r$.

### 3.7.1 Validation

Suppose that we have an independent second data set $(\tilde{X}_i, \tilde{Y}_i)$ for $1 \leq i \leq \tilde{n}$ which are $(S \times \mathbb{R}, \mathcal{S} \otimes \mathcal{B}(\mathbb{R}))$-valued random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let the $(\tilde{X}_i, \tilde{Y}_i)$ be i.i.d. with $\tilde{X}_i \sim P$ and $\mathbb{E}(\tilde{Y}_i | \tilde{X}_i) = g(\tilde{X}_i)$ almost surely. Let $\rho \geq 0$ and $R \subseteq [0, \rho]$ be non-empty and compact. Furthermore, let $F = \{V\hat{h}_r : r \in R\}$. We estimate a value of $r$ which makes the squared $L^2(P)$ error of $V\hat{h}_r$ small by

$$\hat{r} = \operatorname*{arg\,min}_{r \in R} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (V\hat{h}_r(\tilde{X}_i) - \tilde{Y}_i)^2.$$

The minimum is attained because Lemma 3.15.2 shows that it is the minimum of a continuous function over a compact set. In the event of ties, we may take $\hat{r}$ to be the infimum of all points attaining the minimum. Lemma 3.15.3 shows that the estimator $\hat{r}$ is a random variable on $(\Omega, \mathcal{F})$. Hence, by Lemma 3.6.1, $\hat{h}_{\hat{r}}$ is a $(H, \mathcal{B}(H))$-valued random variable on $(\Omega, \mathcal{F})$.

The definition of $\hat{r}$ means that we can analyse $V\hat{h}_{\hat{r}}$ using Lemma 3.5.1. The expectation of the supremum in (3.5.1) with $A = F$ can be bounded using chaining (Theorem 2.3.7 of Giné and Nickl, 2016). The diameter of $(F, \|\cdot\|_\infty)$ is $2C$, which is an important bound for the use of chaining. Hence, this form of analysis can only be performed under the assumption $(g2)$. After symmetrisation, the expectation of the supremum in (3.5.2) with $A = F$ can be bounded in the same way. In order to perform chaining, we need to make an assumption on the behaviour of the errors of the response variables $\tilde{Y}_i$ for $1 \le i \le \tilde{n}$. Let $U$ and $V$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say $U$ is $\sigma^2$-subgaussian if

$$\mathbb{E}(\exp(tU)) \le \exp(\sigma^2 t^2/2)$$

for all $t \in \mathbb{R}$. We say $U$ is $\sigma^2$-subgaussian given $V$ if

$$\mathbb{E}(\exp(tU)|V) \le \exp(\sigma^2 t^2/2)$$

almost surely for all $t \in \mathbb{R}$. We assume

$(\tilde{Y})$ $\qquad$ $\tilde{Y}_i - g(\tilde{X}_i)$ is $\tilde{\sigma}^2$-subgaussian given $\tilde{X}_i$ for $1 \le i \le \tilde{n}$.

This is stronger than the equivalent of the assumption $(Y1)$, that $\operatorname{var}(\tilde{Y}_i|\tilde{X}_i) \le \tilde{\sigma}^2$ almost surely.

**Theorem 3.7.5** *Assume (H) and $(\tilde{Y})$. Let $r_0 \in R$. We have*

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2\right)$$

*is at most*

$$\frac{32C(4C + \tilde{\sigma})}{\tilde{n}^{1/2}}\left(\left(2\log\left(2 + \frac{\|k\|_\infty^2 \rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right) + 10\,\mathbb{E}\left(\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2\right).$$

In order for us to apply the validation result in Theorem 3.7.5 to the initial bound in Theorem 3.7.4, we need to make an assumption on $R$. We assume either

$$(R1) \qquad R = [0, \rho] \text{ for } \rho = an^{1/2} \text{ and } a > 0$$

or

$$(R2) \quad R = \{bi : 0 \leq i \leq I - 1\} \cup \{an^{1/2}\} \text{ and } \rho = an^{1/2} \text{ for } a, b > 0 \text{ and } I = \lceil an^{1/2}/b \rceil.$$

The assumption $(R1)$ is mainly of theoretical interest and would make it difficult to calculate $\hat{r}$ in practice. The estimator $\hat{r}$ can be computed under the assumption $(R2)$, since in this case $R$ is finite. We obtain an expectation bound on the squared $L^2(P)$ error of $V\hat{h}_{\hat{r}}$ of order $n^{-\beta/(1+\beta)}$. This is the same order in $n$ as the second bound in Theorem 3.7.4.

**Theorem 3.7.6** *Assume (Y1), (H), (g1), (g2) and ($\tilde{Y}$). Also assume (R1) or (R2) and that $\tilde{n}$ increases at least linearly in $n$. We have*

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2\right) \leq D_1 n^{-\beta/(1+\beta)}$$

*for a constant $D_1 > 0$ not depending on $n$ or $\tilde{n}$.*

## 3.8 High-Probability Bounds

In this section, we look at how to extend our expectation bounds on our estimators to high-probability bounds. In order to do this, we must control the second term of the bound in Lemma 3.5.1 with $A = rB_H$ for $r > 0$, which is

$$\|h_r - g\|_{L^2(P_n)}^2 \tag{3.8.1}$$

for $h_r \in rB_H$. There is no way to bound (3.8.1) in high-probability without strict assumptions on $g$. In fact, the most natural assumption is ($g2$) that $\|g\|_\infty \leq C$ for $C > 0$, which we assume throughout this section. Bounding (3.8.1) also requires us to introduce a new measure of how well $g$ can be approximated by elements of $H$. We define

$$I_\infty(g, r) = \inf\left\{\|h_r - g\|_\infty^2 : h_r \in rB_H\right\}$$

for $r > 0$. Note that $I_\infty(g, r) \geq I_2(g, r)$. Using $I_\infty(g, r)$ instead of $I_2(g, r)$ means that we do not have to control (3.8.1) by relying on $\|h_r - g\|_\infty \leq \|k\|_\infty r + C$. Using Hoeffding's inequality, this would add a term of order $r^2 t^{1/2}/n^{1/2}$ for $t \geq 1$ to the bound in Theorem 3.8.1 below, which holds with probability $1 - 3e^{-t}$, substantially increasing its size.

It may be possible to avoid this problem by instead considering the Ivanov-regularised least-squares estimator

$$\hat{f}_r = \operatorname*{arg\,min}_{f \in V(rB_H)} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

for $r > 0$, where $V(rB_H) = \{Vh_r : h_r \in rB_H\}$. The second term of the bound in Lemma 3.5.1 with $A = V(rB_H)$ is

$$\|Vh_r - g\|_{L^2(P_n)}^2 \tag{3.8.2}$$

for $h_r \in rB_H$. Since $\|Vh_r - g\|_\infty \leq 2C$, using Hoeffding's inequality to bound (3.8.2) would only add a term of order $C^2 t^{1/2}/n^{1/2}$ to the bound in Theorem 3.8.1 below, which would not alter its size. However, the calculation and analysis of the estimator $\hat{f}_r$ is outside the scope of this chapter. This is because the calculation of $\hat{f}_r$ involves minimising a quadratic form subject to a series of linear constraints, and its analysis requires a bound on the supremum in (3.5.1) with $A = V(rB_H)$.

The rest of the analysis of $V\hat{h}_r$ is similar to that of the expectation bound. The supremum in (3.5.1) with $A = rB_H$ can again be bounded by

$$8r\left(\frac{1}{n^2}\sum_{i,j=1}^{n}(Y_i - g(X_i))(Y_j - g(X_j))k(X_i, X_j)\right)^{1/2}.$$

The quadratic form can be bounded using Lemma 3.16.2, under an assumption on the behaviour of the errors of the response variables $Y_i$ for $1 \leq i \leq n$. The proof of Theorem 3.8.1 below uses a very similar argument to this one. The supremum in (3.5.2) with $A = rB_H$ can be bounded using Talagrand's inequality (Theorem A.9.1 of Steinwart and Christmann, 2008). In order to use Lemma 3.16.2, we must assume

$(Y2)$ $\qquad Y_i - g(X_i)$ is $\sigma^2$-subgaussian given $X_i$ for $1 \leq i \leq n$.

This assumption is stronger than $(Y1)$. In particular, Theorem 3.7.3 still holds under the assumptions $(Y2)$, $(H)$ and $(g2)$.

**Theorem 3.8.1** *Assume (Y2), (H) and (g2). Let $r > 0$ and $t \geq 1$. With probability at least $1 - 3e^{-t}$, we have*

$$\|V\hat{h}_r - g\|_{L^2(P)}^2$$

*is at most*

$$\frac{8\left(2C^2 + 8\|k\|_\infty^{1/2}C^{3/2}r^{1/2} + \|k\|_\infty(16C + 5\sigma)r\right)t^{1/2}}{n^{1/2}} + \frac{16C^2t}{3n} + 10I_\infty(g, r).$$

We can obtain rates of convergence for our estimator $V\hat{h}_r$, but we must make a new assumption about how well $g$ can be approximated by elements of $H$, instead of $(g1)$. We now assume

$(g3)$ $\qquad g \in [L^\infty, H]_{\beta,\infty}$ with norm at most $B$ for $\beta \in (0, 1)$ and $B > 0$, instead of $g \in [L^2(P), H]_{\beta,\infty}$ with norm at most $B$. This assumption is stronger than $(g1)$, as it implies that the norm of $g \in [L^2(P), H]_{\beta,\infty}$ is

$$\sup_{t>0}(t^{-\beta}\inf_{h\in H}(\|g-h\|_{L^2(P)} + t\|h\|_H)) \le \sup_{t>0}(t^{-\beta}\inf_{h\in H}(\|g-h\|_{L^\infty} + t\|h\|_H)) \le B.$$

In particular, Theorem 3.7.4 still holds under the assumptions $(Y1)$, $(H)$, $(g2)$ and $(g3)$ or $(Y2)$, $(H)$, $(g2)$ and $(g3)$. The assumption $(g3)$, together with Lemma 3.1.1, give

$$I_\infty(g,r) \le \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}}. \tag{3.8.3}$$

We obtain a high-probability bound on the squared $L^2(P)$ error of $V\hat{h}_r$ of order $t^{\beta/(1+\beta)}n^{-\beta/(1+\beta)}$ with probability at least $1-e^{-t}$.

**Theorem 3.8.2** *Assume (Y2), (H), (g2) and (g3). Let $r > 0$ and $t \ge 1$. With probability at least $1-3e^{-t}$, we have*

$$\|V\hat{h}_r - g\|^2_{L^2(P)}$$

*is at most*

$$\frac{8\left(2C^2 + 8\|k\|_\infty^{1/2}C^{3/2}r^{1/2} + \|k\|_\infty(16C+5\sigma)r\right)t^{1/2}}{n^{1/2}} + \frac{16C^2t}{3n} + \frac{10B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}}.$$

*Let $D_1 > 0$. Setting*

$$r = D_1\|k\|_\infty^{-(1-\beta)/(1+\beta)}B^{2/(1+\beta)}(16C+5\sigma)^{-(1-\beta)/(1+\beta)}t^{-(1-\beta)/(2(1+\beta))}n^{(1-\beta)/(2(1+\beta))}$$

*gives*

$$\|V\hat{h}_r - g\|^2_{L^2(P)}$$

*is at most*

$$\begin{aligned}&D_2\|k\|_\infty^{2\beta/(1+\beta)}B^{2/(1+\beta)}(16C+5\sigma)^{2\beta/(1+\beta)}t^{\beta/(1+\beta)}n^{-\beta/(1+\beta)}\\&+D_3\|k\|_\infty^{\beta/(1+\beta)}B^{1/(1+\beta)}C^{3/2}(16C+5\sigma)^{-(1-\beta)/(2(1+\beta))}t^{(1+3\beta)/(4(1+\beta))}n^{-(1+3\beta)/(4(1+\beta))}\\&+D_4C^2t^{1/2}n^{-1/2}+D_5C^2tn^{-1}\end{aligned}$$

*for constants $D_2, D_3, D_4, D_5 > 0$ depending only on $D_1$ and $\beta$.*

Since we must let $r \to \infty$ for the initial bound in Theorem 3.8.2 to tend to 0, the asymptotically largest terms in the bound are

$$\frac{8\|k\|_\infty(16C+5\sigma)rt^{1/2}}{n^{1/2}}+\frac{10B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}}.$$

If we minimise this over $r > 0$, we get $r$ of the form in Theorem 3.8.2 with

$$D_1=\left(\frac{5\beta}{2(1-\beta)}\right)^{(1-\beta)/(1+\beta)}.$$

### 3.8.1 Validation

We now extend our expectation bound on $V\hat{h}_{\hat{r}}$ to a high-probability bound. The supremum in (3.5.1) with $A = F$ can be bounded using chaining (Exercise 1 of Section 2.3 of Giné and Nickl, 2016), while the supremum in (3.5.2) with $A = F$ can be bounded using Talagrand's inequality.

**Theorem 3.8.3** *Assume (H) and ($\tilde{Y}$). Let $r_0 \in R$ and $t \geq 1$. With probability at least $1 - 3e^{-t}$, we have*

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2$$

*is at most*

$$\frac{20C(C+\tilde{\sigma})t^{1/2}}{\tilde{n}^{1/2}}\left(1+32\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)\right)$$
$$+\frac{48C^2t^{1/2}}{\tilde{n}^{1/2}}+\frac{16C^2t}{3\tilde{n}}+10\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2.$$

We can apply the validation result in Theorem 3.8.3 to the initial bound in Theorem 3.8.2 by assuming either $(R1)$ or $(R2)$. We obtain a high-probability bound on the squared $L^2(P)$ error of $V\hat{h}_{\hat{r}}$ of order $t^{1/2}n^{-\beta/(1+\beta)}$ with probability at least $1-e^{-t}$. This is the same order in $n$ as the second bound in Theorem 3.8.2.

**Theorem 3.8.4** *Assume (Y2), (H), (g2), (g3) and ($\tilde{Y}$). Let $t \geq 1$. Also assume (R1) or (R2) and that $\tilde{n}$ increases at least linearly in $n$. With probability at least $1-6e^{-t}$, we have*

$$\|V\hat{h}_{\hat{r}}-g\|_{L^2(P)}^2 \leq D_1 t^{1/2} n^{-\beta/(1+\beta)} + D_2 t n^{-1}$$

*for constants $D_1, D_2 > 0$ not depending on $n$, $\tilde{n}$ or $t$.*

## 3.9 Discussion

In this chapter, we show how Ivanov regularisation can be used to produce smooth estimators which have a small squared $L^2(P)$ error. We first consider the case in which the regression function lies in an interpolation space between $L^2(P)$ and an RKHS $H$. We achieve bounds on the squared $L^2(P)$ under the assumption that $H$ is separable, with a bounded and measurable kernel. Under the weak assumption that the response variables have bounded variance, we prove an expectation bound on the squared $L^2(P)$ error of our estimator of order $n^{-\beta/2}$. Here, $\beta$ parametrises the interpolation space between $L^2(P)$ and $H$ containing the regression function. As far as we are aware, the analysis of an estimator in this setting has not previously been considered.

If we assume that the regression function is bounded, then we can clip the estimator and show that the clipped estimator has an expected squared $L^2(P)$ error of order $n^{-\beta/(1+\beta)}$. Under the stronger assumption that the response variables have subgaussian errors and that the regression function comes from an interpolation space between $L^\infty$ and $H$, we show that the squared $L^2(P)$ error is of order $n^{-\beta/(1+\beta)}$ with high probability. For the settings in which the regression function is bounded, we can use training and validation on the data set to obtain bounds of the same order of $n^{-\beta/(1+\beta)}$. This allows us to select the size of the norm constraint for our Ivanov regularisation without knowing which interpolation space contains the regression function. The response variables in the validation set must have subgaussian errors.

The expectation bounds of order $n^{-\beta/(1+\beta)}$ for bounded regression functions is optimal in the sense discussed at the end of the introduction. We use Ivanov regularisation instead of Tikhonov regularisation to control empirical processes over balls in the RKHS. By contrast, the analysis of Tikhonov-regularised estimators usually uses the spectral decomposition of the kernel operator (Mendelson and Neeman, 2010; Steinwart et al., 2009). Analysing the Ivanov-regularised estimator using this decomposition would give a more complete picture of the differences between Ivanov and Tikhonov regularisation for RKHS regression.

It would be useful to extend the lower bound of order $n^{-\beta/(1+\beta)}$, discussed at the end of the introduction, to the case in which the regression function lies in an interpolation space between $L^\infty$ and the RKHS. This would show that our high-probability bounds are also of optimal order. However, it is possible that estimation can be performed with a high-probability bound on the squared $L^2(P)$ error of smaller order.

## 3.10 Proof of Expectation Bound for Unbounded Regression Function

The proofs of all of the bounds in this chapter follow the outline in Section 3.5. We first prove Lemma 3.5.1.

**Proof of Lemma 3.5.1** Since $f \in A$, the definition of $\hat{f}$ gives

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - Y_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

Expanding

$$(\hat{f}(X_i) - Y_i)^2 = \Big((\hat{f}(X_i) - f(X_i)) + (f(X_i) - Y_i)\Big)^2,$$

substituting into the above and rearranging gives

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - f(X_i))^2 \leq \frac{2}{n}\sum_{i=1}^{n}(Y_i - f(X_i))(\hat{f}(X_i) - f(X_i)).$$

Substituting

$$Y_i - f(X_i) = (Y_i - g(X_i)) + (g(X_i) - f(X_i))$$

into the above and applying the Cauchy–Schwarz inequality to the second term gives

$$\begin{aligned}\|\hat{f} - f\|^2_{L^2(P_n)} \leq{}& \frac{2}{n}\sum_{i=1}^{n}(Y_i - g(X_i))(\hat{f}(X_i) - f(X_i)) \\ &+ 2\|g - f\|_{L^2(P_n)}\|\hat{f} - f\|_{L^2(P_n)}.\end{aligned}$$

For constants $a, b \in \mathbb{R}$ and a variable $x \in \mathbb{R}$, we have

$$x^2 \leq a + 2bx \implies x^2 \leq 2a + 4b^2$$

by completing the square and rearranging. Applying this result to the above inequality proves the lemma. ■

The following lemma is useful for bounding the expectation of both of the suprema in Section 3.5.

**Lemma 3.10.1** *Assume (H). Let the* $\varepsilon_i$ *be random variables on* $(\Omega, \mathcal{F}, \mathbb{P})$ *such that* $\mathbb{E}(\varepsilon_i|X) = 0$ *almost surely and* $\operatorname{var}(\varepsilon_i|X) \leq \sigma^2$ *almost surely for* $1 \leq i \leq n$ *and* $\operatorname{cov}(\varepsilon_i, \varepsilon_j|X) = 0$ *almost surely for* $1 \leq i, j \leq n$ *with* $i \neq j$*. Then*

$$\mathbb{E}\left(\sup_{f \in rB_H} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right) \leq \frac{\|k\|_\infty \sigma r}{n^{1/2}}.$$

**Proof** This proof method is due to Remark 6.1 of Sriperumbudur (2016). By the reproducing kernel property and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\sup_{f \in rB_H} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| &= \sup_{f \in rB_H} \left|\left\langle \frac{1}{n}\sum_{i=1}^n \varepsilon_i k_{X_i}, f\right\rangle_H\right| \\ &= r\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i k_{X_i}\right\|_H \\ &= r\left(\frac{1}{n^2}\sum_{i,j=1}^n \varepsilon_i \varepsilon_j k(X_i, X_j)\right)^{1/2}.\end{aligned}$$

By Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}\left(\sup_{f \in rB_H} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| \middle| X\right) &\leq r\left(\frac{1}{n^2}\sum_{i,j=1}^n \operatorname{cov}(\varepsilon_i, \varepsilon_j|X) k(X_i, X_j)\right)^{1/2} \\ &\leq r\left(\frac{\sigma^2}{n^2}\sum_{i=1}^n k(X_i, X_i)\right)^{1/2}\end{aligned}$$

almost surely and again, by Jensen's inequality, we have

$$\mathbb{E}\left(\sup_{f\in rB_H}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right)\leq r\left(\frac{\sigma^2\|k\|_\infty^2}{n}\right)^{1/2}.$$

The result follows. ■

We bound the distance between $\hat{h}_r$ and $h_r$ in the $L^2(P_n)$ norm for $r>0$ and $h_r\in rB_H$.

**Lemma 3.10.2** *Assume (Y1) and (H). Let $h_r\in rB_H$. We have*

$$\mathbb{E}\left(\|\hat{h}_r-h_r\|_{L^2(P_n)}^2\right)\leq\frac{4\|k\|_\infty\sigma r}{n^{1/2}}+4\|h_r-g\|_{L^2(P)}^2.$$

**Proof** By Lemma 3.5.1 with $A=rB_H$, we have

$$\|\hat{h}_r-h_r\|_{L^2(P_n)}^2\leq\frac{4}{n}\sum_{i=1}^{n}(Y_i-g(X_i))(\hat{h}_r(X_i)-h_r(X_i))+4\|h_r-g\|_{L^2(P_n)}^2.$$

We now bound the expectation of the right-hand side. We have

$$\mathbb{E}\left(\|h_r-g\|_{L^2(P_n)}^2\right)=\|h_r-g\|_{L^2(P)}^2.$$

Furthermore,

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i-g(X_i))h_r(X_i)\right)=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i-g(X_i)|X_i)h_r(X_i)\right)=0.$$

Finally, by Lemma 3.10.1 with $\varepsilon_i=Y_i-g(X_i)$, we have

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i-g(X_i))\hat{h}_r(X_i)\right)&\leq\mathbb{E}\left(\sup_{f\in rB_H}\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i-g(X_i))f(X_i)\right|\right)\\&\leq\frac{\|k\|_\infty\sigma r}{n^{1/2}}.\end{aligned}$$

The result follows. ■

The following lemma is useful for moving the bound on the distance between $\hat{h}_r$ and $h_r$ from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $r > 0$ and $h_r \in rB_H$.

**Lemma 3.10.3** *Assume (H). We have*

$$\mathbb{E}\left(\sup_{f \in rB_H} \left| \|f\|^2_{L^2(P_n)} - \|f\|^2_{L^2(P)} \right| \right) \leq \frac{8\|k\|^2_\infty r^2}{n^{1/2}}.$$

**Proof** Let the $\varepsilon_i$ for $1 \leq i \leq n$ be i.i.d. Rademacher random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, independent of the $X_i$. Lemma 2.3.1 of van der Vaart and Wellner (1996) shows

$$\mathbb{E}\left(\sup_{f \in rB_H} \left| \frac{1}{n}\sum_{i=1}^n f(X_i)^2 - \int f^2 dP \right| \right) \leq 2\,\mathbb{E}\left(\sup_{f \in rB_H} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)^2 \right| \right)$$

by symmetrisation. Since $|f(X_i)| \leq \|k\|_\infty r$ for all $f \in rB_H$, we find

$$\frac{f(X_i)^2}{2\|k\|_\infty r}$$

is a contraction vanishing at 0 as a function of $f(X_i)$ for all $1 \leq i \leq n$. By Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}\left(\sup_{f \in rB_H} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{f(X_i)^2}{2\|k\|_\infty r} \right| \middle| X \right) \leq 2\,\mathbb{E}\left(\sup_{f \in rB_H} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) \right| \middle| X \right)$$

almost surely. By Lemma 3.10.1 with $\sigma^2 = 1$, we have

$$\mathbb{E}\left(\sup_{f \in rB_H} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) \right| \right) \leq \frac{\|k\|_\infty r}{n^{1/2}}.$$

The result follows. ■

We move the bound on the distance between $\hat{h}_r$ and $h_r$ from the $L^2(P_n)$ norm to the

$L^2(P)$ norm for $r > 0$ and $h_r \in rB_H$.

**Corollary 3.10.4** *Assume (Y1) and (H). Let $h_r \in rB_H$. We have*

$$\mathbb{E}\left(\|\hat{h}_r - h_r\|^2_{L^2(P)}\right) \leq \frac{4\|k\|_\infty \sigma r}{n^{1/2}} + \frac{32\|k\|^2_\infty r^2}{n^{1/2}} + 4\|h_r - g\|^2_{L^2(P)}.$$

**Proof** By Lemma 3.10.2, we have

$$\mathbb{E}\left(\|\hat{h}_r - h_r\|^2_{L^2(P_n)}\right) \leq \frac{4\|k\|_\infty \sigma r}{n^{1/2}} + 4\|h_r - g\|^2_{L^2(P)}.$$

Since $\hat{h}_r - h_r \in 2rB_H$, by Lemma 3.10.3 we have

$$\begin{aligned}\mathbb{E}\left(\|\hat{h}_r - h_r\|^2_{L^2(P)} - \|\hat{h}_r - h_r\|^2_{L^2(P_n)}\right) &\leq \mathbb{E}\left(\sup_{f \in 2rB_H} \left|\|f\|^2_{L^2(P_n)} - \|f\|^2_{L^2(P)}\right|\right) \\ &\leq \frac{32\|k\|^2_\infty r^2}{n^{1/2}}.\end{aligned}$$

The result follows. ■

We bound the distance between $\hat{h}_r$ and $g$ in the $L^2(P)$ norm for $r > 0$ to prove Theorem 3.7.1.

**Proof of Theorem 3.7.1** Fix $h_r \in rB_H$. We have

$$\begin{aligned}\|\hat{h}_r - g\|^2_{L^2(P)} &\leq \left(\|\hat{h}_r - h_r\|^2_{L^2(P)} + \|h_r - g\|^2_{L^2(P)}\right)^2 \\ &\leq 2\|\hat{h}_r - h_r\|^2_{L^2(P)} + 2\|h_r - g\|^2_{L^2(P)}.\end{aligned}$$

By Corollary 3.10.4, we have

$$\mathbb{E}\left(\|\hat{h}_r - h_r\|^2_{L^2(P)}\right) \leq \frac{4\|k\|_\infty \sigma r}{n^{1/2}} + \frac{32\|k\|^2_\infty r^2}{n^{1/2}} + 4\|h_r - g\|^2_{L^2(P)}.$$

Hence,

$$\mathbb{E}\left(\|\hat{h}_r - g\|_{L^2(P)}^2\right) \leq \frac{8\|k\|_\infty \sigma r}{n^{1/2}} + \frac{64\|k\|_\infty^2 r^2}{n^{1/2}} + 10\|h_r - g\|_{L^2(P)}^2.$$

Taking an infimum over $h_r \in rB_H$ proves the result. ■

We assume $(g1)$ to prove Theorem 3.7.2.

**Proof of Theorem 3.7.2** The initial bound follows from Theorem 3.7.1 and (3.7.1). Based on this bound, setting

$$r = D_1 \|k\|_\infty^{-(1-\beta)} B n^{(1-\beta)/4}$$

gives

$$\mathbb{E}\left(\|\hat{h}_r - g\|_{L^2(P)}^2\right) \leq \left(64D_1^2 + 10D_1^{-2\beta/(1-\beta)}\right) \|k\|_\infty^{2\beta} B^2 n^{-\beta/2} + 8D_1 \|k\|_\infty^\beta B \sigma n^{-(1+\beta)/4}.$$

Hence, the next bound follows with

$$D_2 = 64D_1^2 + 10D_1^{-2\beta/(1-\beta)} \text{ and } D_3 = 8D_1.$$

■

## 3.11 Proof of Expectation Bound for Bounded Regression Function

We can obtain a bound on the distance between $V\hat{h}_r$ and $Vh_r$ in the $L^2(P_n)$ norm for $r > 0$ and $h_r \in rB_H$ from Lemma 3.10.2. The following lemma is useful for moving the bound on the distance between $V\hat{h}_r$ and $Vh_r$ from the $L^2(P_n)$ norm to the $L^2(P)$ norm.

**Lemma 3.11.1** *Assume (H). We have*

$$\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\|Vf_1-Vf_2\|^2_{L^2(P_n)}-\|Vf_2-Vf_2\|^2_{L^2(P)}\right|\right)\leq\frac{64\|k\|_\infty Cr}{n^{1/2}}.$$

**Proof** Let the $\varepsilon_i$ for $1\leq i\leq n$ be i.i.d. Rademacher random variables on $(\Omega,\mathcal{F},\mathbb{P})$, independent of the $X_i$. Lemma 2.3.1 of van der Vaart and Wellner (1996) shows

$$\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n(Vf_1(X_i)-Vf_2(X_i))^2-\int(Vf_1-Vf_2)^2dP\right|\right)$$

is at most

$$2\,\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(Vf_1(X_i)-Vf_2(X_i))^2\right|\right)$$

by symmetrisation. Since $|Vf_1(X_i)-Vf_2(X_i)|\leq 2C$ for all $f_1,f_2\in rB_H$, we find

$$\frac{(Vf_1(X_i)-Vf_2(X_i))^2}{4C}$$

is a contraction vanishing at 0 as a function of $Vf_1(X_i)-Vf_2(X_i)$ for all $1\leq i\leq n$. By Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\frac{(Vf_1(X_i)-Vf_2(X_i))^2}{4C}\right|\middle|X\right)$$

is at most

$$2\,\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(Vf_1(X_i)-Vf_2(X_i))\right|\middle|X\right)$$

almost surely. Therefore,

$$\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n(Vf_1(X_i)-Vf_2(X_i))^2-\int(Vf_1-Vf_2)^2dP\right|\right)$$

is at most

$$16C\,\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(Vf_1(X_i)-Vf_2(X_i))\right|\right)\leq 32C\,\mathbb{E}\left(\sup_{f\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_iVf(X_i)\right|\right)$$

by the triangle inequality. Again, by Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n(Vf_1(X_i)-Vf_2(X_i))^2-\int(Vf_1-Vf_2)^2dP\right|\right)$$

is at most

$$64C\,\mathbb{E}\left(\sup_{f\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_if(X_i)\right|\right)$$

since $V$ is a contraction vanishing at 0. The result follows from Lemma 3.10.1 with $\sigma^2=1$. ■

We move the bound on the distance between $V\hat{h}_r$ and $Vh_r$ from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $r>0$ and $h_r\in rB_H$.

**Corollary 3.11.2** *Assume (Y1) and (H). Let $h_r\in rB_H$. We have*

$$\mathbb{E}\left(\|V\hat{h}_r-Vh_r\|_{L^2(P)}^2\right)\leq\frac{4\|k\|_\infty(16C+\sigma)r}{n^{1/2}}+4\|h_r-g\|_{L^2(P)}^2.$$

**Proof** By Lemma 3.10.2, we have

$$\mathbb{E}\left(\|\hat{h}_r-h_r\|_{L^2(P_n)}^2\right)\leq\frac{4\|k\|_\infty\sigma r}{n^{1/2}}+4\|h_r-g\|_{L^2(P)}^2,$$

so

$$\mathbb{E}\left(\|V\hat{h}_r - Vh_r\|^2_{L^2(P_n)}\right) \leq \frac{4\|k\|_\infty \sigma r}{n^{1/2}} + 4\|h_r - g\|^2_{L^2(P)}.$$

Since $\hat{h}_r, h_r \in rB_H$, by Lemma 3.11.1 we have

$$\begin{aligned}
&\mathbb{E}\left(\|V\hat{h}_r - Vh_r\|^2_{L^2(P)} - \|V\hat{h}_r - Vh_r\|^2_{L^2(P_n)}\right) \\
\leq\ &\mathbb{E}\left(\sup_{f_1, f_2 \in rB_H} \left|\|Vf_1 - Vf_2\|^2_{L^2(P_n)} - \|Vf_2 - Vf_2\|^2_{L^2(P)}\right|\right) \\
\leq\ &\frac{64\|k\|_\infty C r}{n^{1/2}}.
\end{aligned}$$

The result follows. ■

We assume $(g2)$ to bound the distance between $V\hat{h}_r$ and $g$ in the $L^2(P)$ norm for $r > 0$ and prove Theorem 3.7.3.

**Proof of Theorem 3.7.3** Fix $h_r \in rB_H$. We have

$$\begin{aligned}
\|V\hat{h}_r - g\|^2_{L^2(P)} &\leq \left(\|V\hat{h}_r - Vh_r\|^2_{L^2(P)} + \|Vh_r - g\|^2_{L^2(P)}\right)^2 \\
&\leq 2\|V\hat{h}_r - Vh_r\|^2_{L^2(P)} + 2\|Vh_r - g\|^2_{L^2(P)} \\
&\leq 2\|V\hat{h}_r - Vh_r\|^2_{L^2(P)} + 2\|h_r - g\|^2_{L^2(P)}.
\end{aligned}$$

By Corollary 3.11.2, we have

$$\mathbb{E}\left(\|V\hat{h}_r - Vh_r\|^2_{L^2(P)}\right) \leq \frac{4\|k\|_\infty (16C + \sigma) r}{n^{1/2}} + 4\|h_r - g\|^2_{L^2(P)}.$$

Hence,

$$\mathbb{E}\left(\|V\hat{h}_r - g\|^2_{L^2(P)}\right) \leq \frac{8\|k\|_\infty (16C + \sigma) r}{n^{1/2}} + 10\|h_r - g\|^2_{L^2(P)}.$$

Taking an infimum over $h_r \in rB_H$ proves the result. ■

We assume $(g1)$ to prove Theorem 3.7.4.

**Proof of Theorem 3.7.4** The initial bound follows from Theorem 3.7.3 and (3.7.1). Based on this bound, setting

$$r = D_1 \|k\|_\infty^{-(1-\beta)/(1+\beta)} B^{2/(1+\beta)} (16C + \sigma)^{-(1-\beta)/(1+\beta)} n^{(1-\beta)/(2(1+\beta))}$$

gives

$$\mathbb{E}\left(\|V\hat{h}_r - g\|_{L^2(P)}^2\right)$$

is at most

$$\left(8D_1 + 10D_1^{-2\beta/(1-\beta)}\right) \|k\|_\infty^{2\beta/(1+\beta)} B^{2/(1+\beta)} (16C + \sigma)^{2\beta/(1+\beta)} n^{-\beta/(1+\beta)}.$$

Hence, the next bound follows with

$$D_2 = 8D_1 + 10D_1^{-2\beta/(1-\beta)}.$$

■

## 3.12 Proof of Expectation Bound for Validation

We need to introduce some definitions for stochastic processes. A stochastic process $W$ on $(\Omega, \mathcal{F})$ indexed by a metric space $(M, d)$ is $d^2$-subgaussian if it is centred and $W(s) - W(t)$ is $d(s,t)^2$-subgaussian for all $s, t \in M$. $W$ is separable if there exists a countable set $M_0 \subseteq M$ such that the following holds for all $\omega \in \Omega_0$, where $\mathbb{P}(\Omega_0) = 1$. For all $s \in M$ and $\varepsilon > 0$, $W(s)$ is in the closure of $\{W(t) : t \in M_0, d(s,t) \leq \varepsilon\}$.

We also need to introduce the concept of covering numbers for the next result. The covering number $N(M, d, \varepsilon)$ is the minimum number of $d$-balls of size $\varepsilon > 0$ needed to cover $M$.

The following lemma is useful for bounding the expectation of both of the suprema in Section 3.5.

**Lemma 3.12.1** *Assume (H). Let the $\varepsilon_i$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $(\tilde{X}_i, \varepsilon_i)$ is i.i.d. for $1 \leq i \leq \tilde{n}$ and let $\varepsilon_i$ be $\tilde{\sigma}^2$-subgaussian given $\tilde{X}_i$. Let $r_0 \in R$, $f_0 = V\hat{h}_{r_0}$ and*

$$W(f) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \varepsilon_i (f(\tilde{X}_i) - f_0(\tilde{X}_i))$$

*for $f \in F$. Then $W$ is $\tilde{\sigma}^2 \|\cdot\|_\infty^2 / \tilde{n}$-subgaussian given $\tilde{X}$ and separable on $(F, \tilde{\sigma}\|\cdot\|_\infty / \tilde{n}^{1/2})$. Furthermore,*

$$\mathbb{E}\left( \sup_{f \in F} |W(f)| \right) \leq \frac{4C\tilde{\sigma}}{\tilde{n}^{1/2}} \left( \left( 2 \log \left( 2 + \frac{\|k\|_\infty^2 \rho^2}{C^2} \right) \right)^{1/2} + \pi^{1/2} \right).$$

**Proof** Let $W_i(f) = \varepsilon_i(f(\tilde{X}_i) - f_0(\tilde{X}_i))$ for $1 \leq i \leq \tilde{n}$ and $f \in F$. Note that the $W_i$ are independent and centred. Since $W_i(f_1) - W_i(f_2)$ is $\tilde{\sigma}^2 \|f_1 - f_2\|_\infty^2$-subgaussian given $\tilde{X}_i$ for $1 \leq i \leq \tilde{n}$ and $f_1, f_2 \in F$, the process $W$ is $\tilde{\sigma}^2 \|\cdot\|_\infty^2 / \tilde{n}$-subgaussian given $\tilde{X}$. By Lemma 3.15.2, we have that $(F, \tilde{\sigma}\|\cdot\|_\infty / \tilde{n}^{1/2})$ is separable. Hence, $W$ is separable on $(F, \tilde{\sigma}\|\cdot\|_\infty / \tilde{n}^{1/2})$ since it is continuous. The diameter of $(F, \tilde{\sigma}\|\cdot\|_\infty / \tilde{n}^{1/2})$ is

$$D = \sup_{f_1, f_2 \in F} \tilde{\sigma} \|f_1 - f_2\|_\infty / \tilde{n}^{1/2} \leq 2C\tilde{\sigma} / \tilde{n}^{1/2}.$$

We have

$$\begin{aligned} \int_0^\infty (\log(N(F, \tilde{\sigma}\|\cdot\|_\infty / \tilde{n}^{1/2}, \varepsilon)))^{1/2} d\varepsilon &= \int_0^\infty (\log(N(F, \|\cdot\|_\infty, \tilde{n}^{1/2}\varepsilon / \tilde{\sigma})))^{1/2} d\varepsilon \\ &= \frac{\tilde{\sigma}}{\tilde{n}^{1/2}} \int_0^\infty (\log(N(F, \|\cdot\|_\infty, u)))^{1/2} du. \end{aligned}$$

This is finite by Lemma 3.17.2. Hence, by Theorem 2.3.7 of Giné and Nickl (2016) and Lemma 3.17.2, we have

$$\mathbb{E}\left(\sup_{f\in F}|W(f)|\middle|\tilde{X},X,Y\right)$$

is at most

$$\begin{aligned}
&\mathbb{E}(|W(f_0)|\,|\tilde{X},X,Y)+2^{5/2}\int_0^{C\tilde{\sigma}/\tilde{n}^{1/2}}(\log(2N(F,\tilde{\sigma}\|\cdot\|_\infty/\tilde{n}^{1/2},\varepsilon)))^{1/2}d\varepsilon\\
&=2^{5/2}\int_0^{C\tilde{\sigma}/\tilde{n}^{1/2}}(\log(2N(F,\|\cdot\|_\infty,\tilde{n}^{1/2}\varepsilon/\tilde{\sigma})))^{1/2}d\varepsilon\\
&=\frac{2^{5/2}\tilde{\sigma}}{\tilde{n}^{1/2}}\int_0^{C}(\log(2N(F,\|\cdot\|_\infty,u)))^{1/2}du\\
&\leq\frac{2^{5/2}\tilde{\sigma}}{\tilde{n}^{1/2}}\left(\left(\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}C+\left(\frac{\pi}{2}\right)^{1/2}C\right)\\
&=\frac{4C\tilde{\sigma}}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)
\end{aligned}$$

almost surely, noting $W(f_0)=0$. The result follows. ■

We bound the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ in the $L^2(\tilde{P}_{\tilde{n}})$ norm for $r_0\in R$.

**Lemma 3.12.2** *Assume (H) and ($\tilde{Y}$). Let $r_0\in R$. We have*

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}}-V\hat{h}_{r_0}\|_{L^2(\tilde{P}_{\tilde{n}})}^2\right)$$

*is at most*

$$\frac{16C\tilde{\sigma}}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)+4\,\mathbb{E}\left(\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right).$$

**Proof** By Lemma 3.5.1 with $A=F$ and $n$, $X$, $Y$ and $P_n$ replaced by $\tilde{n}$, $\tilde{X}$, $\tilde{Y}$ and

$\tilde{P}_{\tilde{n}}$, we have

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|^2_{L^2(\tilde{P}_{\tilde{n}})} \leq \frac{4}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{Y}_i - g(\tilde{X}_i))(V\hat{h}_{\hat{r}}(\tilde{X}_i) - V\hat{h}_{r_0}(\tilde{X}_i)) + 4\|V\hat{h}_{r_0} - g\|^2_{L^2(\tilde{P}_{\tilde{n}})}.$$

We now bound the expectation of the right-hand side. We have

$$\mathbb{E}\left(\|V\hat{h}_{r_0} - g\|^2_{L^2(\tilde{P}_{\tilde{n}})}\right) = \mathbb{E}\left(\|V\hat{h}_{r_0} - g\|^2_{L^2(P)}\right).$$

Let $f_0 = V\hat{h}_{r_0}$. By Lemma 3.12.1 with $\varepsilon_i = Y_i - g(X_i)$, we have

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{Y}_i - g(\tilde{X}_i))(V\hat{h}_{\hat{r}}(\tilde{X}_i) - V\hat{h}_{r_0}(\tilde{X}_i))\right)\\
&\leq \mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{Y}_i - g(\tilde{X}_i))(f(\tilde{X}_i) - f_0(\tilde{X}_i))\right|\right)\\
&\leq \frac{4C\tilde{\sigma}}{\tilde{n}^{1/2}}\left(\left(2\log\left(2 + \frac{\|k\|^2_\infty\rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right).
\end{aligned}$$

The result follows. ■

The following lemma is useful for moving the bound on the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ from the $L^2(\tilde{P}_{\tilde{n}})$ norm to the $L^2(P)$ norm for $r_0 \in R$.

**Lemma 3.12.3** *Assume (H). Let $r_0 \in R$ and $f_0 = V\hat{h}_{r_0}$. We have*

$$\mathbb{E}\left(\sup_{f\in F}\left|\|f - f_0\|^2_{L^2(\tilde{P}_{\tilde{n}})} - \|f - f_0\|^2_{L^2(P)}\right|\right) \leq \frac{64C^2}{\tilde{n}^{1/2}}\left(\left(2\log\left(2 + \frac{\|k\|^2_\infty\rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right).$$

**Proof** Let the $\varepsilon_i$ be i.i.d. Rademacher random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ for $1 \leq i \leq \tilde{n}$, independent of $\tilde{X}$, $X$ and $Y$. Lemma 2.3.1 of van der Vaart and Wellner (1996) shows

$$\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(f(\tilde{X}_i) - f_0(\tilde{X}_i))^2 - \int(f - f_0)^2dP\right|\,\middle|\,X, Y\right)$$

is at most

$$2\,\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\varepsilon_i(f(\tilde{X}_i)-f_0(\tilde{X}_i))^2\right|\,\middle|X,Y\right)$$

almost surely by symmetrisation. Since $|f(\tilde{X}_i)-f_0(\tilde{X}_i)|\le 2C$ for all $f\in F$, we find

$$\frac{(f(\tilde{X}_i)-f_0(\tilde{X}_i))^2}{4C}$$

is a contraction vanishing at 0 as a function of $f(\tilde{X}_i)-f_0(\tilde{X}_i)$ for all $1\le i\le\tilde{n}$. By Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\varepsilon_i\frac{(f(\tilde{X}_i)-f_0(\tilde{X}_i))^2}{4C}\right|\,\middle|\tilde{X},X,Y\right)$$

is at most

$$2\,\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\varepsilon_i(f(\tilde{X}_i)-f_0(\tilde{X}_i))\right|\,\middle|\tilde{X},X,Y\right)$$

almost surely. Therefore,

$$\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(f(\tilde{X}_i)-f_0(\tilde{X}_i))^2-\int(f-f_0)^2dP\right|\,\middle|X,Y\right)$$

is at most

$$16C\,\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\varepsilon_i(f(\tilde{X}_i)-f_0(\tilde{X}_i))\right|\,\middle|X,Y\right)$$

almost surely. The result follows from Lemma 3.12.1 with $\tilde{\sigma}^2=1$. ■

We move the bound on the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ from the $L^2(\tilde{P}_{\tilde{n}})$ norm to the $L^2(P)$ norm for $r_0\in R$.

**Corollary 3.12.4** *Assume (H) and ($\tilde{Y}$). Let $r_0\in R$. We have*

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}}-V\hat{h}_{r_0}\|^2_{L^2(P)}\right)$$

*is at most*

$$\frac{16C(4C+\tilde{\sigma})}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)+4\,\mathbb{E}\left(\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right).$$

**Proof** By Lemma 3.12.2, we have

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}}-V\hat{h}_{r_0}\|_{L^2(\tilde{P}_{\tilde{n}})}^2\right)$$

is at most

$$\frac{16C\tilde{\sigma}}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)+4\,\mathbb{E}\left(\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right).$$

Let $f_0=V\hat{h}_{r_0}$. Since $\hat{h}_{\hat{r}}\in F$, by Lemma 3.12.3 we have

$$\begin{aligned}&\mathbb{E}\left(\|V\hat{h}_{\hat{r}}-V\hat{h}_{r_0}\|_{L^2(P)}^2-\|V\hat{h}_{\hat{r}}-V\hat{h}_{r_0}\|_{L^2(\tilde{P}_{\tilde{n}})}^2\right)\\ \leq\ &\mathbb{E}\left(\sup_{f\in F}\left|\|f-f_0\|_{L^2(\tilde{P}_{\tilde{n}})}^2-\|f-f_0\|_{L^2(P)}^2\right|\right)\\ \leq\ &\frac{64C^2}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right).\end{aligned}$$

The result follows. ■

We bound the distance between $V\hat{h}_{\hat{r}}$ and $g$ in the $L^2(P)$ norm to prove Theorem 3.7.5.

**Proof of Theorem 3.7.5** We have

$$\begin{aligned}\|V\hat{h}_{\hat{r}}-g\|_{L^2(P)}^2&\leq\left(\|V\hat{h}_{\hat{r}}-Vh_{r_0}\|_{L^2(P)}^2+\|Vh_{r_0}-g\|_{L^2(P)}^2\right)^2\\&\leq 2\|V\hat{h}_{\hat{r}}-Vh_{r_0}\|_{L^2(P)}^2+2\|Vh_{r_0}-g\|_{L^2(P)}^2.\end{aligned}$$

By Corollary 3.12.4, we have

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|_{L^2(P)}^2\right)$$

is at most

$$\frac{16C(4C+\tilde{\sigma})}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right) + 4\,\mathbb{E}\left(\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2\right).$$

Hence,

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2\right)$$

is at most

$$\frac{32C(4C+\tilde{\sigma})}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right) + 10\,\mathbb{E}\left(\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2\right).$$

■

We assume the conditions of Theorem 3.7.4 to prove Theorem 3.7.6.

**Proof of Theorem 3.7.6** If we assume $(R1)$, then $r_0 = an^{(1-\beta)/(2(1+\beta))} \in R$ and

$$\mathbb{E}\left(\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2\right) \leq \frac{8\|k\|_\infty(16C+\sigma)an^{(1-\beta)/(2(1+\beta))}}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}$$

by Theorem 3.7.4. If we assume $(R2)$, then there is at least one $r_0 \in R$ such that

$$an^{(1-\beta)/(2(1+\beta))} \leq r_0 < an^{(1-\beta)/(2(1+\beta))} + b$$

and

$$\mathbb{E}\left(\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2\right) \leq \frac{8\|k\|_\infty(16C+\sigma)r_0}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{r_0^{2\beta/(1-\beta)}}$$

$$\leq \frac{8\|k\|_\infty(16C+\sigma)\left(an^{(1-\beta)/(2(1+\beta))}+b\right)}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}$$

by Theorem 3.7.4. In either case,

$$\mathbb{E}\left(\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2\right) \leq D_2 n^{-\beta/(1+\beta)}$$

for some constant $D_2 > 0$ not depending on $n$ or $\tilde{n}$. By Theorem 3.7.5, we have

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2\right) \leq D_3 \log(n)^{1/2}\tilde{n}^{-1/2} + 10D_2 n^{-\beta/(1+\beta)}$$

for some constant $D_3 > 0$ not depending on $n$ or $\tilde{n}$. Since $\tilde{n}$ increases at least linearly in $n$, there exists some constant $D_4 > 0$ such that $\tilde{n} \geq D_4 n$. We then have

$$\begin{aligned}\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2\right) &\leq D_4^{-1/2} D_3 \log(n)^{1/2} n^{-1/2} + 10D_2 n^{-\beta/(1+\beta)} \\ &\leq D_1 n^{-\beta/(1+\beta)}\end{aligned}$$

for some constant $D_1 > 0$ not depending on $n$ or $\tilde{n}$. ■

## 3.13 Proof of High-Probability Bound for Bounded Regression Function

We bound the distance between $\hat{h}_r$ and $h_r$ in the $L^2(P_n)$ norm for $r > 0$ and $h_r \in rB_H$.

**Lemma 3.13.1** *Assume (Y2) and (H). Let $r > 0$, $h_r \in rB_H$ and $t \geq 1$. With probability at least $1 - 2e^{-t}$, we have*

$$\|\hat{h}_r - h_r\|_{L^2(P_n)}^2 \leq \frac{20\|k\|_\infty \sigma r t^{1/2}}{n^{1/2}} + 4\|h_r - g\|_\infty^2.$$

**Proof** By Lemma 3.5.1 with $A = rB_H$, we have

$$\|\hat{h}_r - h_r\|_{L^2(P_n)}^2 \leq \frac{4}{n}\sum_{i=1}^n (Y_i - g(X_i))(\hat{h}_r(X_i) - h_r(X_i)) + 4\|h_r - g\|_{L^2(P_n)}^2.$$

We now bound the right-hand side. We have

$$\|h_r - g\|_{L^2(P_n)}^2 \leq \|h_r - g\|_\infty^2.$$

Furthermore,

$$-\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))h_r(X_i)$$

is subgaussian given $X$ with parameter

$$\frac{1}{n^2}\sum_{i=1}^n \sigma^2 h_r(X_i)^2 \leq \frac{\|k\|_\infty^2 \sigma^2 r^2}{n}.$$

So we have

$$-\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))h_r(X_i) \leq \frac{\|k\|_\infty \sigma r (2t)^{1/2}}{n^{1/2}} \leq \frac{2\|k\|_\infty \sigma r t^{1/2}}{n^{1/2}}$$

with probability at least $1 - e^{-t}$ by Chernoff bounding. Finally, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))\hat{h}_r(X_i) &\leq \sup_{f \in rB_H} \left| \frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i)) f(X_i) \right| \\
&= \sup_{f \in rB_H} \left| \left\langle \frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i)) k_{X_i}, f \right\rangle_H \right| \\
&= r \left\| \frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i)) k_{X_i} \right\|_H \\
&= r \left( \frac{1}{n^2}\sum_{i,j=1}^n (Y_i - g(X_i))(Y_j - g(X_j)) k(X_i, X_j) \right)^{1/2}
\end{aligned}$$

by the reproducing kernel property and the Cauchy–Schwarz inequality. Let $K$ be the

$n \times n$ matrix with $K_{i,j} = k(X_i, X_j)$ and let $\varepsilon$ be the vector of the $Y_i - g(X_i)$. Then

$$\frac{1}{n^2} \sum_{i,j=1}^{n} (Y_i - g(X_i))(Y_j - g(X_j)) k(X_i, X_j) = \varepsilon^{\mathsf{T}} (n^{-2} K) \varepsilon.$$

Furthermore, since $k$ is a measurable function on $(S \times S, \mathcal{S} \otimes \mathcal{S})$, we have that $n^{-2}K$ is an $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$ and non-negative-definite. Let $a_i$ for $1 \leq i \leq n$ be the eigenvalues of $n^{-2}K$. Then

$$\max_i a_i \leq \operatorname{tr}(n^{-2}K) \leq n^{-1} \|k\|_\infty^2$$

and

$$\operatorname{tr}((n^{-2}K)^2) = \|a\|_2^2 \leq \|a\|_1^2 \leq n^{-2} \|k\|_\infty^4.$$

Therefore, by Lemma 3.16.2 with $M = n^{-2}K$, we have

$$\varepsilon^{\mathsf{T}} (n^{-2} K) \varepsilon \leq \|k\|_\infty^2 \sigma^2 n^{-1} (1 + 2t + 2(t^2 + t)^{1/2})$$

and

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) \hat{h}_r(X_i) \leq \frac{3\|k\|_\infty \sigma r t^{1/2}}{n^{1/2}}$$

with probability at least $1 - e^{-t}$. The result follows. ■

The following lemma is useful for bounding the supremum in (3.5.2).

**Lemma 3.13.2** *Let $D > 0$ and $A \subseteq L^\infty$ be separable with $\|f\|_\infty \leq D$ for all $f \in A$. Let*

$$Z = \sup_{f \in A} \left| \|f\|_{L^2(P_n)}^2 - \|f\|_{L^2(P)}^2 \right|.$$

*Then, for $t > 0$, we have*

$$Z \leq \mathbb{E}(Z) + \left( \frac{2D^4 t}{n} + \frac{4D^2\, \mathbb{E}(Z) t}{n} \right)^{1/2} + \frac{2D^2 t}{3n}$$

*with probability at least* $1 - e^{-t}$.

**Proof** We have

$$Z = \sup_{f \in A} \left| \sum_{i=1}^{n} n^{-1} \left( f(X_i)^2 - \|f\|_{L^2(P)}^2 \right) \right|$$

and

$$\begin{aligned} \mathbb{E} \left( n^{-1} \left( f(X_i)^2 - \|f\|_{L^2(P)}^2 \right) \right) &= 0, \\ n^{-1} \left| f(X_i)^2 - \|f\|_{L^2(P)}^2 \right| &\leq \frac{D^2}{n}, \\ \mathbb{E} \left( n^{-2} \left( f(X_i)^2 - \|f\|_{L^2(P)}^2 \right)^2 \right) &\leq \frac{D^4}{n^2} \end{aligned}$$

for all $1 \leq i \leq n$ and $f \in A$. Furthermore, $A$ is separable, so $Z$ is a random variable on $(\Omega, \mathcal{F})$ and we can use Talagrand's inequality (Theorem A.9.1 of Steinwart and Christmann, 2008) to show

$$Z > \mathbb{E}(Z) + \left( 2t \left( \frac{D^4}{n} + \frac{2D^2 \, \mathbb{E}(Z)}{n} \right) \right)^{1/2} + \frac{2tD^2}{3n}$$

with probability at most $e^{-t}$. The result follows. ■

The following lemma is useful for moving the bound on the distance between $V\hat{h}_r$ and $Vh_r$ from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $r > 0$ and $h_r \in rB_H$.

**Lemma 3.13.3** *Assume (H). Let* $r > 0$ *and* $t \geq 1$. *With probability at least* $1 - e^{-t}$, *we have*

$$\sup_{f_1, f_2 \in rB_H} \left| \|Vf_1 - Vf_2\|_{L^2(P_n)}^2 - \|Vf_1 - Vf_2\|_{L^2(P)}^2 \right|$$

*is at most*

$$\frac{8 \left( C^2 + 4\|k\|_\infty^{1/2} C^{3/2} r^{1/2} + 8\|k\|_\infty Cr \right) t^{1/2}}{n^{1/2}} + \frac{8C^2 t}{3n}.$$

**Proof** Let $A = \{Vf_1 - Vf_2 : f_1, f_2 \in rB_H\}$ and

$$Z = \sup_{f_1, f_2 \in rB_H} \left| \|Vf_1 - Vf_2\|^2_{L^2(P_n)} - \|Vf_1 - Vf_2\|^2_{L^2(P)} \right|.$$

Then $A \subseteq L^\infty$ is separable because $H$ is separable and has a bounded kernel $k$. Furthermore, $\|Vf_1 - Vf_2\|_\infty \leq 2C$ for all $f_1, f_2 \in rB_H$. By Lemma 3.13.2, we have

$$Z \leq \mathbb{E}(Z) + \left( \frac{32C^4 t}{n} + \frac{16C^2\,\mathbb{E}(Z)t}{n} \right)^{1/2} + \frac{8C^2 t}{3n}$$

with probability at least $1 - e^{-t}$. By Lemma 3.11.1, we have

$$\mathbb{E}(Z) \leq \frac{64\|k\|_\infty Cr}{n^{1/2}}.$$

The result follows. ■

We move the bound on the distance between $V\hat{h}_r$ and $Vh_r$ from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $r > 0$ and $h_r \in rB_H$.

**Corollary 3.13.4** *Assume (Y2) and (H). Let $r > 0$, $h_r \in rB_H$ and $t \geq 1$. With probability at least $1 - 3e^{-t}$, we have*

$$\|V\hat{h}_r - Vh_r\|^2_{L^2(P)}$$

*is at most*

$$\frac{4\left(2C^2 + 8\|k\|_\infty^{1/2} C^{3/2} r^{1/2} + \|k\|_\infty (16C + 5\sigma) r\right) t^{1/2}}{n^{1/2}} + \frac{8C^2 t}{3n} + 4\|h_r - g\|^2_\infty.$$

**Proof** By Lemma 3.13.1, we have

$$\|\hat{h}_r - h_r\|^2_{L^2(P_n)} \leq \frac{20\|k\|_\infty \sigma r t^{1/2}}{n^{1/2}} + 4\|h_r - g\|^2_\infty.$$

with probability at least $1 - 2e^{-t}$, so

$$\|V\hat{h}_r - Vh_r\|_{L^2(P_n)}^2 \leq \frac{20\|k\|_\infty \sigma r t^{1/2}}{n^{1/2}} + 4\|h_r - g\|_\infty^2.$$

Since $\hat{h}_r, h_r \in rB_H$, by Lemma 3.13.3 we have

$$\begin{aligned}
&\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 - \|V\hat{h}_r - Vh_r\|_{L^2(P_n)}^2 \\
\leq{}& \sup_{f_1, f_2 \in rB_H} \left| \|Vf_1 - Vf_2\|_{L^2(P_n)}^2 - \|Vf_2 - Vf_2\|_{L^2(P)}^2 \right| \\
\leq{}& \frac{8\left(C^2 + 4\|k\|_\infty^{1/2} C^{3/2} r^{1/2} + 8\|k\|_\infty Cr\right) t^{1/2}}{n^{1/2}} + \frac{8C^2 t}{3n}
\end{aligned}$$

with probability at least $1 - e^{-t}$. The result follows. ■

We assume $(g2)$ to bound the distance between $V\hat{h}_r$ and $g$ in the $L^2(P)$ norm for $r > 0$ and prove Theorem 3.8.1.

**Proof of Theorem 3.8.1** Fix $h_r \in rB_H$. We have

$$\begin{aligned}
\|V\hat{h}_r - g\|_{L^2(P)}^2 &\leq \left(\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + \|Vh_r - g\|_{L^2(P)}^2\right)^2 \\
&\leq 2\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + 2\|Vh_r - g\|_{L^2(P)}^2 \\
&\leq 2\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + 2\|h_r - g\|_{L^2(P)}^2.
\end{aligned}$$

By Corollary 3.13.4, we have

$$\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2$$

is at most

$$\frac{4\left(2C^2 + 8\|k\|_\infty^{1/2} C^{3/2} r^{1/2} + \|k\|_\infty (16C + 5\sigma) r\right) t^{1/2}}{n^{1/2}} + \frac{8C^2 t}{3n} + 4\|h_r - g\|_\infty^2.$$

with probability at least $1 - 3e^{-t}$. Hence,

$$\|V\hat{h}_r - g\|_{L^2(P)}^2$$

is at most

$$\frac{8\left(2C^2 + 8\|k\|_\infty^{1/2}C^{3/2}r^{1/2} + \|k\|_\infty(16C + 5\sigma)r\right)t^{1/2}}{n^{1/2}} + \frac{16C^2t}{3n} + 10\|h_r - g\|_\infty^2.$$

Taking a sequence of $h_{r,n} \in rB_H$ for $n \geq 1$ with

$$\|h_{r,n} - g\|_\infty^2 \downarrow I_\infty(g, r)$$

as $n \to \infty$ proves the result. ■

We assume $(g3)$ to prove Theorem 3.8.2.

**Proof of Theorem 3.8.2** The initial bound follows from Theorem 3.8.1 and (3.8.3). Based on this bound, setting

$$r = D_1\|k\|_\infty^{-(1-\beta)/(1+\beta)}B^{2/(1+\beta)}(16C + 5\sigma)^{-(1-\beta)/(1+\beta)}t^{-(1-\beta)/(2(1+\beta))}n^{(1-\beta)/(2(1+\beta))}$$

gives

$$\|V\hat{h}_r - g\|_{L^2(P)}^2$$

is at most

$$\left(8D_1 + 10D_1^{-2\beta/(1-\beta)}\right)\|k\|_\infty^{2\beta/(1+\beta)}B^{2/(1+\beta)}(16C + 5\sigma)^{2\beta/(1+\beta)}t^{\beta/(1+\beta)}n^{-\beta/(1+\beta)}$$
$$+ 64D_1^{1/2}\|k\|_\infty^{\beta/(1+\beta)}B^{1/(1+\beta)}C^{3/2}(16C + 5\sigma)^{-(1-\beta)/(2(1+\beta))}t^{(1+3\beta)/(4(1+\beta))}n^{-(1+3\beta)/(4(1+\beta))}$$
$$+ 16C^2t^{1/2}n^{-1/2} + 16C^2tn^{-1}/3.$$

Hence, the next bound follows with

$$D_2 = 8D_1 + 10D_1^{-2\beta/(1-\beta)},\ D_3 = 64D_1^{1/2},\ D_4 = 16 \text{ and } D_5 = 16/3.$$

■

## 3.14 Proof of High-Probability Bound for Validation

We need to introduce some new notation for the next result. Let $U$ and $V$ be random variables on $(\Omega, \mathcal{F})$. Then

$$\|U\|_{\psi_2} = \inf\{a \in (0,\infty) : \mathbb{E}\,\psi_2(|U|/a) \leq 1\},$$
$$\|U|V\|_{\psi_2} = \inf\{a \in (0,\infty) : \mathbb{E}(\psi_2(|U|/a)|V) \leq 1 \text{ almost surely}\},$$

where $\psi_2(x) = \exp(x^2) - 1$ for $x \in \mathbb{R}$. Note that these infima are attained by the monotone convergence theorem. Exercise 5 of Section 2.3 of Giné and Nickl (2016) shows that $\|U\|_{\psi_2}$ is a norm on the space of $U$ such that $\|U\|_{\psi_2} < \infty$ and $\|U|V\|_{\psi_2}$ is a norm on the space of $U$ such that $\|U|V\|_{\psi_2} < \infty$.

We bound the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ in the $L^2(\tilde{P}_{\tilde{n}})$ norm for $r_0 \in R$.

**Lemma 3.14.1** *Assume (H) and ($\tilde{Y}$). Let $r_0 \in R$ and $t \geq 1$. With probability at least $1 - 2e^{-t}$, we have*

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|^2_{L^2(\tilde{P}_{\tilde{n}})}$$

*is at most*

$$\frac{292C\tilde{\sigma}t^{1/2}}{\tilde{n}^{1/2}}\left(\left(2\log\left(1+\frac{\|k\|_\infty^2\rho^2}{8C^2}\right)\right)^{1/2}+\pi^{1/2}\right)+\frac{24C^2t^{1/2}}{\tilde{n}^{1/2}}+4\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2.$$

**Proof** By Lemma 3.5.1 with $A = F$ and $n$, $X$, $Y$ and $P_n$ replaced by $\tilde{n}$, $\tilde{X}$, $\tilde{Y}$ and $\tilde{P}_{\tilde{n}}$, we have

$$\|V\hat{h}_{\hat{r}}-V\hat{h}_{r_0}\|_{L^2(\tilde{P}_{\tilde{n}})}^2\le\frac{4}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{Y}_i-g(\tilde{X}_i))(V\hat{h}_{\hat{r}}(\tilde{X}_i)-V\hat{h}_{r_0}(\tilde{X}_i))+4\|V\hat{h}_{r_0}-g\|_{L^2(\tilde{P}_{\tilde{n}})}^2.$$

We now bound the right-hand side. We have

$$\|V\hat{h}_{r_0}-g\|_{L^2(\tilde{P}_{\tilde{n}})}^2=\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\left((V\hat{h}_{r_0}(\tilde{X}_i)-g(\tilde{X}_i))^2-\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right)+\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2.$$

Since

$$\left|(V\hat{h}_{r_0}(\tilde{X}_i)-g(\tilde{X}_i))^2-\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right|\le 4C^2$$

for all $1\le i\le\tilde{n}$, we find

$$\|V\hat{h}_{r_0}-g\|_{L^2(\tilde{P}_{\tilde{n}})}^2-\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2>t$$

with probability at most

$$\exp\left(-\frac{\tilde{n}t^2}{32C^4}\right).$$

by Hoeffding's inequality. Therefore, we have

$$\|V\hat{h}_{r_0}-g\|_{L^2(\tilde{P}_{\tilde{n}})}^2-\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\le\frac{32^{1/2}C^2t^{1/2}}{\tilde{n}^{1/2}}\le\frac{6C^2t^{1/2}}{\tilde{n}^{1/2}}$$

with probability at least $1 - e^{-t}$. Now let $f_0 = V\hat{h}_{r_0}$ and

$$W(f) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (\tilde{Y}_i - g(\tilde{X}_i))(f(\tilde{X}_i) - f_0(\tilde{X}_i))$$

for $f \in F$. $W$ is $\tilde{\sigma}^2 \|\cdot\|_\infty^2/\tilde{n}$-subgaussian given $\tilde{X}$ and separable on $(F, \tilde{\sigma}\|\cdot\|_\infty/\tilde{n}^{1/2})$ by Lemma 3.12.1. The diameter of $(F, \tilde{\sigma}\|\cdot\|_\infty/\tilde{n}^{1/2})$ is

$$D = \sup_{f_1, f_2 \in F} \tilde{\sigma}\|f_1 - f_2\|_\infty/\tilde{n}^{1/2} \leq 2C\tilde{\sigma}/\tilde{n}^{1/2}.$$

From Lemma 3.17.2, we have

$$\begin{aligned}\int_0^\infty (\log(N(F, \tilde{\sigma}\|\cdot\|_\infty/\tilde{n}^{1/2}, \varepsilon)))^{1/2} d\varepsilon &= \int_0^\infty (\log(N(F, \|\cdot\|_\infty, \tilde{n}^{1/2}\varepsilon/\tilde{\sigma})))^{1/2} d\varepsilon \\ &= \frac{\tilde{\sigma}}{\tilde{n}^{1/2}} \int_0^\infty (\log(N(F, \|\cdot\|_\infty, u)))^{1/2} du\end{aligned}$$

is finite. Hence, by Exercise 1 of Section 2.3 of Giné and Nickl (2016) and Lemma 3.17.2, we have

$$\left\| \sup_{f \in F} |W(f)| \middle| \tilde{X}, X, Y \right\|_{\psi_2}$$

is at most

$$\begin{aligned}&\left\| W(f_0) \middle| \tilde{X}, X, Y \right\|_{\psi_2} + 1536^{1/2} \int_0^{2C\tilde{\sigma}/\tilde{n}^{1/2}} (\log N(F, \tilde{\sigma}\|\cdot\|_\infty/\tilde{n}^{1/2}, \varepsilon))^{1/2} d\varepsilon \\ &= 1536^{1/2} \int_0^{2C\tilde{\sigma}/\tilde{n}^{1/2}} (\log N(F, \|\cdot\|_\infty, \tilde{n}^{1/2}\varepsilon/\tilde{\sigma}))^{1/2} d\varepsilon \\ &= \frac{1536^{1/2}\tilde{\sigma}}{\tilde{n}^{1/2}} \int_0^{2C} (\log N(F, \|\cdot\|_\infty, u))^{1/2} du \\ &\leq \frac{1536^{1/2}\tilde{\sigma}}{\tilde{n}^{1/2}} \left( 2 \left( \log \left( 1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2} \right) \right)^{1/2} C + (2\pi)^{1/2} C \right) \\ &= \frac{3072^{1/2} C\tilde{\sigma}}{\tilde{n}^{1/2}} \left( \left( 2 \log \left( 1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2} \right) \right)^{1/2} + \pi^{1/2} \right),\end{aligned}$$

noting $W(f_0) = 0$. By Chernoff bounding, we have $\sup_{f \in F} |W(f)|$ is at most

$$\frac{3072^{1/2} C\tilde{\sigma}(t + \log(2))^{1/2}}{\tilde{n}^{1/2}} \left( \left( 2 \log \left( 1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2} \right) \right)^{1/2} + \pi^{1/2} \right)$$
$$\leq \frac{73C\tilde{\sigma}t^{1/2}}{\tilde{n}^{1/2}} \left( \left( 2 \log \left( 1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2} \right) \right)^{1/2} + \pi^{1/2} \right)$$

with probability at least $1 - e^{-t}$. In particular,

$$\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (\tilde{Y}_i - g(\tilde{X}_i))(V\hat{h}_{\hat{r}}(\tilde{X}_i) - V\hat{h}_{r_0}(\tilde{X}_i))$$

is at most

$$\frac{73C\tilde{\sigma}t^{1/2}}{\tilde{n}^{1/2}} \left( \left( 2 \log \left( 1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2} \right) \right)^{1/2} + \pi^{1/2} \right)$$

with probability at least $1 - e^{-t}$. The result follows. ■

The following lemma is useful for moving the bound on the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ from the $L^2(\tilde{P}_{\tilde{n}})$ norm to the $L^2(P)$ norm for $r_0 \in R$.

**Lemma 3.14.2** *Assume (H). Let $r_0 \in R$, $f_0 = V\hat{h}_{r_0}$ and $t \geq 1$. With probability at least $1 - e^{-t}$, we have*

$$\sup_{f \in F} \left| \|f - f_0\|_{L^2(\tilde{P}_{\tilde{n}})}^2 - \|f - f_0\|_{L^2(P)}^2 \right|$$

*is at most*

$$\frac{10C^2 t^{1/2}}{\tilde{n}^{1/2}} \left( 1 + 32 \left( \left( 2 \log \left( 2 + \frac{\|k\|_\infty^2 \rho^2}{C^2} \right) \right)^{1/2} + \pi^{1/2} \right) \right) + \frac{8C^2 t}{3\tilde{n}}.$$

**Proof** Let $A = \{f - f_0 : f \in F\}$ and

$$Z = \sup_{f \in F} \left| \|f - f_0\|_{L^2(\tilde{P}_{\tilde{n}})}^2 - \|f - f_0\|_{L^2(P)}^2 \right|.$$

Then $A \subseteq L^\infty$ is separable by Lemma 3.15.2. Furthermore, $\|f - f_0\|_\infty \leq 2C$ for all $f \in F$. By Lemma 3.13.2 with $n$ and $P_n$ replaced by $\tilde{n}$ and $\tilde{P}_{\tilde{n}}$, we have

$$Z \leq \mathbb{E}(Z) + \left(\frac{32C^4 t}{\tilde{n}} + \frac{16C^2\,\mathbb{E}(Z)t}{\tilde{n}}\right)^{1/2} + \frac{8C^2 t}{3\tilde{n}}$$

with probability at least $1 - e^{-t}$. By Lemma 3.12.3, we have

$$\mathbb{E}(Z) \leq \frac{64C^2}{\tilde{n}^{1/2}}\left(\left(2\log\left(2 + \frac{\|k\|_\infty^2 \rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right).$$

The result follows. ■

We move the bound on the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ from the $L^2(\tilde{P}_{\tilde{n}})$ norm to the $L^2(P)$ norm for $r_0 \in R$.

**Corollary 3.14.3** *Assume (H) and ($\tilde{Y}$). Let $r_0 \in R$ and $t \geq 1$. With probability at least $1 - 3e^{-t}$, we have*

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|_{L^2(P)}^2$$

*is at most*

$$\frac{10C(C + \tilde{\sigma})t^{1/2}}{\tilde{n}^{1/2}}\left(1 + 32\left(\left(2\log\left(2 + \frac{\|k\|_\infty^2 \rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right)\right)$$
$$+ \frac{24C^2 t^{1/2}}{\tilde{n}^{1/2}} + \frac{8C^2 t}{3\tilde{n}} + 4\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2.$$

**Proof** By Lemma 3.14.1, we have

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|_{L^2(\tilde{P}_{\tilde{n}})}^2$$

is at most

$$\frac{292C\tilde{\sigma}t^{1/2}}{\tilde{n}^{1/2}}\left(\left(2\log\left(1+\frac{\|k\|_\infty^2\rho^2}{8C^2}\right)\right)^{1/2}+\pi^{1/2}\right)+\frac{24C^2t^{1/2}}{\tilde{n}^{1/2}}+4\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2$$

with probability at least $1-2e^{-t}$. Let $f_0 = V\hat{h}_{r_0}$. Since $\hat{h}_{\hat{r}} \in F$, by Lemma 3.14.2 we have

$$\begin{aligned}
&\|V\hat{h}_{\hat{r}}-V\hat{h}_{r_0}\|_{L^2(P)}^2-\|V\hat{h}_{\hat{r}}-V\hat{h}_{r_0}\|_{L^2(\tilde{P}_{\tilde{n}})}^2\\
\le{}& \sup_{f\in F}\left|\|f-f_0\|_{L^2(\tilde{P}_{\tilde{n}})}^2-\|f-f_0\|_{L^2(P)}^2\right|\\
\le{}& \frac{10C^2t^{1/2}}{\tilde{n}^{1/2}}\left(1+32\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)\right)+\frac{8C^2t}{3\tilde{n}}
\end{aligned}$$

with probability at least $1-e^{-t}$. The result follows. ■

We bound the distance between $V\hat{h}_{\hat{r}}$ and $g$ in the $L^2(P)$ norm to prove Theorem 3.8.3.

**Proof of Theorem 3.8.3** We have

$$\begin{aligned}
\|V\hat{h}_{\hat{r}}-g\|_{L^2(P)}^2 &\le \left(\|V\hat{h}_{\hat{r}}-Vh_{r_0}\|_{L^2(P)}^2+\|Vh_{r_0}-g\|_{L^2(P)}^2\right)^2\\
&\le 2\|V\hat{h}_{\hat{r}}-Vh_{r_0}\|_{L^2(P)}^2+2\|Vh_{r_0}-g\|_{L^2(P)}^2.
\end{aligned}$$

By Corollary 3.14.3, we have

$$\|V\hat{h}_{\hat{r}}-V\hat{h}_{r_0}\|_{L^2(P)}^2$$

is at most

$$\begin{aligned}
&\frac{10C(C+\tilde{\sigma})t^{1/2}}{\tilde{n}^{1/2}}\left(1+32\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)\right)\\
&+\frac{24C^2t^{1/2}}{\tilde{n}^{1/2}}+\frac{8C^2t}{3\tilde{n}}+4\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2.
\end{aligned}$$

with probability at least $1 - 3e^{-t}$. Hence,

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2$$

is at most

$$\frac{20C(C+\tilde{\sigma})t^{1/2}}{\tilde{n}^{1/2}}\left(1 + 32\left(\left(2\log\left(2 + \frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right)\right)$$
$$+ \frac{48C^2t^{1/2}}{\tilde{n}^{1/2}} + \frac{16C^2t}{3\tilde{n}} + 10\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2.$$

The result follows. ■

We assume the conditions of Theorem 3.8.2 to prove Theorem 3.8.4.

**Proof of Theorem 3.8.4** If we assume $(R1)$, then $r_0 = an^{(1-\beta)/(2(1+\beta))} \in R$ and

$$\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2$$

is at most

$$\frac{8\left(2C^2 + 8\|k\|_\infty^{1/2}C^{3/2}a^{1/2}n^{(1-\beta)/(4(1+\beta))} + \|k\|_\infty(16C + 5\sigma)an^{(1-\beta)/(2(1+\beta))}\right)t^{1/2}}{n^{1/2}}$$
$$+ \frac{16C^2t}{3n} + \frac{10B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}.$$

with probability at least $1 - 3e^{-t}$ by Theorem 3.8.2. If we assume $(R2)$, then there is at least one $r_0 \in R$ such that

$$an^{(1-\beta)/(2(1+\beta))} \leq r_0 < an^{(1-\beta)/(2(1+\beta))} + b$$

and

$$\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2$$

is at most

$$\begin{aligned}
&\frac{8\left(2C^2+8\|k\|_\infty^{1/2}C^{3/2}r_0^{1/2}+\|k\|_\infty(16C+5\sigma)r_0\right)t^{1/2}}{n^{1/2}}+\frac{16C^2t}{3n}+\frac{10B^{2/(1-\beta)}}{r_0^{2\beta/(1-\beta)}}\\
\leq\ &\frac{8\left(2C^2+8\|k\|_\infty^{1/2}C^{3/2}\left(a^{1/2}n^{(1-\beta)/(4(1+\beta))}+b^{1/2}\right)\right)t^{1/2}}{n^{1/2}}\\
+\ &\frac{8\|k\|_\infty(16C+5\sigma)\left(an^{(1-\beta)/(2(1+\beta))}+b\right)t^{1/2}}{n^{1/2}}+\frac{16C^2t}{3n}+\frac{10B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}
\end{aligned}$$

with probability at least $1-3e^{-t}$ by Theorem 3.8.2. In either case,

$$\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\leq D_3t^{1/2}n^{-\beta/(1+\beta)}+D_4tn^{-1}$$

for some constants $D_3, D_4>0$ not depending on $n$, $\tilde{n}$ or $t$. By Theorem 3.8.3, we have

$$\|V\hat{h}_{\hat{r}}-g\|_{L^2(P)}^2\leq D_5t^{1/2}\log(n)^{1/2}\tilde{n}^{-1/2}+D_6t\tilde{n}^{-1}+10D_3t^{1/2}n^{-\beta/(1+\beta)}+10D_4tn^{-1}$$

with probability at least $1-6e^{-t}$ for some constants $D_5, D_6>0$ not depending on $n$, $\tilde{n}$ or $t$. Since $\tilde{n}$ increases at least linearly in $n$, there exists some constant $D_7>0$ such that $\tilde{n}\geq D_7n$. We then have

$$\|V\hat{h}_{\hat{r}}-g\|_{L^2(P)}^2$$

is at most

$$\begin{aligned}
&D_7^{-1/2}D_5t^{1/2}\log(n)^{1/2}n^{-1/2}+D_7^{-1}D_6tn^{-1}+10D_3t^{1/2}n^{-\beta/(1+\beta)}+10D_4tn^{-1}\\
\leq\ &D_1t^{1/2}n^{-\beta/(1+\beta)}+D_2tn^{-1}
\end{aligned}$$

for some constants $D_1, D_2>0$ not depending on $n$, $\tilde{n}$ or $t$. ■

## 3.15 Estimator Calculation and Measurability

The following result is essentially Theorem 2.1 from Quintana and Rodríguez (2014). The authors show that that a strictly-positive-definite matrix which is a $(\mathbb{C}^{n\times n}, \mathcal{B}(\mathbb{C}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$ can be diagonalised by an unitary matrix and a diagonal matrix which are both $(\mathbb{C}^{n\times n}, \mathcal{B}(\mathbb{C}^{n\times n}))$-valued measurable matrices on $(\Omega, \mathcal{F})$. The result holds for non-negative-definite matrices by adding the identity matrix before diagonalisation and subtracting it afterwards. Furthermore, the construction of the unitary matrix produces a matrix with real entries, which is to say an orthogonal matrix, when the strictly-positive-definite matrix has real entries.

**Lemma 3.15.1** *Let $M$ be a non-negative-definite matrix which is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$. There exist an orthogonal matrix $A$ and a diagonal matrix $D$ which are both $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrices on $(\Omega, \mathcal{F})$ such that $M = ADA^\mathsf{T}$.*

We prove Lemma 3.6.1.

**Proof of Lemma 3.6.1** Let $H_n = \operatorname{sp}\{k_{X_i} : 1 \leq i \leq n\}$. The subspace $H_n$ is closed in $H$, so there is an orthogonal projection $Q : H \to H_n$. Since $f - Qf \in H_n^\perp$ for all $f \in H$, we have

$$f(X_i) - (Qf)(X_i) = \langle f - Qf, k_{X_i} \rangle = 0$$

for all $1 \leq i \leq n$. Hence,

$$\begin{aligned}\inf_{f\in rB_H} \frac{1}{n}\sum_{i=1}^n (f(X_i) - Y_i)^2 &= \inf_{f\in rB_H} \frac{1}{n}\sum_{i=1}^n ((Qf)(X_i) - Y_i)^2 \\ &= \inf_{f\in (rB_H)\cap H_n} \frac{1}{n}\sum_{i=1}^n (f(X_i) - Y_i)^2.\end{aligned}$$

Let $f \in (rB_H) \cap H_n$ and write

$$f = \sum_{i=1}^{n} a_i k_{X_i}$$

for some $a \in \mathbb{R}^n$. Then

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 = n^{-1}(Ka - Y)^\mathsf{T}(Ka - Y)$$

and $\|f\|_H^2 = a^\mathsf{T} K a$, so we can write the norm constraint as $a^\mathsf{T} K a + s = r^2$, where $s \geq 0$ is a slack variable. The Lagrangian can be written as

$$\begin{aligned} L(a, s; \mu) &= n^{-1}(Ka - Y)^\mathsf{T}(Ka - Y) + \mu(a^\mathsf{T} K a + s - r^2) \\ &= a^\mathsf{T}(n^{-1}K^2 + \mu K)a - 2n^{-1}Y^\mathsf{T} K a + \mu s + n^{-1}Y^\mathsf{T} Y - \mu r^2, \end{aligned}$$

where $\mu$ is the Lagrangian multiplier for the norm constraint. We seek to minimise the Lagrangian for a fixed value of $\mu$. Note that we require $\mu \geq 0$ for the Lagrangian to have a finite minimum, due to the term in $s$. We have

$$\frac{\partial L}{\partial a} = 2(n^{-1}K^2 + \mu K)a - 2n^{-1}KY.$$

This being 0 is equivalent to $K((K + n\mu I)a - Y) = 0$.

Since the kernel $k$ is a measurable function on $(S \times S, \mathcal{S} \otimes \mathcal{S})$ and the $X_i$ are $(S, \mathcal{S})$-valued random variables on $(\Omega, \mathcal{F})$, we find that $K$ is an $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$. Furthermore, since the kernel $k$ takes real values and is non-negative definite, $K$ is non-negative definite with real entries. By Lemma 3.15.1, there exist an orthogonal matrix $A$ and a diagonal matrix $D$ which are both $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrices on $(\Omega, \mathcal{F})$ such that $K = ADA^\mathsf{T}$. Note that the diagonal entries of $D$ must be non-negative and we may assume that they are non-increasing. Inserting this diagonalisation into $K((K+n\mu I)a-Y)=0$ gives

$$AD((D+n\mu I)A^\mathsf{T}a-A^\mathsf{T}Y)=0.$$

Since $A$ has the inverse $A^\mathsf{T}$, this is equivalent to

$$D((D+n\mu I)A^\mathsf{T}a-A^\mathsf{T}Y)=0.$$

This in turn is equivalent to

$$(A^\mathsf{T}a)_i=(D_{i,i}+n\mu)^{-1}(A^\mathsf{T}Y)_i$$

for $1\le i\le m$. The same $f$ is produced for all such $a$, because if $w$ is the difference between two such $a$, then $(A^\mathsf{T}w)_i=0$ for $1\le i\le m$ and the squared $H$ norm of

$$\sum_{i=1}^{n} w_i k_{X_i}$$

is $w^\mathsf{T}Kw=w^\mathsf{T}ADA^\mathsf{T}w=0$. Hence, we are free to set $(A^\mathsf{T}a)_i=0$ for $m+1\le i\le n$. This uniquely defines $A^\mathsf{T}a$, which in turn uniquely defines $a$, since $A^\mathsf{T}$ has the inverse $A$. Note that this definition of $a$ is measurable on $(\Omega\times[0,\infty),\mathcal{F}\otimes\mathcal{B}([0,\infty)))$, where $\mu$ varies in $[0,\infty)$.

We now search for a value of $\mu$ such that $a$ and $s$ satisfy the norm constraint. We call this value $\mu(r)$. There are two cases. If

$$r^2<\sum_{i=1}^{m} D_{i,i}^{-1}(A^\mathsf{T}Y)_i^2,$$

then the $a$ above and $s=0$ minimise $L$ for $\mu=\mu(r)>0$ and satisfy the norm constraint, where $\mu(r)$ satisfies

$$\sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu(r))^2} (A^\mathsf{T} Y)_i^2 = r^2.$$

Otherwise, the $a$ above and

$$s = r^2 - \sum_{i=1}^{m} D_{i,i}^{-1} (A^\mathsf{T} Y)_i^2 \geq 0$$

minimise $L$ for $\mu = \mu(r) = 0$ and satisfy the norm constraint. Hence, the Lagrangian sufficiency theorem shows

$$\hat{h}_r = \sum_{i=1}^{n} a_i k_{X_i}$$

for the $a$ above with $\mu = \mu(r)$ for $r > 0$. We also have $\hat{h}_0 = 0$.

Since $\mu(r) > 0$ is strictly decreasing for

$$r^2 < \sum_{i=1}^{m} D_{i,i}^{-1} (A^\mathsf{T} Y)_i^2$$

and $\mu(r) = 0$ otherwise, we find

$$\{\mu(r) \leq \mu\} = \left\{ \sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu)^2} (A^\mathsf{T} Y)_i^2 \leq r^2 \right\}$$

for $\mu \in [0, \infty)$. Therefore, $\mu(r)$ is measurable on $(\Omega \times [0, \infty), \mathcal{F} \otimes \mathcal{B}((0, \infty)))$, where $r$ varies in $(0, \infty)$. Hence, the $a$ above with $\mu = \mu(r)$ for $r > 0$ is measurable on $(\Omega \times [0, \infty), \mathcal{F} \otimes \mathcal{B}((0, \infty)))$, where $r$ varies in $(0, \infty)$. By Lemma 4.25 of Steinwart and Christmann (2008), the function $\Phi : S \to H$ by $\Phi(x) = k_x$ is a $(H, \mathcal{B}(H))$-valued measurable function on $(S, \mathcal{S})$. Hence, $k_{X_i}$ for $1 \leq i \leq n$ are $(H, \mathcal{B}(H))$-valued random variables on $(\Omega, \mathcal{F})$. Together, these show that $\hat{h}_r$ is a $(H, \mathcal{B}(H))$-valued measurable function on $(\Omega \times [0, \infty), \mathcal{F} \otimes \mathcal{B}([0, \infty)))$, where $r$ varies in $[0, \infty)$, recalling that $\hat{h}_0 =$

0. ■

We prove a continuity result about our estimator.

**Lemma 3.15.2** *Let $r, s \in [0, \infty)$. We have $\|\hat{h}_r - \hat{h}_s\|_H^2 \leq |r^2 - s^2|$.*

**Proof** Recall the diagonalisation of $K = ADA^\mathsf{T}$ from Lemma 3.6.1. If $u, v \in \mathbb{R}^n$ and

$$h_1 = \sum_{i=1}^n u_i k_{X_i} \text{ and } h_2 = \sum_{i=1}^n v_i k_{X_i},$$

then $\langle h_1, h_2 \rangle_H = u^\mathsf{T} K v = (A^\mathsf{T} u)^\mathsf{T} D (A^\mathsf{T} v)$. Let $s > r$. If $r > 0$ then, by Lemma 3.6.1, we have

$$\begin{aligned}\langle \hat{h}_r, \hat{h}_s \rangle_H &= \sum_{i=1}^m \frac{D_{i,i}}{(D_{i,i} + n\mu(r))(D_{i,i} + n\mu(s))} (A^\mathsf{T} Y)_i^2 \\ &\geq \sum_{i=1}^m \frac{D_{i,i}}{(D_{i,i} + n\mu(r))^2} (A^\mathsf{T} Y)_i^2 \\ &= \|\hat{h}_r\|_H^2.\end{aligned}$$

Furthermore, again by Lemma 3.6.1, if $\mu(r) > 0$ then $\|\hat{h}_r\|_H^2 = r^2$ and

$$\begin{aligned}\|\hat{h}_r - \hat{h}_s\|_H^2 &= \|\hat{h}_r\|_H^2 + \|\hat{h}_s\|_H^2 - 2\langle \hat{h}_r, \hat{h}_s \rangle_H \\ &\leq \|\hat{h}_s\|_H^2 - \|\hat{h}_r\|_H^2 \\ &= \|\hat{h}_s\|_H^2 - r^2 \\ &\leq s^2 - r^2.\end{aligned}$$

Otherwise, $\mu(r) = 0$ and so $\mu(s) = 0$ by Lemma 3.6.1, which means $\hat{h}_r = \hat{h}_s$. If $r = 0$ then $\hat{h}_r = 0$ and $\|\hat{h}_r - \hat{h}_s\|_H^2 = \|\hat{h}_s\|_H^2 \leq s^2$. Hence, whenever $r < s$, we have $\|\hat{h}_r - \hat{h}_s\|_H^2 \leq s^2 - r^2$. The result follows. ■

We also have the estimator $\hat{r}$ when performing validation.

**Lemma 3.15.3** *We have that $\hat{r}$ is a random variable on $(\Omega, \mathcal{F})$.*

**Proof** Let

$$W(s) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (V\hat{h}_s(\tilde{X}_i) - \tilde{Y}_i)^2$$

for $s \in R$. Note that $W(s)$ is a random variable on $(\Omega, \mathcal{F})$ and continuous in $s$ by Lemma 3.15.2. Since $R \subseteq \mathbb{R}$, it is separable. Let $R_0$ be a countable dense subset of $R$. Then $\inf_{s \in R} W(s) = \inf_{s \in R_0} W(s)$ is a random variable on $(\Omega, \mathcal{F})$ as the right-hand side is the infimum of countably many random variables on $(\Omega, \mathcal{F})$. Let $r \in [0, \rho]$. By the definition of $\hat{r}$, we have

$$\{\hat{r} \leq r\} = \bigcup_{s \in R \cap [0,r]} \{W(s) \leq \inf_{t \in R} W(t)\}.$$

Since $R \cap [0, r] \subseteq \mathbb{R}$, it is separable. Let $A_r$ be a countable dense subset of $R \cap [0, r]$. By the sequential compactness of $R \cap [0, r]$ and continuity of $W(s)$, we have

$$\{\hat{r} \leq r\} = \bigcap_{a=1}^{\infty} \bigcup_{s \in A_r} \{W(s) \leq \inf_{t \in R} W(t) + a^{-1}\}.$$

This set is an element of $\mathcal{F}$. ■

## 3.16 Subgaussian Random Variables

We need the definition of a sub-$\sigma$-algebra for the next result. The $\sigma$-algebra $\mathcal{G}$ is a sub-$\sigma$-algebra of the $\sigma$-algebra $\mathcal{F}$ if $\mathcal{G} \subseteq \mathcal{F}$. The following lemma relates a quadratic form of subgaussians to that of centred normal random variables.

**Lemma 3.16.1** *Let $\varepsilon_i$ for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are* *independent conditional on some sub-$\sigma$-algebra $\mathcal{G} \subseteq \mathcal{F}$ and let*

$$\mathbb{E}(\exp(t\varepsilon_i)|\mathcal{G}) \leq \exp(\sigma^2 t^2/2)$$

*almost surely for all $t \in \mathbb{R}$. Also, let $\delta_i$ for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent of each other and $\mathcal{G}$ with $\delta_i \sim \mathrm{N}(0, \sigma^2)$. Let $M$ be an $n \times n$ non-negative-definite matrix which is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{G})$. We have*

$$\mathbb{E}(\exp(z\varepsilon^\mathsf{T} M\varepsilon)|\mathcal{G}) \leq \mathbb{E}(\exp(z\delta^\mathsf{T} M\delta)|\mathcal{G})$$

*almost surely for all $z \geq 0$.*

**Proof** This proof method uses techniques from the proof of Lemma 9 of Abbasi-Yadkori, Pál, and Szepesvári (2011). We have

$$\mathbb{E}(\exp(t_i\varepsilon_i/\sigma)|\mathcal{G}) \leq \exp(t_i^2/2)$$

almost surely for all $1 \leq i \leq n$ and $t_i \in \mathbb{R}$. Furthermore, the $\varepsilon_i$ are independent conditional on $\mathcal{G}$, so

$$\mathbb{E}(\exp(t^\mathsf{T}\varepsilon/\sigma)|\mathcal{G}) \leq \exp(\|t\|_2^2/2)$$

almost surely. By Lemma 3.15.1 with $\mathcal{F}$ replaced by $\mathcal{G}$, there exist an orthogonal matrix $A$ and a diagonal matrix $D$ which are both $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrices on $(\Omega, \mathcal{G})$ such that $M = ADA^\mathsf{T}$. Hence, $M$ has a square root $M^{1/2} = AD^{1/2}A^\mathsf{T}$ which is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{G})$, where $D^{1/2}$ is the diagonal matrix with entries equal to the square root of those of $D$. Note that these entries are non-negative because $M$ is non-negative definite. We can then replace $t$ with $sM^{1/2}u$ for $s \in \mathbb{R}$ and $u \in \mathbb{R}^n$ to get

$$\mathbb{E}(\exp(su^{\mathsf{T}}M^{1/2}\varepsilon/\sigma)|\mathcal{G}) \leq \exp(s^2\|M^{1/2}u\|_2^2/2)$$

almost surely. Integrating over $u$ with respect to the distribution of $\delta$ gives

$$\mathbb{E}(\exp(s^2\varepsilon^{\mathsf{T}}M\varepsilon/2)|\mathcal{G}) \leq \mathbb{E}(\exp(s^2\delta^{\mathsf{T}}M\delta/2)|\mathcal{G})$$

almost surely. The result follows. ■

Having established this relationship, we can now obtain a probability bound on a quadratic form of subgaussians by using Chernoff bounding. The following result is a conditional subgaussian version of the Hanson–Wright inequality.

**Lemma 3.16.2** *Let $\varepsilon_i$ for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent conditional on some sub-$\sigma$-algebra $\mathcal{G} \subseteq \mathcal{F}$ and let*

$$\mathbb{E}(\exp(t\varepsilon_i)|\mathcal{G}) \leq \exp(\sigma^2 t^2/2)$$

*almost surely for all $t \in \mathbb{R}$. Let $M$ be an $n \times n$ non-negative-definite matrix which is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{G})$ and $t \geq 0$. We have*

$$\varepsilon^{\mathsf{T}}M\varepsilon \leq \sigma^2 \operatorname{tr}(M) + 2\sigma^2\|M\|t + 2\sigma^2(\|M\|^2t^2 + \operatorname{tr}(M^2)t)^{1/2}$$

*with probability at least $1 - e^{-t}$ almost surely conditional on $\mathcal{G}$. Here, $\|M\|$ is the operator norm of $M$, which is a random variable on $(\Omega, \mathcal{G})$.*

**Proof** This proof method follows that of Theorem 3.1.9 of Giné and Nickl (2016). By Lemma 3.15.1 with $\mathcal{F}$ replaced by $\mathcal{G}$, there exist an orthogonal matrix $A$ and a diagonal matrix $D$ which are both $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrices on

$(\Omega, \mathcal{G})$ such that $M = ADA^\mathsf{T}$. Let $\delta_i$ for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent of each other and $\mathcal{G}$, with $\delta_i \sim \mathrm{N}(0, \sigma^2)$. By Lemma 3.16.1 and the fact that $A^\mathsf{T}\delta$ has the same distribution as $\delta$, we have

$$\mathbb{E}(\exp(t\varepsilon^\mathsf{T} M \varepsilon)|\mathcal{G}) \leq \mathbb{E}(\exp(t\delta^\mathsf{T} M \delta)|\mathcal{G}) = \mathbb{E}(\exp(t\delta^\mathsf{T} D \delta)|\mathcal{G})$$

almost surely for all $t \geq 0$. Furthermore,

$$\mathbb{E}(\exp(t\delta_i^2/\sigma^2)) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp(tx^2 - x^2/2)dx = \frac{1}{(1-2t)^{1/2}}$$

for $0 \leq t < 1/2$ and $1 \leq i \leq n$, so

$$\mathbb{E}(\exp(t(\delta_i^2/\sigma^2 - 1))) = \exp(-(\log(1-2t) + 2t)/2).$$

We have

$$-2(\log(1-2t) + 2t) \leq \sum_{i=2}^{\infty} (2t)^i (2/i) \leq 4t^2/(1-2t)$$

for $0 \leq t \leq 1/2$. Therefore, since the $\delta_i$ are independent of $\mathcal{G}$, we have

$$\mathbb{E}(\exp(tD_{i,i}(\delta_i^2 - \sigma^2))|\mathcal{G}) \leq \exp\left(\frac{\sigma^4 D_{i,i}^2 t^2}{1 - 2\sigma^2 D_{i,i} t}\right)$$

almost surely for $0 \leq t < 1/(2\sigma^2 D_{i,i})$ and $1 \leq i \leq n$. Since the $D_{i,i}$ are random variables on $(\Omega, \mathcal{G})$ and the $D_{i,i}\delta_i$ for $1 \leq i \leq n$ are independent conditional on $\mathcal{G}$, we have

$$\mathbb{E}(\exp(t(\delta^\mathsf{T} D \delta - \sigma^2 \operatorname{tr}(D)))|\mathcal{G}) \leq \exp\left(\frac{\sigma^4 \operatorname{tr}(D^2) t^2}{1 - 2\sigma^2 (\max_i D_{i,i}) t}\right)$$

almost surely for $0 \leq t < 1/(2\sigma^2(\max_i D_{i,i}))$. Combining this with $\mathbb{E}(\exp(t\varepsilon^\mathsf{T} M \varepsilon)|\mathcal{G}) \leq \mathbb{E}(\exp(t\delta^\mathsf{T} D \delta)|\mathcal{G})$, we find

$$\mathbb{E}(\exp(t(\varepsilon^\mathsf{T} M \varepsilon - \sigma^2 \operatorname{tr}(M)))|\mathcal{G}) \leq \exp\left(\frac{\sigma^4 \operatorname{tr}(M^2) t^2}{1 - 2\sigma^2 \|M\| t}\right)$$

almost surely for $0 \leq t < 1/(2\sigma^2 \|M\|)$. By Chernoff bounding, we have

$$\varepsilon^{\mathsf{T}} M \varepsilon - \sigma^2 \operatorname{tr}(M) > s$$

for $s \geq 0$ with probability at most

$$\exp\left( \frac{\sigma^4 \operatorname{tr}(M^2) t^2}{1 - 2\sigma^2 \|M\| t} - ts \right)$$

almost surely conditional on $\mathcal{G}$ for $0 \leq t < 1/(2\sigma^2 \|M\|)$. Letting

$$t = \frac{s}{2\sigma^4 \operatorname{tr}(M^2) + 2\sigma^2 \|M\| s}$$

gives the bound

$$\exp\left( - \frac{s^2}{4\sigma^4 \operatorname{tr}(M^2) + 4\sigma^2 \|M\| s} \right).$$

Rearranging gives the result. ■

## 3.17 Covering Numbers

The following lemma gives a bound on the covering numbers of $F$.

**Lemma 3.17.1** *Let $\varepsilon > 0$. We have*

$$N(F, \|\cdot\|_\infty, \varepsilon) \leq 1 + \frac{\|k\|_\infty^2 \rho^2}{2\varepsilon^2}.$$

**Proof** Let $a \geq 1$ and $r_i \in R$ and $f_i = V\hat{h}_{r_i} \in F$ for $1 \leq i \leq a$. Also, let $f = V\hat{h}_r \in F$ for $r \in R$. Since $V$ is a contraction, we have $\|f - f_i\|_\infty \leq \varepsilon$ whenever $\|\hat{h}_r - \hat{h}_{r_i}\|_\infty \leq \varepsilon$.

By Lemma 3.15.2, we have $\|\hat{h}_r - \hat{h}_{r_i}\|_\infty \leq \varepsilon$ whenever $|r^2 - r_i^2| \leq \varepsilon^2/\|k\|_\infty^2$. Hence, if we let $r_i^2 = \varepsilon^2(2i-1)/\|k\|_\infty^2$ and let $\rho$ be such that

$$\rho^2 - \varepsilon^2(2a-1)/\|k\|_\infty^2 \leq \varepsilon^2/\|k\|_\infty^2,$$

then we find $N(F, \|\cdot\|_\infty, \varepsilon) \leq a$. Rearranging the above shows that we can choose

$$a = \left\lceil \frac{\|k\|_\infty^2 \rho^2}{2\varepsilon^2} \right\rceil$$

and the result follows. ■

We also calculate integrals of these covering numbers.

**Lemma 3.17.2** *Let $a \geq 1$. We have*

$$\int_0^L (\log(aN(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon \leq \left(\log\left(\left(1 + \frac{\|k\|_\infty^2 \rho^2}{2L^2}\right) a\right)\right)^{1/2} L + \left(\frac{\pi}{2}\right)^{1/2} L$$

*for $L \in (0, \infty)$. When $a = 1$, we have*

$$\int_0^L (\log(N(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon \leq 2\left(\log\left(1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2}\right)\right)^{1/2} C + (2\pi)^{1/2} C$$

*for $L \in (0, \infty]$.*

**Proof** Let $L \in (0, \infty)$. Then

$$\int_0^L (\log(aN(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon \leq \int_0^L \left(\log\left(a\left(1 + \frac{\|k\|_\infty^2 \rho^2}{2\varepsilon^2}\right)\right)\right)^{1/2} d\varepsilon$$

by Lemma 3.17.1. Changing variables to $u = \varepsilon/L$ gives

$$L \int_0^1 \left(\log\left(a\left(1 + \frac{\|k\|_\infty^2 \rho^2}{2L^2 u^2}\right)\right)\right)^{1/2} du$$

$$
\begin{aligned}
&\le L\int_0^1 \left(\log\left(a\left(1+\frac{\|k\|_\infty^2\rho^2}{2L^2}\right)\frac{1}{u^2}\right)\right)^{1/2} du \\
&= L\int_0^1 \left(\log\left(a\left(1+\frac{\|k\|_\infty^2\rho^2}{2L^2}\right)\right)+\log\left(\frac{1}{u^2}\right)\right)^{1/2} du.
\end{aligned}
$$

For $b, c \ge 0$ we have $(b+c)^{1/2} \le b^{1/2} + c^{1/2}$, so the above is at most

$$
\begin{aligned}
&L\int_0^1 \left(\log\left(a\left(1+\frac{\|k\|_\infty^2\rho^2}{2L^2}\right)\right)\right)^{1/2} du + L\int_0^1 \left(\log\left(\frac{1}{u^2}\right)\right)^{1/2} du \\
&= L\left(\log\left(a\left(1+\frac{\|k\|_\infty^2\rho^2}{2L^2}\right)\right)\right)^{1/2} + L\int_0^1 \left(\log\left(\frac{1}{u^2}\right)\right)^{1/2} du.
\end{aligned}
$$

Changing variables to

$$
s = \left(\log\left(\frac{1}{u^2}\right)\right)^{1/2}
$$

shows

$$
\begin{aligned}
\int_0^1 \left(\log\left(\frac{1}{u^2}\right)\right)^{1/2} du &= \int_0^\infty s^2 \exp(-s^2/2) ds \\
&= \frac{1}{2}\int_{-\infty}^\infty s^2 \exp(-s^2/2) ds \\
&= \left(\frac{\pi}{2}\right)^{1/2},
\end{aligned}
$$

since the last integral is a multiple of the variance of an $\mathrm{N}(0,1)$ random variable. The first result follows. Note that $N(F, \|\cdot\|_\infty, \varepsilon) = 1$ whenever $\varepsilon \ge 2C$, as the ball of radius $2C$ about any point in $F$ is the whole of $F$. Hence, when $a = 1$, we have

$$
\begin{aligned}
\int_0^L (\log(N(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon &\le \int_0^\infty (\log(N(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon \\
&= \int_0^{2C} (\log(N(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon \\
&\le 2\left(\log\left(1+\frac{\|k\|_\infty^2\rho^2}{8C^2}\right)\right)^{1/2} C + (2\pi)^{1/2} C
\end{aligned}
$$

for $L \in (0, \infty]$. ■

# Chapter 4

# The Goldenshluger–Lepski Method for Constrained Least-Squares Estimators over RKHSs

In nonparametric statistics, it is assumed that the estimand belongs to a very large parameter space in order to avoid model misspecification. Such misspecification can lead to large approximation errors and poor estimator performance. However, it is often challenging to produce estimators which are robust against such large parameter spaces. An important tool which allows us to achieve this aim is adaptive estimation. Adaptive estimators behave as if they know the true model from a collection of models, despite being a function of the data. In particular, adaptive estimators can often achieve the same optimal rates of convergence as the best estimators when the true model is known.

There are many ways of creating adaptive estimators. One way is to pass information on the true model from the data to a non-adaptive estimator through tuning parameters. For example, a Gaussian kernel estimator depends on the width parameter of the Gaussian kernel. The different width parameters define different sets of functions and represent different assumptions about the estimand.

In this chapter, we study an adaptive estimation procedure called the Goldenshluger–Lepski method in the context of reproducing-kernel Hilbert space (RKHS) regression. The Goldenshluger–Lepski method works by performing pairwise comparisons between non-adaptive estimators with a range of values for the tuning parameters. As far as we are aware, this is the first time that this method has been applied in the context of RKHS regression. The Goldenshluger–Lepski method (Goldenshluger and Lepski, 2008, 2009, 2011, 2013) is an extension of Lepski's method. While Lepski's method focusses on adaptation over a single parameter, the Goldenshluger–Lepski method can be used to perform adaptation over multiple parameters.

The Goldenshluger–Lepski method operates by selecting an estimator which minimises the sum of a proxy for the unknown bias and an inflated variance term. The proxy for the bias is calculated by performing pairwise comparisons between the estimator in question and all estimators which are in some sense less smooth than this estimator. A key challenge in applying the Goldenshluger–Lepski method is proving a high-probability bound on all of these pairwise comparisons simultaneously. This bound is known as a majorant.

A popular alternative to the Goldenshluger–Lepski method for constructing adaptive estimators is training and validation. Here, the data is split into a training set and a validation set. The training set is used to produce a collection of non-adaptive estimators for a range of different values for the tuning parameters and the validation set is used to select the best estimator from this collection. This selection is performed by calculating a proxy for the cost function that we wish to minimise. The estimator with the smallest value of the proxy is selected as our final estimator. One important advantage of the Goldenshluger–Lepski method in comparison to training and validation is that it uses all of the data to calculate the non-adaptive estimators. This is because it does not require data for calculating a proxy cost function. However, the Goldenshluger–Lepski method does require us to calculate a majorant, as discussed above, which is often a challenging task.

We now describe the RKHS regression problem studied in this chapter in more detail. We assume that the regression function lies in an interpolation space between $L^\infty$ and an RKHS. Depending on the setting, this RKHS may be fixed or we may perform adaptation over a collection of RKHSs. The non-adaptive estimators we use in this context are clipped versions of least-squares estimators which are constrained to lie in a ball of predefined radius in an RKHS. These estimators are discussed in detail in Chapter 3. Constraining an estimator to lie in a ball of predefined radius is a form of Ivanov regularisation (see Oneto et al., 2016).

One advantage of the estimators that we consider is that there is a clear way of producing a majorant for them, especially when the RKHS is fixed. This is because we can control the estimator constrained to lie in a ball of radius $r$ by bounding quantities of the form $rZ$ for some random variable $Z$ which does not depend on $r$, such as in the proof of Lemma 4.5.2. It may be possible to use different non-adaptive estimators to address our RKHS regression problem, however this would require the calculation of a majorant for such estimators, which would generally be more difficult than the calculation of the majorant for the Ivanonv-regularised estimators considered in this chapter.

When the RKHS is fixed, the only tuning parameter to be selected is the radius of the ball in which the least-squares estimator is constrained to lie. Estimators for which the radius is larger are considered to be less smooth. In order to provide a majorant for the Goldenshluger–Lepski method, we must prove regression results which control these estimators for all radii simultaneously. When we perform adaptation over a collection of RKHSs, we must prove regression results which control the same estimators for all RKHSs and all ball radii in these RKHSs simultaneously. We demonstrate this approach for a collection of RKHSs with Gaussian kernels. Estimators for which both the width parameter of the Gaussian kernel is smaller and the radius of the ball in the RKHS is larger are considered to be less smooth. These results extend those of Chapter 3.

One of the main difficulties in applying the Goldenshluger–Lepski method to our RKHS regression problem is that the covariate distribution $P$, and hence the $L^2(P)$ norm, is unknown. This is a problem when trying to control the squared $L^2(P)$ error of our adaptive estimator, because the Goldenshluger–Lepski method generally requires the corresponding norm to be known. This is so that the pairwise comparisons can be performed when calculating the proxy for the unknown bias of the non-adaptive estimators. In order to get around this problem, we replace the $L^2(P)$ norm in the pairwise comparisons with its empirical counterpart, the $L^2(P_n)$ norm. Here, $P_n$ is the empirical distribution of the covariates. The terms added to our bound when moving our control on the squared $L^2(P_n)$ error of our adaptive estimator to the squared $L^2(P)$ error do not significantly increase its size.

Our main results are Theorems 4.6.5 (page 129) and 4.8.7 (page 138). These show that a fixed quantile of the squared $L^2(P)$ error of a clipped version of the estimator produced by the Goldenshluger–Lepski method is of order $n^{-\beta/(1+\beta)}$. Here, $n$ is the number of data points and $\beta$ parametrises the interpolation space between $L^\infty$ and the RKHS containing the regression function. We use $L^\infty$ when interpolating so that we have direct control over approximation errors in the $L^2(P_n)$ norm. Theorem 4.6.5 addresses the case in which the RKHS is fixed and Theorem 4.8.7 addresses the case in which we perform adaptation over a collection of RKHSs with Gaussian kernels. The order $n^{-\beta/(1+\beta)}$ for the squared $L^2(P)$ error of the adaptive estimators matches the order of the smallest bounds obtained in Chapter 3 for the squared $L^2(P)$ error of the non-adaptive estimators. In the sense discussed in Chapter 3, this order is the optimal power of $n$ if we make the slightly weaker assumption that the regression function is an element of the interpolation space between $L^2(P)$ and the RKHS parametrised by $\beta$.

## 4.1 Literature Review

Lepski's method (Lepski, 1991a,b, 1993) is a method for adaptation over a single parameter. Since its introduction it has been studied by, for example, Birgé (2001) and Giné and Nickl (2016). Lepski's method selects the smoothest non-adaptive estimator from a collection, subject to a bound on a series of pairwise comparisons involving all estimators at most as smooth as the resulting estimator. The method can only adapt to one parameter because of the need for an ordering of the collection of non-adaptive estimators.

Lepski's method has been applied to RKHS regression under the name of the balancing principle. However, as far as we are aware, Lepski's method has not been used to target the true regression function, but instead an RKHS element which approximates the true regression function. De Vito et al. (2010) note the difficulty in using Lespki's method to control the squared $L^2(P)$ error of an adaptive estimator. This difficulty arises because Lepski's method generally requires the norm we are interested in controlling to be known in order to perform the pairwise comparisons. However, $P$ is unknown in this situation.

De Vito et al. (2010) get around the problem that $P$ is unknown as follows. Lepski's method is used to control the known squared $L^2(P_n)$ error and squared RKHS error of two different adaptive estimators. The results of these procedures are combined to produce an adaptive estimator whose squared $L^2(P)$ error is bounded. The above alteration is also noted by Lu, Mathé, and Pereverzev (2018). Furthermore, the authors show that it is possible to greatly reduce the number of pairwise comparisons which must be performed to produce an adaptive estimator. This is done by only comparing each estimator to the estimator which is next less smooth.

The Goldenshluger–Lepski method extends Lepski's method in order to perform adaptation over multiple parameters. Goldenshluger and Lepski (2008, 2009) concentrate on function estimation in the presence of white noise. The first paper considers the problem of pointwise estimation, while the second paper examines estimation in the $L^p$ norm for $p \in [1, \infty]$. Goldenshluger and Lepski (2011) produce adaptive bandwidth estimators for kernel density estimation and Goldenshluger and Lepski (2013) consider general methodology for selecting a linear estimator from a collection.

An example of using training and validation to perform adaptation over a Gaussian kernel parameter for a support vector machine is examined by Eberts and Steinwart (2013). The procedure produces an adaptive estimator of a bounded regression function from a range of Sobolev spaces. This estimator is analysed using union bounding, as opposed to the chaining techniques used to analyse the Goldenshluger–Lepski method in this chapter.

## 4.2 Contribution

In this chapter, we use the Goldenshluger–Lepski method to produce an adaptive estimator from a collection of clipped versions of least-squares estimators which are constrained to lie in a ball of predefined radius in a fixed RKHS $H$, which is separable with a bounded and measurable kernel $k$. The estimator, defined by (4.6.1) on page 126, adapts over the radius of the ball. As far as we are aware, the Goldenshluger–Lepski method has not previously been applied in the context of RKHS regression. Under the assumption that the regression function comes from an interpolation space between $L^\infty$ and $H$, we prove a bound on a fixed quantile of the squared $L^2(P)$ error of this adaptive estimator of order $n^{-\beta/(1+\beta)}$ (Theorem 4.6.5 on page 129). Here, $P$ is the covariate distribution, $n$ is the number of data points and $\beta$ parametrises the interpolation space between $L^\infty$ and $H$. The order $n^{-\beta/(1+\beta)}$ matches the order of the smallest bounds obtained in Chapter 3 for the squared $L^2(P)$ error of the non-adaptive estimators. It is the optimal power of $n$, in the sense discussed in Chapter 3, if we make the closely-related weaker assumption that the regression function is an element of the interpolation space between $L^2(P)$ and the RKHS parametrised by $\beta$.

We then extend this result to the case in which we perform adaptation over a collection of RKHSs. In particular, we provide guarantees when the RKHSs in the collection have Gaussian kernels. We again use the Goldenshluger–Lepski method to produce an adaptive estimator, defined by (4.8.1), however this estimator adapts over both the RKHS and the radius of the ball. Under the assumption that the regression function comes from an interpolation space between $L^\infty$ and and some RKHS $H$ from the collection, we obtain a bound on a fixed quantile of the squared $L^2(P)$ error of the same order $n^{-\beta/(1+\beta)}$ (Theorem 4.8.7 on page 138).

## 4.3 RKHSs and Their Interpolation Spaces

An RKHS $H$ on $S$ is a Hilbert space of real-valued functions on $S$ such that, for all $x \in S$, there is some $k_x \in H$ such that $h(x) = \langle h, k_x \rangle_H$ for all $h \in H$. The function

$k(x_1, x_2) = \langle k_{x_1}, k_{x_2} \rangle_H$ for $x_1, x_2 \in S$ is known as the kernel and is symmetric and positive-definite.

We now define interpolation spaces between a Banach space $(Z, \|\cdot\|_Z)$ and a dense subspace $(V, \|\cdot\|_V)$ (see Bergh and Löfström, 1976). The $K$-functional of $(Z, V)$ is

$$K(z,t) = \inf_{v \in V} (\|z - v\|_Z + t\|v\|_V)$$

for $z \in Z$ and $t > 0$. We define

$$\|z\|_{\beta,q} = \left( \int_0^\infty (t^{-\beta} K(z,t))^q t^{-1} dt \right)^{1/q} \quad \text{and } \|z\|_{\beta,\infty} = \sup_{t>0} (t^{-\beta} K(z,t))$$

for $z \in Z$, $\beta \in (0,1)$ and $1 \leq q < \infty$. We then define the interpolation space $[Z, V]_{\beta,q}$ to be the set of $z \in Z$ such that $\|z\|_{\beta,q} < \infty$. The size of $[Z, V]_{\beta,q}$ decreases as $\beta$ increases. Recall Lemma 3.1.1, which is essentially Theorem 3.1 of Smale and Zhou (2003).

**Lemma 4.3.1** *Let $(Z, \|\cdot\|_Z)$ be a Banach space, $(V, \|\cdot\|_V)$ be a dense subspace of $Z$ and $z \in [Z, V]_{\beta,\infty}$. We have*

$$\inf\{\|v - z\|_Z : v \in V, \|v\|_V \leq r\} \leq \frac{\|z\|_{\beta,\infty}^{1/(1-\beta)}}{r^{\beta/(1-\beta)}}.$$

From the above, when $H$ is dense in $L^\infty$, we can define the interpolation spaces $[L^\infty, H]_{\beta,q}$, where $L^\infty$ is the space of bounded measurable functions on $(S, \mathcal{S})$. We set $q = \infty$ and work with the largest space of functions for a fixed $\beta \in (0,1)$. We are then able to apply the approximation result in Lemma 4.3.1.

## 4.4 Problem Definition

We give a formal definition of the RKHS regression problem. For a topological space $T$, let $\mathcal{B}(T)$ be its Borel $\sigma$-algebra. Let $(S, \mathcal{S})$ be a measurable space and $(X_i, Y_i)$ for $1 \leq i \leq n$ be i.i.d. $(S \times \mathbb{R}, \mathcal{S} \otimes \mathcal{B}(\mathbb{R}))$-valued random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume $X_i \sim P$ and $\mathbb{E}(Y_i^2) < \infty$, where $\mathbb{E}$ denotes integration with respect to $\mathbb{P}$. We have $\mathbb{E}(Y_i | X_i) = g(X_i)$ almost surely for some function $g$ which is measurable on $(S, \mathcal{S})$ (Section A3.2 of Williams, 1991). Since $\mathbb{E}(Y_i^2) < \infty$, it follows that $g \in L^2(P)$ by Jensen's inequality. We assume throughout that

$$(g1) \qquad \|g\|_\infty \leq C \text{ for } C > 0.$$

We also need to make an assumption on the behaviour of the errors of the response variables $Y_i$ for $1 \leq i \leq n$. Let $U$ and $V$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say $U$ is $\sigma^2$-subgaussian if

$$\mathbb{E}(\exp(tU)) \leq \exp(\sigma^2 t^2 / 2)$$

for all $t \in \mathbb{R}$. We say $U$ is $\sigma^2$-subgaussian given $V$ if

$$\mathbb{E}(\exp(tU) | V) \leq \exp(\sigma^2 t^2 / 2)$$

almost surely for all $t \in \mathbb{R}$. We assume

$$(Y) \qquad Y_i - g(X_i) \text{ is } \sigma^2\text{-subgaussian given } X_i \text{ for } 1 \leq i \leq n.$$

## 4.5 Regression for a Fixed RKHS

We continue by providing simultaneous bounds on our collection of non-adaptive estimators for a fixed RKHS. Our results in this section depend on how well the regression function $g$ can be approximated by elements of an RKHS $H$ with kernel $k$. We make the following assumptions.

$(H)$ The RKHS $H$ with kernel $k$ has the following properties:

- The RKHS $H$ is separable.
- The kernel $k$ is bounded.
- The kernel $k$ is a measurable function on $(S \times S, \mathcal{S} \otimes \mathcal{S})$.

We define

$$\|k\|_{\mathsf{diag}} = \sup_{x \in S} k(x, x) < \infty.$$

We use the notation $\|k\|_{\mathsf{diag}}$ in this chapter in place of $\|k\|_\infty^2$ from Chapter 3. We can guarantee that $H$ is separable by, for example, assuming that $k$ is continuous and $S$ is a separable topological space (Lemma 4.33 of Steinwart and Christmann, 2008). The fact that $H$ has a kernel $k$ which is measurable on $(S \times S, \mathcal{S} \otimes \mathcal{S})$ guarantees that all functions in $H$ are measurable on $(S, \mathcal{S})$ (Lemma 4.24 of Steinwart and Christmann, 2008).

Let $B_H$ be the closed unit ball of $H$ and $r > 0$. We define the estimator

$$\hat{h}_r = \operatorname*{arg\,min}_{f \in rB_H} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

of the regression function $g$. We make this definition unique by demanding that $\hat{h}_r \in \operatorname{sp}\{k_{X_i} : 1 \leq i \leq n\}$ (see Lemma 3.6.1). We also define $\hat{h}_0 = 0$. The following combines parts of Lemmas 3.6.1 and 3.15.2.

**Lemma 4.5.1** *Assume (H). We have that $\hat{h}_r$ is a $(H, \mathcal{B}(H))$-valued measurable function on $(\Omega \times [0,\infty), \mathcal{F} \otimes \mathcal{B}([0,\infty)))$, where $r$ varies in $[0,\infty)$. Furthermore, $\|\hat{h}_r - \hat{h}_s\|_H^2 \leq |r^2 - s^2|$ for $r, s \in [0,\infty)$.*

Since we assume $(g1)$, that $g$ is bounded in $[-C, C]$, we can make $\hat{h}_r$ closer to $g$ by constraining it to lie in the same interval. As in Chapter 3, we define the projection $V : \mathbb{R} \to [-C, C]$ by

$$V(t) = \begin{cases} -C & \text{if} \quad t < -C \\ t & \text{if} \quad |t| \leq C \\ C & \text{if} \quad t > C \end{cases}$$

for $t \in \mathbb{R}$.

We now prove a series of result which allow us to control $\hat{h}_r$ for $r \geq 0$ simultaneously, extending the results of Chapter 3 while using similar proof techniques. This is crucial in order to apply the Goldenshluger–Lepski method to these estimators. The results assign probabilities to events which occur for all $r \geq 0$ and all $h_r \in rB_H$. These events are measurable due to the separability of $[0,\infty)$ and $rB_H$, as well as the continuity in $r$ of the quantities in question, including $\hat{h}_r$ by Lemma 4.5.1. By Lemma 3.5.1, we have

$$\|\hat{h}_r - h_r\|_{L^2(P_n)}^2 \leq \frac{4}{n} \sum_{i=1}^{n} (Y_i - g(X_i))(\hat{h}_r(X_i) - h_r(X_i)) + 4\|h_r - g\|_{L^2(P_n)}^2$$

for all $r > 0$ and all $h_r \in rB_H$. We can get rid of $\hat{h}_r$ in the first term on the right-hand side by taking a supremum over $rB_H$. After applying the reproducing kernel property and the Cauchy–Schwarz inequality, we obtain a quadratic form of subgaussians which can be controlled using Lemma 3.16.2.

**Lemma 4.5.2** *Assume (Y) and (H). Let $t \geq 1$ and $A_{1,t} \in \mathcal{F}$ be the set on which*

$$\|\hat{h}_r - h_r\|^2_{L^2(P_n)} \leq \frac{20\|k\|^{1/2}_{\mathsf{diag}}\sigma r t^{1/2}}{n^{1/2}} + 4\|h_r - g\|^2_\infty$$

*simultaneously for all $r \geq 0$ and all $h_r \in rB_H$. We have $\mathbb{P}(A_{1,t}) \geq 1 - e^{-t}$.*

It is useful to be able to transfer a bound on the squared $L^2(P_n)$ error of an estimator, including the result above, to a bound on the squared $L^2(P)$ error of the estimator. By using Talagrand's inequality, we can obtain a high-probability bound on

$$\sup_{r>0} \sup_{f_1, f_2 \in rB_H} \frac{1}{r} \left| \|Vf_1 - Vf_2\|^2_{L^2(P_n)} - \|Vf_1 - Vf_2\|^2_{L^2(P)} \right|$$

by proving an expectation bound on the same quantity. By using symmetrisation (Lemma 2.3.1 of van der Vaart and Wellner, 1996) and the contraction principle for Rademacher processes (Theorem 3.2.1 of Giné and Nickl, 2016), we again obtain a quadratic form of subgaussians, which in this case are Rademacher random variables.

**Lemma 4.5.3** *Assume (H). Let $t \geq 1$ and $A_{2,t} \in \mathcal{F}$ be the set on which*

$$\sup_{f_1, f_2 \in rB_H} \left| \|Vf_1 - Vf_2\|^2_{L^2(P_n)} - \|Vf_1 - Vf_2\|^2_{L^2(P)} \right| \leq \frac{97\|k\|^{1/2}_{\mathsf{diag}} C r t^{1/2}}{n^{1/2}} + \frac{8\|k\|^{1/2}_{\mathsf{diag}} C r t}{3n}$$

*simultaneously for all $r \geq 0$. We have $\mathbb{P}(A_{2,t}) \geq 1 - e^{-t}$.*

To capture how well $g$ can be approximated by elements of $H$, we define

$$I_\infty(g, r) = \inf\left\{\|h_r - g\|^2_\infty : h_r \in rB_H\right\}$$

for $r \geq 0$. We use this measure of approximation as it is compatible with the use of the bound

$$\|h_r - g\|^2_{L^2(P_n)} \leq \|h_r - g\|^2_\infty$$

in the proof of Lemma 4.5.2. We show that $I_\infty(g,r)$ is continuous.

**Lemma 4.5.4** *Assume (H). Let $s \geq r \geq 0$. We have*

$$I_\infty(g,s) \leq I_\infty(g,r) \leq \left( I_\infty(g,s)^{1/2} + \|k\|_{\mathsf{diag}}^{1/2}(s-r) \right)^2 .$$

We obtain a bound on the squared $L^2(P)$ error of $V\hat{h}_r$ by combining Lemmas 4.5.2 and 4.5.3.

**Theorem 4.5.5** *Assume (g1), (Y) and (H). Let $t \geq 1$ and recall the definitions of $A_{1,t}$ and $A_{2,t}$ from Lemmas 4.5.2 and 4.5.3. On the set $A_{1,t} \cap A_{2,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{1,t} \cap A_{2,t}) \geq 1 - 2e^{-t}$, we have*

$$\|V\hat{h}_r - g\|_{L^2(P)}^2 \leq \frac{2\|k\|_{\mathsf{diag}}^{1/2}(97C + 20\sigma)rt^{1/2}}{n^{1/2}} + \frac{16\|k\|_{\mathsf{diag}}^{1/2}Crt}{3n} + 10I_\infty(g,r)$$

*simultaneously for all $r \geq 0$.*

## 4.6 The Goldenshluger–Lepski Method for a Fixed RKHS

We now produce bounds on our adaptive estimator for a fixed RKHS. The following result, which is a simple consequence of Lemma 4.5.2, can be used to define the majorant of the non-adaptive estimators. This motivates the definition of the adaptive estimator used in the Goldenshluger–Lepski method.

**Lemma 4.6.1** *Assume (Y) and (H). Let $t \geq 1$ and recall the definition of $A_{1,t}$ from*

*Lemma 4.5.2. On the set $A_{1,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{1,t}) \geq 1 - e^{-t}$, we have*

$$\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} \leq \frac{80\|k\|^{1/2}_{\mathsf{diag}}\sigma(r+s)t^{1/2}}{n^{1/2}} + 40 I_\infty(g, r)$$

*simultaneously for all $s \geq r \geq 0$.*

Let $R \subseteq [0, \infty)$ be closed and non-empty. The Goldenshluger–Lepski method defines an adaptive estimator using

$$\hat{r} = \operatorname*{arg\,min}_{r \in R} \left( \sup_{s \in R, s \geq r} \left( \|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r+s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau r}{n^{1/2}} \right) \tag{4.6.1}$$

for tuning parameters $\tau, \nu > 0$. The supremum of pairwise comparisons can be viewed as a proxy for the unknown bias, while the other term is an inflated variance term. Note that the supremum is at least the value at $r$, so

$$\sup_{s \in R, s \geq r} \left( \|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r+s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau r}{n^{1/2}} \geq \frac{2\nu\tau r}{n^{1/2}}. \tag{4.6.2}$$

The role of the tuning parameter $\nu$ is simply to control this bound. The parameter $\tau$ controls the probability with which our bound on the squared $L^2(P)$ error of $V\hat{h}_{\hat{r}}$ holds. We give a unique definition of $\hat{r}$.

**Lemma 4.6.2** *Let $\hat{r}$ be the infimum of all points attaining the minimum in* (4.6.1)*. Then $\hat{r}$ is well-defined.*

It may be that $\hat{r}$ is not a random variable on $(\Omega, \mathcal{F})$ in some cases, but we assume

$(\hat{r})$ $\quad$ $\hat{r}$ is a well-defined random variable on $(\Omega, \mathcal{F})$

throughout. Later, we assume that $R$ is finite, in which case $\hat{r}$ is certainly a random variable on $(\Omega, \mathcal{F})$. If $\hat{r}$ is a random variable on $(\Omega, \mathcal{F})$, then $\hat{h}_{\hat{r}}$ is a $(H, \mathcal{B}(H))$-valued measurable function on $(\Omega, \mathcal{F})$ by Lemma 4.5.1.

By Lemma 4.6.1, the supremum in the definition of $\hat{r}$ is at most $40I_\infty(g,r)$ for an appropriate value of $\tau$. The definition of $\hat{r}$ then gives us control over the squared $L^2(P_n)$ norm of $\hat{h}_{\hat{r}} - \hat{h}_r$ when $\hat{r} \le r$. When $\hat{r} \ge r$, we can control the squared $L^2(P_n)$ norm of $\hat{h}_{\hat{r}} - \hat{h}_r$ using Lemma 4.6.1. However, we must control a term of order $\hat{r}/n^{1/2}$ using (4.6.2) and the definition of $\hat{r}$. In both cases, this gives a bound on the squared $L^2(P_n)$ norm of $V\hat{h}_{\hat{r}} - V\hat{h}_r$. Extra terms appear when moving to a bound on the squared $L^2(P)$ norm of $V\hat{h}_{\hat{r}} - V\hat{h}_r$ using Lemma 4.5.3. However, these terms are very similar to the inflated variance term, and can be controlled in the same way. Applying

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \le 2\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P)} + 2\|V\hat{h}_r - g\|^2_{L^2(P)}$$

gives the following result.

**Theorem 4.6.3** *Assume (Y), (H) and ($\hat{r}$). Let $\tau \ge 80\|k\|^{1/2}_{\mathsf{diag}}\sigma$, $\nu > 0$ and*

$$t = \left(\frac{\tau}{80\|k\|^{1/2}_{\mathsf{diag}}\sigma}\right)^2 \ge 1.$$

*Recall the definitions of $A_{1,t}$ and $A_{2,t}$ from Lemmas 4.5.2 and 4.5.3. On the set $A_{1,t} \cap A_{2,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{1,t} \cap A_{2,t}) \ge 1 - 2e^{-t}$, we have*

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)}$$

*is at most*

$$\inf_{r\in R}\left(\max\left\{\frac{2\tau r}{n^{1/2}} + \left(\frac{1}{\nu} + \frac{97C}{80\sigma\nu} + \frac{C\tau}{2400\|k\|^{1/2}_{\mathsf{diag}}\sigma^2\nu n^{1/2}}\right)\left(40I_\infty(g,r) + \frac{2(1+\nu)\tau r}{n^{1/2}}\right),\right.\right.$$
$$\left.\left.\frac{4(2+\nu)\tau r}{n^{1/2}} + \frac{97C\tau r}{40\sigma n^{1/2}} + \frac{C\tau^2 r}{1200\|k\|^{1/2}_{\mathsf{diag}}\sigma^2 n}\right\} + 80I_\infty(g,r) + 2\|V\hat{h}_r - g\|^2_{L^2(P)}\right).$$

We now combine Theorems 4.5.5 and 4.6.3.

**Theorem 4.6.4** *Assume (g1), (Y), (H) and ($\hat{r}$). Let $\tau \geq 80\|k\|_{\mathsf{diag}}^{1/2}\sigma$, $\nu > 0$ and*

$$t = \left(\frac{\tau}{80\|k\|_{\mathsf{diag}}^{1/2}\sigma}\right)^2 \geq 1.$$

*Recall the definitions of $A_{1,t}$ and $A_{2,t}$ from Lemmas 4.5.2 and 4.5.3. On the set $A_{1,t} \cap A_{2,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{1,t} \cap A_{2,t}) \geq 1 - 2e^{-t}$, we have*

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 \leq \inf_{r \in R}\left((1 + D_1\tau n^{-1/2})(D_2\tau r n^{-1/2} + D_3 I_\infty(g, r))\right)$$

*for constants $D_1, D_2, D_3 > 0$ not depending on $\tau$, $r$ or $n$.*

We can obtain rates of convergence for our estimator $V\hat{h}_{\hat{r}}$ if we make an assumption about how well $g$ can be approximated by elements of $H$. Let us assume

($g2$) $\quad g \in [L^\infty, H]_{\beta,\infty}$ with norm at most $B$ for $\beta \in (0, 1)$ and $B > 0$.

The assumption ($g2$), together with Lemma 4.3.1, give

$$I_\infty(g, r) \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} \tag{4.6.3}$$

for $r > 0$. In order for us to apply Theorem 4.6.4 to this setting, we need to make an assumption on $R$. We assume either

($R1$) $\quad R = [0, \infty)$

or

($R2$) $R = \{bi : 0 \leq i \leq I - 1\} \cup \{an^{1/2}\}$ and $\rho = an^{1/2}$ for $a, b > 0$ and $I = \lceil an^{1/2}/b \rceil$.

The assumption ($R1$) is mainly of theoretical interest and would make it difficult to calculate $\hat{r}$ in practice. The estimator $\hat{r}$ can be computed under the assumption $(R2)$, since in this case $R$ is finite. We obtain a high-probability bound on a fixed quantile of the squared $L^2(P)$ error of $V\hat{h}_{\hat{r}}$ of order $t^{1/2}n^{-\beta/(1+\beta)}$ with probability at least $1-e^{-t}$ when $\tau$ is an appropriate multiple of $t^{1/2}$.

**Theorem 4.6.5** *Assume (g1), (g2), (Y) and (H). Let* $\tau \geq 80\|k\|_{\mathsf{diag}}^{1/2}\sigma$, $\nu > 0$ *and*

$$t = \left(\frac{\tau}{80\|k\|_{\mathsf{diag}}^{1/2}\sigma}\right)^2 \geq 1.$$

*Also assume (R1) and ($\hat{r}$), or (R2). Recall the definitions of $A_{1,t}$ and $A_{2,t}$ from Lemmas 4.5.2 and 4.5.3. On the set $A_{1,t} \cap A_{2,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{1,t} \cap A_{2,t}) \geq 1-2e^{-t}$, we have*

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 \leq D_1 \tau n^{-\beta/(1+\beta)} + D_2 \tau^2 n^{-(1+3\beta)/(2(1+\beta))}$$

*for constants $D_1, D_2 > 0$ not depending on $n$ or $\tau$.*

## 4.7 Regression for a Collection of RKHSs

In this section, we again provide simultaneous bounds on our collection of non-adaptive estimators. Our results still depend on how well the regression function $g$ can be approximated by elements of an RKHS. However, this RKHS now comes from a collection instead of being fixed. Let $\mathcal{K}$ be a set of kernels on $S \times S$. We make the following assumptions.

$(\mathcal{K}1)$ The covariate set $S$ and the set of kernels $\mathcal{K}$ have the following properties:

- The covariate set $S$ is a separable topological space.

- The set of kernels $(\mathcal{K}, \|\cdot\|_\infty)$ is separable.

- The kernel $k$ is bounded for all $k \in \mathcal{K}$.
- The kernel $k$ is continuous for all $k \in \mathcal{K}$.

Since $(\mathcal{K}, \|\cdot\|_\infty)$ is a separable set of kernels, we have that $\mathcal{K}$ has a countable dense subset $\mathcal{K}_0$. For all $\varepsilon > 0$ and all $k \in \mathcal{K}$, there exists $k_0 \in \mathcal{K}_0$ such that

$$\|k_0 - k\|_\infty = \sup_{x_1, x_2 \in S} |k_0(x_1, x_2) - k(x_1, x_2)| < \varepsilon.$$

Let $H_k$ be the RKHS with kernel $k$ for $k \in \mathcal{K}$. Since $k$ is continuous and $S$ is a separable topological space, we have that $H_k$ is separable by Lemma 4.33 of Steinwart and Christmann (2008). Hence, the assumption $(H)$ holds for $H_k$. We use the notation $\|\cdot\|_k$ and $\langle \cdot, \cdot \rangle_k$ for the norm and inner product of $H_k$.

Let $B_k$ be the closed unit ball of $H_k$ for $k \in \mathcal{K}$ and $r > 0$. We define the estimator

$$\hat{h}_{k,r} = \operatorname*{arg\,min}_{f \in rB_k} \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2$$

of the regression function $g$. We make this definition unique by demanding that $\hat{h}_{k,r} \in \operatorname{sp}\{k_{X_i} : 1 \leq i \leq n\}$ (see Lemma 3.6.1). We also define $\hat{h}_{k,0} = 0$. Since we assume $(g1)$, that $g$ is bounded in $[-C, C]$, we can make $\hat{h}_{k,r}$ closer to $g$ by clipping it to obtain $V\hat{h}_{k,r}$.

**Lemma 4.7.1** *Assume ($\mathcal{K}1$). We have that $\hat{h}_{k,r}$ is an $(L^\infty, \mathcal{B}(L^\infty))$-valued measurable function on $(\Omega \times \mathcal{K} \times [0,\infty), \mathcal{F} \otimes \mathcal{B}(\mathcal{K}) \otimes \mathcal{B}([0,\infty)))$, where $k$ varies in $\mathcal{K}$ and $r$ varies in $[0,\infty)$.*

Let

$$\mathcal{L} = \{k/\|k\|_{\mathsf{diag}} : k \in \mathcal{K}\} \cup \{0\}$$

and

$$D = \sup_{f_1, f_2 \in \mathcal{L}} \|f_1 - f_2\|_\infty \leq 2.$$

We include 0 in the definition of $\mathcal{L}$ so that, when analysing stochastic processes over $\mathcal{L}$ using chaining, we can start all chains at 0. Note that $(\mathcal{L}, \|\cdot\|_\infty)$ is separable since $\mathcal{L} \setminus \{0\}$ is the image of a continuous function on $(\mathcal{K}, \|\cdot\|_\infty)$, which is itself separable. Let $N(a, M, d)$ be the minimum size of an $a > 0$ cover of a metric space $(M, d)$, and let

$$J = \left( 162 \int_0^{D/2} \log(2N(a, \mathcal{L}, \|\cdot\|_\infty)) da + 1 \right)^{1/2}.$$

The next result is proved using the same method as Lemma 4.5.2. However, instead of one quadratic form of subgaussians, we obtain a supremum over $\mathcal{K}$ of quadratic forms of subgaussians. This can be controlled by chaining using Lemma 4.12.2.

**Lemma 4.7.2** *Assume (Y) and ($\mathcal{K}$1). Let $t \geq 1$. There exists a set $A_{3,t} \in \mathcal{F}$ with $\mathbb{P}(A_{3,t}) \geq 1 - e^{-t}$ on which*

$$\|\hat{h}_{k,r} - h_{k,r}\|_{L^2(P_n)}^2 \leq \frac{21 J \|k\|_{\mathsf{diag}}^{1/2} \sigma r t^{1/2}}{n^{1/2}} + 4\|h_{k,r} - g\|_\infty^2$$

*simultaneously for all $k \in \mathcal{K}$, all $r \geq 0$ and all $h_{k,r} \in rB_k$.*

It is again useful to be able to transfer a bound on the squared $L^2(P_n)$ error of an estimator to a bound on the squared $L^2(P)$ error of the estimator. The result below is proved using the same method as Lemma 4.5.3, although we again obtain a supremum of quadratic forms of subgaussians which are controlled using chaining. The event in the result is measurable by Lemma 4.12.3.

**Lemma 4.7.3** *Assume ($\mathcal{K}$1). Let $t \geq 1$ and $A_{4,t} \in \mathcal{F}$ be the set on which*

$$\sup_{f_1, f_2 \in rB_k} \left| \|Vf_1 - Vf_2\|_{L^2(P_n)}^2 - \|Vf_1 - Vf_2\|_{L^2(P)}^2 \right| \leq \frac{151 J \|k\|_{\mathsf{diag}}^{1/2} C r t^{1/2}}{n^{1/2}} + \frac{8\|k\|_{\mathsf{diag}}^{1/2} C r t}{3n}$$

*simultaneously for all $k \in \mathcal{K}$ and all $r \geq 0$. We have $\mathbb{P}(A_{4,t}) \geq 1 - e^{-t}$.*

To capture how well $g$ can be approximated by elements of $H_k$, we define

$$I_\infty(g, k, r) = \inf \left\{ \|h_{k,r} - g\|_\infty^2 : h_{k,r} \in rB_k \right\}$$

for $k \in \mathcal{K}$ and $r \geq 0$. We obtain a bound on the squared $L^2(P)$ error of $V\hat{h}_{k,r}$ by combining Lemmas 4.7.2 and 4.7.3.

**Theorem 4.7.4** *Assume (g1), (Y) and ($\mathcal{K}$1). Let $t \geq 1$ and recall the definitions of $A_{3,t}$ and $A_{4,t}$ from Lemmas 4.7.2 and 4.7.3. On the set $A_{3,t} \cap A_{4,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{3,t} \cap A_{4,t}) \geq 1 - 2e^{-t}$, we have*

$$\|V\hat{h}_{k,r} - g\|_{L^2(P)}^2 \leq \frac{2J\|k\|_{\mathsf{diag}}^{1/2}(151C + 21\sigma)rt^{1/2}}{n^{1/2}} + \frac{16\|k\|_{\mathsf{diag}}^{1/2}Crt}{3n} + 10I_\infty(g, k, r)$$

*simultaneously for all $k \in \mathcal{K}$ and all $r \geq 0$.*

## 4.8 The Goldenshluger–Lepski Method for a Collection of RKHSs with Gaussian Kernels

We now apply the Goldenshluger–Lepski method again in the context of RKHS regression. However, we now produce an estimator which adapts over a collection of RKHSs with Gaussian kernels. We make the following assumptions on $S$ and $\mathcal{K}$.

($\mathcal{K}$2) The covariate set $S$ and the set of kernels $\mathcal{K}$ have the following properties:

- The covariate set $S \subseteq \mathbb{R}^d$ for $d \geq 1$.

- The set of kernels

  $$\mathcal{K} = \left\{k_\gamma(x_1, x_2) = \gamma^{-d} \exp\left(-\|x_1 - x_2\|_2^2/\gamma^2\right) : \gamma \in \Gamma \text{ and } x_1, x_2 \in S\right\}$$

  for $\Gamma \subseteq [u, v]$ non-empty for $v \geq u > 0$.

Recalling the definitions from the previous section, we have

$$\mathcal{L} = \left\{f_\gamma(x_1, x_2) = \exp\left(-\|x_1 - x_2\|_2^2/\gamma^2\right) : \gamma \in \Gamma \text{ and } x_1, x_2 \in S\right\} \cup \{0\}.$$

The assumption ($\mathcal{K}2$) implies the assumption ($\mathcal{K}1$). This is because Lemma 4.14.1 shows that $(\mathcal{L}, \|\cdot\|_\infty)$, and hence $(\mathcal{K}, \|\cdot\|_\infty)$, is separable. We change notation slightly. Let $H_\gamma$ be the RKHS with kernel $k_\gamma$ for $\gamma \in \Gamma$, let $\|\cdot\|_\gamma$ and $\langle\cdot,\cdot\rangle_\gamma$ be the norm and inner product of $H_\gamma$, and let $B_\gamma$ be the closed unit ball of $H_\gamma$. Furthermore, we write $\hat{h}_{\gamma,r}$ in place of $\hat{h}_{k_\gamma,r}$ and $I_\infty(g, \gamma, r)$ in place of $I_\infty(g, k_\gamma, r)$.

The scaling of the kernels is selected so that the following lemma holds. The result is immediate from Proposition 4.46 of Steinwart and Christmann (2008) and the way that the norm of an RKHS scales with its kernel (Theorem 4.21 of Steinwart and Christmann, 2008).

**Lemma 4.8.1** *Assume ($\mathcal{K}2$). Let $\gamma, \eta \in \Gamma$ with $\gamma \geq \eta$. We have $B_\gamma \subseteq B_\eta$.*

By Lemma 4.14.1, the function $F : \Gamma \to \mathcal{L} \setminus \{0\}$ by $F(\gamma) = f_\gamma$ is continuous. Hence, the function $G : \Gamma \to \mathcal{K}$ by $G(\gamma) = k_\gamma$ is continuous. The next result then follows from Lemma 4.7.1.

**Lemma 4.8.2** *Assume ($\mathcal{K}2$). We have that $\hat{h}_{\gamma,r}$ is an $(L^\infty, \mathcal{B}(L^\infty))$-valued measurable function on $(\Omega \times \Gamma \times [0, \infty), \mathcal{F} \otimes \mathcal{B}(\Gamma) \otimes \mathcal{B}([0, \infty)))$, where $\gamma$ varies in $\Gamma$ and $r$ varies in $[0, \infty)$.*

Recall the definition of $J$ from the previous section. Lemma 4.14.2 provides us with a bound on $J$.

**Lemma 4.8.3** *Assume ($\mathcal{K}2$). We have*

$$J \leq \left(81(\log(8\log(v/u)+4)+2)+1\right)^{1/2}.$$

The following result can be used to define the majorant of the non-adaptive estimators and is a simple consequence of Lemma 4.7.2. This motivates the definition of the adaptive estimator used in the Goldenshluger–Lepski method.

**Lemma 4.8.4** *Assume ($Y$) and ($\mathcal{K}2$). Let $t \geq 1$ and recall the definition of $A_{3,t}$ from Lemma 4.7.2. On the set $A_{3,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{3,t}) \geq 1 - e^{-t}$, we have*

$$\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L^2(P_n)} \leq \frac{84J\sigma(\gamma^{-d/2}r + \eta^{-d/2}s)t^{1/2}}{n^{1/2}} + 40I_\infty(g,\gamma,r)$$

*simultaneously for all $\gamma, \eta \in \Gamma$ such that $\eta \leq \gamma$ and all $s \geq r \geq 0$.*

Let $R \subseteq [0,\infty)$ be non-empty. The Goldenshluger–Lepski method creates an adaptive estimator by defining $(\hat{\gamma}, \hat{r})$ to be the minimiser of

$$\sup_{\eta\in\Gamma,\eta\leq\gamma}\ \sup_{s\in R,s\geq r}\left(\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L_2(P_n)} - \frac{\tau(\gamma^{-d/2}r + \eta^{-d/2}s)}{n^{1/2}}\right) + \frac{2(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}} \tag{4.8.1}$$

over $(\gamma, r) \in \Gamma \times R$ for tuning parameters $\tau, \nu > 0$. Again, the supremum of pairwise comparisons can be viewed as a proxy for the unknown bias, while the other term is an inflated variance term. Note that the supremum is at least the value at $(\gamma, r)$, which means that (4.8.1) is at least

$$\frac{2\nu\tau\gamma^{-d/2}r}{n^{1/2}}. \tag{4.8.2}$$

Again, the role of the tuning parameter $\nu$ is simply to control this bound. The parameter $\tau$ controls the probability with which our bound on the squared $L^2(P)$ error of $V\hat{h}_{\hat{\gamma},\hat{r}}$ holds. It may be that $\hat{\gamma}$ is not a well-defined random variable on $(\Omega, \mathcal{F})$ in some cases, but we assume

$$(\hat{\gamma}) \qquad\qquad \hat{\gamma} \text{ is a well-defined random variable on } (\Omega, \mathcal{F})$$

throughout. Later, we assume that $R$ and $\Gamma$ are finite, in which case $\hat{\gamma}$ and $\hat{r}$ are certainly well-defined random variables on $(\Omega, \mathcal{F})$. If $\hat{\gamma}$ and $\hat{r}$ are well-defined random variables on $(\Omega, \mathcal{F})$, then $\hat{h}_{\hat{\gamma},\hat{r}}$ is an $(L^\infty, \mathcal{B}(L^\infty))$-valued measurable function on $(\Omega, \mathcal{F})$ by Lemma 4.8.2.

By Lemma 4.8.4, the supremum in the definition of $(\hat{\gamma}, \hat{r})$ is at most $40I_\infty(g, \gamma, r)$ for an appropriate value of $\tau$. The definition of $(\hat{\gamma}, \hat{r})$ then gives us control over the squared $L^2(P_n)$ norm of $\hat{h}_{\hat{\gamma},\hat{r}} - \hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}$. We can control the squared $L^2(P_n)$ norm of $\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r} - h_{\gamma,r}$ using Lemma 4.8.4. In both cases, we use the boundedness of $\Gamma$ when controlling the squared $L^2(P_n)$ norm before clipping the estimators using $V$. Extra terms appear when moving from bounds on the squared $L^2(P_n)$ norm to bounds on the squared $L^2(P)$ norm using Lemma 4.7.3. We must then control terms of order $\hat{\gamma}^{-d/2}\hat{r}/n^{1/2}$ using (4.8.2) and the definition of $(\hat{\gamma}, \hat{r})$. Combining the bounds gives a bound on the squared $L^2(P)$ norm of $Vh_{\hat{\gamma},\hat{r}} - Vh_{\gamma,r}$. Applying

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \le 2\|Vh_{\hat{\gamma},\hat{r}} - Vh_{\gamma,r}\|^2_{L^2(P)} + 2\|Vh_{\gamma,r} - g\|^2_{L^2(P)}$$

gives the following result. Comparisons between $(\hat{r}, \hat{\gamma})$, $(r, \gamma)$ and $(\hat{r} \vee r, \hat{\gamma} \wedge \gamma)$ are demonstrated in Figure 4.1 for two different values of $(r, \gamma)$.

**Theorem 4.8.5** *Assume (Y) and ($\mathcal{K}2$). Let $\tau \ge 84J\sigma$, $\nu > 0$ and*

$$t = \left(\frac{\tau}{84J\sigma}\right)^2 \ge 1.$$

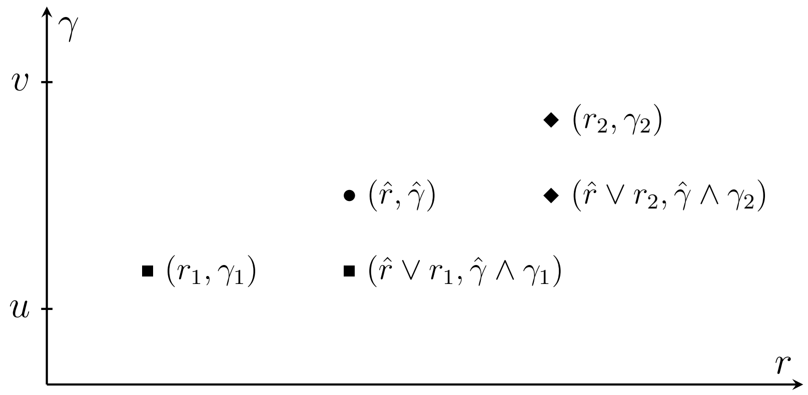
$\gamma$
$v$
$u$
$r$
$(r_2, \gamma_2)$
$(\hat{r}, \hat{\gamma})$
$(\hat{r} \vee r_2, \hat{\gamma} \wedge \gamma_2)$
$(r_1, \gamma_1)$
$(\hat{r} \vee r_1, \hat{\gamma} \wedge \gamma_1)$

Figure 4.1: A demonstration of the parameter comparisons made in the proof of Theorem 4.8.5

*Recall the definitions of* $A_{3,t}$ *and* $A_{4,t}$ *from Lemmas 4.7.2 and 4.7.3. On the set* $A_{3,t} \cap A_{4,t} \in \mathcal{F}$*, for which* $\mathbb{P}(A_{3,t} \cap A_{4,t}) \geq 1 - 2e^{-t}$*, we have*

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L^2(P)}$$

*is at most*

$$\inf_{\gamma \in \Gamma} \inf_{r \in R} \Bigg( 320 I_\infty(g,\gamma,r) + \frac{4v^{d/2}(5+2\nu)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{302Cv^{d/2}\tau\gamma^{-d/2}r}{21u^{d/2}\sigma n^{1/2}} + \frac{4Cv^{d/2}\tau^2\gamma^{-d/2}r}{1323J^2u^{d/2}\sigma^2 n}$$
$$+ \left(\frac{12v^{d/2}}{u^{d/2}\nu} + \frac{302Cv^{d/2}}{21u^{d/2}\sigma\nu} + \frac{4Cv^{d/2}\tau}{1323J^2u^{d/2}\sigma^2\nu n^{1/2}}\right)\left(20I_\infty(g,\gamma,r) + \frac{(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}}\right)$$
$$+ 2\|V\hat{h}_{\gamma,r} - g\|^2_{L_2(P)}\Bigg) .$$

We now combine Theorems 4.7.4 and 4.8.5.

**Theorem 4.8.6** *Assume (g1), (Y) and ($\mathcal{K}$2). Let* $\tau \geq 84J\sigma$*,* $\nu > 0$ *and*

$$t = \left(\frac{\tau}{84J\sigma}\right)^2 \geq 1.$$

*Recall the definitions of* $A_{3,t}$ *and* $A_{4,t}$ *from Lemmas 4.7.2 and 4.7.3. On the set*

*$A_{3,t} \cap A_{4,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{3,t} \cap A_{4,t}) \geq 1 - 2e^{-t}$, we have*

$$\|V\hat{h}_{\gamma,\hat{r}} - g\|^2_{L^2(P)} \leq \inf_{\gamma\in\Gamma}\inf_{r\in R}\big((1 + D_1\tau n^{-1/2})(D_2\tau\gamma^{-d/2}rn^{-1/2} + D_3 I_\infty(g,\gamma,r))\big)$$

*for constants $D_1, D_2, D_3 > 0$ not depending on $\tau$, $\gamma$, $r$ or $n$.*

We can obtain rates of convergence for our estimator $V\hat{h}_{\hat{\gamma},\hat{r}}$ if we make an assumption about how well $g$ can be approximated by elements of $H_\alpha$ for $\alpha \in [u,v]$. Let us assume

$(g3)$ $g \in [L^\infty, H_\alpha]_{\beta,\infty}$ with norm at most $B$ for $\alpha \in [u,v]$, $\beta \in (0,1)$ and $B > 0$.

The assumption $(g3)$, together with Lemma 4.3.1, give

$$I_\infty(g,\alpha,r) \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} \tag{4.8.3}$$

for $r > 0$. In order for us to apply Theorem 4.8.6 to this setting, we need to make assumptions on $\Gamma$ and $R$. We assume either $(R1)$ and

$(\Gamma 1)$ $\Gamma = [u,v],$

or $(R2)$ and

$(\Gamma 2)$ $\Gamma = \{uc^i : 0 \leq i \leq L-1\} \cup \{v\}$ for $c > 1$ and $L = \lceil \log(v/u)/\log(c) \rceil$.

The assumptions $(R1)$ and $(\Gamma1)$ are mainly of theoretical interest and would make it difficult to calculate $(\hat{\gamma},\hat{r})$ in practice. The estimator $(\hat{\gamma},\hat{r})$ can be computed under the assumptions $(R2)$ and $(\Gamma2)$, since in this case $R$ and $\Gamma$ are finite. We obtain a high-probability bound on a fixed quantile of the squared $L^2(P)$ error of $V\hat{h}_{\hat{r},\hat{\gamma}}$ of order $t^{1/2}n^{-\beta/(1+\beta)}$ with probability at least $1-e^{-t}$ when $\tau$ is an appropriate multiple of $t^{1/2}$.

**Theorem 4.8.7** *Assume (g1), (g3), (Y) and ($\mathcal{K}$2). Let $\tau \geq 84J\sigma$, $\nu > 0$ and*

$$t = \left(\frac{\tau}{84J\sigma}\right)^2 \geq 1.$$

*Also assume (R1), ($\Gamma$1), ($\hat{r}$) and ($\hat{\gamma}$), or (R2) and ($\Gamma$2). Recall the definitions of $A_{3,t}$ and $A_{4,t}$ from Lemmas 4.7.2 and 4.7.3. On the set $A_{3,t} \cap A_{4,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{3,t} \cap A_{4,t}) \geq 1 - 2e^{-t}$, we have*

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L^2(P)} \leq D_1 \tau n^{-\beta/(1+\beta)} + D_2 \tau^2 n^{-(1+3\beta)/(2(1+\beta))}$$

*for constants $D_1, D_2 > 0$ not depending on $n$ or $\tau$.*

## 4.9 Discussion

In this chapter, we show how the Goldenshluger–Lepski method can be applied when performing regression over an RKHS $H$, which is separable with a bounded and measurable kernel $k$, or a collection of such RKHSs. We produce an adaptive estimator from a collection of clipped versions of least-squares estimators which are constrained to lie in a ball of predefined radius in $H$. Since the $L^2(P)$ norm is unknown, we use the $L^2(P_n)$ norm when calculating the pairwise comparisons for the proxy for the unknown bias of this collection of non-adaptive estimators. When $H$ is fixed, our estimator need only adapt to the radius of the ball in $H$. However, when $H$ comes from a collection of RKHSs with Gaussian kernels, the estimator must also adapt to the width parameter of the kernel. As far as we are aware, this is the first time that the Goldenshluger–Lepski method has been applied in the context of RKHS regression. In order to apply the Goldenshluger–Lepski method in this context, we must provide a majorant by controlling all of the non-adaptive estimators simultaneously, extending the results of Chapter 3.

By assuming that the regression function lies in an interpolation space between $L^\infty$ and $H$ parametrised by $\beta$, we obtain a bound on a fixed quantile of the squared $L^2(P)$ error of our adaptive estimator of order $n^{-\beta/(1+\beta)}$. This is true for both the case in which $H$ is fixed and the case in which $H$ comes from a collection of RKHSs with Gaussian kernels. The order $n^{-\beta/(1+\beta)}$ for the squared $L^2(P)$ error of the adaptive estimators matches the order of the smallest bounds obtained in Chapter 3 for the squared $L^2(P)$ error of the non-adaptive estimators. In the sense discussed in Chapter 3, this order is the optimal power of $n$ if we make the slightly weaker assumption that the regression function is an element of the interpolation space between $L^2(P)$ and $H$ parametrised by $\beta$.

For the case in which $H$ comes from a collection of RKHSs with Gaussian kernels, our current results rely on the boundedness of the set $\Gamma$ of width parameters of the kernels. This is somewhat limiting as allowing the width parameter to tend to 0 as $n$ tends to infinity would allow us to estimate a greater collection of functions. We hope that in the future the analysis in the proof of Theorem 4.8.5 can be extended to allow for such flexibility.

The results in this chapter warrant the investigation of whether it is possible to extend the use of the Goldenshluger–Lepski method from the case in which $H$ comes from a collection of RKHSs with Gaussian kernels to other cases. The analysis in this chapter relies on the fact that the closed unit ball of the RKHS generated by a Gaussian kernel increases as the width of the kernel decreases. It may be possible to apply a similar analysis to other situations in which $H$ belongs to a collection of RKHSs which also exhibit this nestedness property. If the RKHSs did not exhibit this property, then a new form of analysis would be necessary to apply the Goldenshluger–Lepski method. In particular, we would need a new criterion for deciding on the smoothness of the non-adaptive estimators when performing the pairwise comparisons.

## 4.10 Proof of the Regression Results for a Fixed RKHS

We bound the distance between $\hat{h}_r$ and $h_r$ in the $L^2(P_n)$ norm for $r \geq 0$ and $h_r \in rB_H$ to prove Lemma 4.5.2.

**Proof of Lemma 4.5.2** The result is trivial for $r = 0$. By Lemma 3.5.1, we have

$$\|\hat{h}_r - h_r\|_{L^2(P_n)}^2 \leq \frac{4}{n}\sum_{i=1}^{n}(Y_i - g(X_i))(\hat{h}_r(X_i) - h_r(X_i)) + 4\|h_r - g\|_{L^2(P_n)}^2$$

for all $r > 0$ and all $h_r \in rB_H$. We now bound the right-hand side. We have

$$\|h_r - g\|_{L^2(P_n)}^2 \leq \|h_r - g\|_\infty^2.$$

Furthermore,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))(\hat{h}_r(X_i) - h_r(X_i)) \\
\leq\ & \sup_{f \in 2rB_H}\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))f(X_i)\right| \\
=\ & \sup_{f \in 2rB_H}\left|\left\langle \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))k_{X_i}, f\right\rangle_H\right| \\
=\ & 2r\left\|\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))k_{X_i}\right\|_H \\
=\ & 2r\left(\frac{1}{n^2}\sum_{i,j=1}^{n}(Y_i - g(X_i))(Y_j - g(X_j))k(X_i, X_j)\right)^{1/2}
\end{aligned}$$

by the reproducing kernel property and the Cauchy–Schwarz inequality. Let $K$ be the $n \times n$ matrix with $K_{i,j} = k(X_i, X_j)$ and let $\varepsilon$ be the vector of the $Y_i - g(X_i)$. Then

$$\frac{1}{n^2} \sum_{i,j=1}^{n} (Y_i - g(X_i))(Y_j - g(X_j))k(X_i, X_j) = \varepsilon^{\mathsf{T}}(n^{-2}K)\varepsilon.$$

Furthermore, since $k$ is a measurable function on $(S \times S, \mathcal{S} \otimes \mathcal{S})$, we have that $n^{-2}K$ is an $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$ and non-negative-definite. Let $a_i$ for $1 \leq i \leq n$ be the eigenvalues of $n^{-2}K$. Then

$$\max_{1 \leq i \leq n} a_i \leq \operatorname{tr}(n^{-2}K) \leq n^{-1} \|k\|_{\mathsf{diag}}$$

and

$$\operatorname{tr}((n^{-2}K)^2) = \|a\|_2^2 \leq \|a\|_1^2 \leq n^{-2} \|k\|_{\mathsf{diag}}^2.$$

Therefore, by Lemma 3.16.2, we have

$$\varepsilon^{\mathsf{T}}(n^{-2}K)\varepsilon \leq \|k\|_{\mathsf{diag}} \sigma^2 n^{-1}(1 + 2t + 2(t^2 + t)^{1/2})$$

and

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))(\hat{h}_r(X_i) - h_r(X_i)) \leq \frac{5\|k\|_{\mathsf{diag}}^{1/2} \sigma r t^{1/2}}{n^{1/2}}$$

with probability at least $1 - e^{-t}$. The result follows. ■

Recall Lemma 3.13.2, which is useful for proving Lemma 4.5.3.

**Lemma 4.10.1** *Let $D > 0$ and $A \subseteq L^\infty$ be separable with $\|f\|_\infty \leq D$ for all $f \in A$. Let*

$$Z = \sup_{f \in A} \left| \|f\|_{L^2(P_n)}^2 - \|f\|_{L^2(P)}^2 \right|.$$

*Then, for $t > 0$, we have*

$$Z \leq \mathbb{E}(Z) + \left( \frac{2D^4 t}{n} + \frac{4D^2\,\mathbb{E}(Z) t}{n} \right)^{1/2} + \frac{2D^2 t}{3n}$$

*with probability at least $1 - e^{-t}$.*

We bound the supremum of the difference in the $L^2(P_n)$ norm and the $L^2(P)$ norm over $rB_H$ for $r \geq 0$ to prove Lemma 4.5.3.

**Proof of Lemma 4.5.3** The result is trivial for $r = 0$. Let

$$Z = \sup_{r>0} \sup_{f_1, f_2 \in rB_H} \frac{1}{r} \left| \|Vf_1 - Vf_2\|^2_{L^2(P_n)} - \|Vf_1 - Vf_2\|^2_{L^2(P)} \right|.$$

Furthermore, let the $\varepsilon_i$ for $1 \leq i \leq n$ be i.i.d. Rademacher random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, independent of the $X_i$. Lemma 2.3.1 of van der Vaart and Wellner (1996) shows

$$\mathbb{E}(Z) \leq 2\,\mathbb{E} \left( \sup_{r>0} \sup_{f_1, f_2 \in rB_H} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (r^{-1/2} V f_1(X_i) - r^{-1/2} V f_2(X_i))^2 \right| \right)$$

by symmetrisation. Since

$$|Vf_1(X_i) - Vf_2(X_i)| \leq 2C$$

for all $r > 0$ and all $f_1, f_2 \in rB_H$, we find

$$\frac{(r^{-1/2} V f_1(X_i) - r^{-1/2} V f_2(X_i))^2}{4C}$$

is a contraction vanishing at 0 as a function of $r^{-1} V f_1(X_i) - r^{-1} V f_2(X_i)$ for all

$1 \leq i \leq n$. By Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}\left(\sup_{r>0} \sup_{f_1, f_2 \in rB_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \frac{(r^{-1/2} V f_1(X_i) - r^{-1/2} V f_2(X_i))^2}{4C} \right| \middle| X \right)$$

is at most

$$2\,\mathbb{E}\left(\sup_{r>0} \sup_{f_1, f_2 \in rB_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (r^{-1} V f_1(X_i) - r^{-1} V f_2(X_i)) \right| \middle| X \right)$$

almost surely. Therefore,

$$\begin{aligned}
\mathbb{E}(Z) &\leq 16C\,\mathbb{E}\left(\sup_{r>0} \sup_{f_1, f_2 \in rB_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (r^{-1} V f_1(X_i) - r^{-1} V f_2(X_i)) \right| \right) \\
&\leq 32C\,\mathbb{E}\left(\sup_{r>0} \sup_{f \in rB_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i r^{-1} V f(X_i) \right| \right)
\end{aligned}$$

by the triangle inequality. Again, by Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}(Z) \leq 64C\,\mathbb{E}\left(\sup_{r>0} \sup_{f \in rB_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i r^{-1} f(X_i) \right| \right)$$

since $V$ is a contraction vanishing at 0. We have

$$\begin{aligned}
\sup_{r>0} \sup_{f \in rB_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i r^{-1} f(X_i) \right| &= \sup_{r>0} \sup_{f \in rB_H} \left| \left\langle \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i k_{X_i}, r^{-1} f \right\rangle_H \right| \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i k_{X_i} \right\|_H \\
&= \left( \frac{1}{n^2} \sum_{i,j=1}^{n} \varepsilon_i \varepsilon_j k(X_i, X_j) \right)^{1/2}
\end{aligned}$$

by the reproducing kernel property and the Cauchy–Schwarz inequality. By Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}\left(\sup_{r>0}\sup_{f\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i r^{-1} f(X_i)\right|\,\middle|\, X\right) &\le \left(\frac{1}{n^2}\sum_{i,j=1}^n \operatorname{cov}(\varepsilon_i,\varepsilon_j|X)k(X_i,X_j)\right)^{1/2}\\ &= \left(\frac{1}{n^2}\sum_{i=1}^n k(X_i,X_i)\right)^{1/2}\end{aligned}$$

almost surely and again, by Jensen's inequality, we have

$$\mathbb{E}\left(\sup_{r>0}\sup_{f\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i r^{-1} f(X_i)\right|\right) \le \left(\frac{\|k\|_{\mathsf{diag}}}{n}\right)^{1/2}.$$

Hence, $\mathbb{E}(Z) \le 64\|k\|_{\mathsf{diag}}^{1/2} C n^{-1/2}$.

Let

$$A = \left\{ r^{-1/2} V f_1 - r^{-1/2} V f_2 : r > 0 \text{ and } f_1, f_2 \in rB_H \right\}.$$

We have that $(0,\infty)$, the set indexing $r$, is separable. Furthermore, $H$ is separable and so is separable in $L^\infty$ as it can be continuously embedded in $L^\infty$ due to its bounded kernel. Therefore, $rB_H \subseteq H$ is separable in $L^\infty$ for $r > 0$. Hence, we have that $A \subseteq L^\infty$ is separable. Furthermore,

$$\begin{aligned}\left\|r^{-1/2} V f_1 - r^{-1/2} V f_2\right\|_\infty &\le \min\left(2Cr^{-1/2}, 2\|k\|_{\mathsf{diag}}^{1/2} r^{1/2}\right)\\ &\le 2\|k\|_{\mathsf{diag}}^{1/4} C^{1/2}\end{aligned}$$

for all $r > 0$ and all $f_1, f_2 \in rB_H$. The first term in the minimum comes from clipping using $V$, while the second term comes from the continuous embedding of $H$ in $L^\infty$ due to its bounded kernel. By Lemma 4.10.1, we have

$$Z \le \mathbb{E}(Z) + \left(\frac{32\|k\|_{\mathsf{diag}} C^2 t}{n} + \frac{16\|k\|_{\mathsf{diag}}^{1/2} C\,\mathbb{E}(Z) t}{n}\right)^{1/2} + \frac{8\|k\|_{\mathsf{diag}}^{1/2} C t}{3n}$$

with probability at least $1 - e^{-t}$. We have $\mathbb{E}(Z) \leq 64\|k\|_{\mathsf{diag}}^{1/2} C n^{-1/2}$ from above. The result follows. ■

We move the bound on the distance between $V\hat{h}_r$ and $Vh_r$ from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $r \geq 0$ and $h_r \in rB_H$.

**Corollary 4.10.2** *Assume (Y) and (H). Let $t \geq 1$ and recall the definitions of $A_{1,t}$ and $A_{2,t}$ from Lemmas 4.5.2 and 4.5.3. On the set $A_{1,t} \cap A_{2,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{1,t} \cap A_{2,t}) \geq 1 - 2e^{-t}$, we have*

$$\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 \leq \frac{\|k\|_{\mathsf{diag}}^{1/2}(97C + 20\sigma)rt^{1/2}}{n^{1/2}} + \frac{8\|k\|_{\mathsf{diag}}^{1/2}Crt}{3n} + 4\|h_r - g\|_\infty^2$$

*simultaneously for all $r \geq 0$ and all $h_r \in rB_H$.*

**Proof** By Lemma 4.5.2, we have

$$\|\hat{h}_r - h_r\|_{L^2(P_n)}^2 \leq \frac{20\|k\|_{\mathsf{diag}}^{1/2}\sigma rt^{1/2}}{n^{1/2}} + 4\|h_r - g\|_\infty^2$$

for all $r \geq 0$ and all $h_r \in rB_H$, so

$$\|V\hat{h}_r - Vh_r\|_{L^2(P_n)}^2 \leq \frac{20\|k\|_{\mathsf{diag}}^{1/2}\sigma rt^{1/2}}{n^{1/2}} + 4\|h_r - g\|_\infty^2.$$

Since $\hat{h}_r, h_r \in rB_H$, by Lemma 4.5.3 we have

$$\begin{aligned}
&\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 - \|V\hat{h}_r - Vh_r\|_{L^2(P_n)}^2 \\
\leq\ &\sup_{f_1, f_2 \in rB_H} \left| \|Vf_1 - Vf_2\|_{L^2(P_n)}^2 - \|Vf_1 - Vf_2\|_{L^2(P)}^2 \right| \\
\leq\ &\frac{97\|k\|_{\mathsf{diag}}^{1/2}Crt^{1/2}}{n^{1/2}} + \frac{8\|k\|_{\mathsf{diag}}^{1/2}Crt}{3n}.
\end{aligned}$$

The result follows. ■

We bound the changes in $I_\infty(g, r)$ with $r \geq 0$ to prove Lemma 4.5.4.

**Proof of Lemma 4.5.4** We have $I_\infty(g, s) \leq I_\infty(g, r)$ since $rB_H \subseteq sB_H$. Let $h_s \in sB_H$. We have

$$\left\| \frac{r}{s} h_s - g \right\|_\infty \leq \left\| \frac{r}{s} h_s - h_s \right\|_\infty + \|h_s - g\|_\infty.$$

We have

$$\begin{aligned} \left\| \frac{r}{s} h_s - h_s \right\|_\infty &= \left(1 - \frac{r}{s}\right) \|h_s\|_\infty \\ &\leq (s - r)\|k\|_{\mathsf{diag}}^{1/2}. \end{aligned}$$

The result follows. ■

We assume $(g1)$ to bound the distance between $V\hat{h}_r$ and $g$ in the $L^2(P)$ norm for $r \geq 0$ and prove Theorem 4.5.5.

**Proof of Theorem 4.5.5** Note that $Vg = g$. We have

$$\begin{aligned} \|V\hat{h}_r - g\|_{L^2(P)}^2 &\leq \left( \|V\hat{h}_r - Vh_r\|_{L^2(P)} + \|Vh_r - g\|_{L^2(P)} \right)^2 \\ &\leq 2\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + 2\|Vh_r - g\|_{L^2(P)}^2 \\ &\leq 2\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + 2\|h_r - g\|_{L^2(P)}^2 \end{aligned}$$

for all $r \geq 0$ and all $h_r \in rB_H$. By Corollary 4.10.2, we have

$$\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 \leq \frac{\|k\|_{\mathsf{diag}}^{1/2}(97C + 20\sigma)rt^{1/2}}{n^{1/2}} + \frac{8\|k\|_{\mathsf{diag}}^{1/2}Crt}{3n} + 4\|h_r - g\|_\infty^2.$$

Hence,

$$\|V\hat{h}_r - g\|_{L^2(P)}^2 \leq \frac{2\|k\|_{\mathsf{diag}}^{1/2}(97C + 20\sigma)rt^{1/2}}{n^{1/2}} + \frac{16\|k\|_{\mathsf{diag}}^{1/2}Crt}{3n} + 10\|h_r - g\|_\infty^2.$$

Taking an infimum over $h_r \in rB_H$ proves the result. ■

## 4.11 Proof of the Goldenshluger–Lepski Method for a Fixed RKHS

We bound the distance between $\hat{h}_r$ and $\hat{h}_s$ in the $L^2(P_n)$ norm for $s \geq r \geq 0$ to prove Lemma 4.6.1.

**Proof of Lemma 4.6.1** By Lemma 4.5.2, we have

$$\begin{aligned}\|\hat{h}_r - \hat{h}_s\|_{L^2(P_n)}^2 &\leq 4\|\hat{h}_r - h_r\|_{L^2(P_n)}^2 + 4\|h_r - g\|_{L^2(P_n)}^2 \\ &\quad + 4\|g - h_s\|_{L^2(P_n)}^2 + 4\|h_s - \hat{h}_s\|_{L^2(P_n)}^2 \\ &\leq \frac{80\|k\|_{\mathsf{diag}}^{1/2}\sigma(r+s)t^{1/2}}{n^{1/2}} + 20\|h_r - g\|_\infty^2 + 20\|h_s - g\|_\infty^2\end{aligned}$$

for all $r, s \geq 0$ and all $h_r \in rB_H, h_s \in sB_H$. Taking an infimum over $h_r \in rB_H$ and $h_s \in sB_H$ gives

$$\|\hat{h}_r - \hat{h}_s\|_{L^2(P_n)}^2 \leq \frac{80\|k\|_{\mathsf{diag}}^{1/2}\sigma(r+s)t^{1/2}}{n^{1/2}} + 20I_\infty(g, r) + 20I_\infty(g, s).$$

The result follows. ■

We prove Lemma 4.6.2.

**Proof of Lemma 4.6.2** Let $K$ be the $n \times n$ symmetric matrix with $K_{i,j} = k(X_i, X_j)$. By Lemma 3.6.1, we have that $K$ is an $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$ and that there exist an orthogonal matrix $A$ and a diagonal matrix $D$ which are both $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrices on $(\Omega, \mathcal{F})$ such that $K = ADA^{\mathsf{T}}$. Furthermore, we can demand that the diagonal entries of $D$ are non-negative and non-increasing. Let $m = \operatorname{rk} K$ and

$$\rho = \left( \sum_{i=1}^{m} D_{i,i}^{-1} (A^{\mathsf{T}} Y)_i^2 \right)^{1/2},$$

which are random variables on $(\Omega, \mathcal{F})$. By Lemma 3.6.1, we have that $\hat{h}_r$ is constant in $r$ for $r \geq \rho$. Hence,

$$\begin{aligned} &\inf_{r \in R} \left( \sup_{s \in R, s \geq r} \left( \|\hat{h}_r - \hat{h}_s\|_{L^2(P_n)}^2 - \frac{\tau(r+s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau r}{n^{1/2}} \right) \\ &= \inf_{r \in R \cap [0,\rho]} \left( \sup_{s \in R, s \geq r} \left( \|\hat{h}_r - \hat{h}_s\|_{L^2(P_n)}^2 - \frac{\tau(r+s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau r}{n^{1/2}} \right). \end{aligned} \tag{4.11.1}$$

By Lemma 4.5.1, we have

$$\|\hat{h}_r - \hat{h}_s\|_{L^2(P_n)}^2 - \frac{\tau(r+s)}{n^{1/2}}$$

is continuous in $r$ for all $s \in R$ such that $s \geq r$. The supremum of a collection of lower semicontinuous functions is lower semicontinuous. Therefore,

$$\sup_{s \in R, s \geq r} \left( \|\hat{h}_r - \hat{h}_s\|_{L^2(P_n)}^2 - \frac{\tau(r+s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau r}{n^{1/2}}$$

is lower semicontinuous in $r$. Hence, the infimum (4.11.1) is attained as it is the infimum of a lower semicontinuous function on a compact set. By lower semicontinuity, $\hat{r}$ also attains the infimum and is well-defined. ■

We use the Goldenshluger–Lepski method to prove Theorem 4.6.3.

**Proof of Theorem 4.6.3** Since we assume $(Y)$ and $(H)$, we find that Lemma 4.5.2 holds, which implies that Lemma 4.6.1 holds. By our choice of $t$, we have

$$\|\hat{h}_r - \hat{h}_s\|_{L^2(P_n)}^2 \leq \frac{\tau(r+s)}{n^{1/2}} + 40 I_\infty(g, r) \tag{4.11.2}$$

simultaneously for all $s, r \in R$ such that $s \geq r \geq 0$. Fix $r \in R$ and suppose that $\hat{r} \leq r$. By the definition of $\hat{r}$ in (4.6.1) and (4.11.2), we have

$$\begin{aligned}
\|\hat{h}_{\hat{r}} - \hat{h}_r\|^2_{L^2(P_n)} &= \|\hat{h}_{\hat{r}} - \hat{h}_r\|^2_{L^2(P_n)} - \frac{\tau(\hat{r} + r)}{n^{1/2}} + \frac{\tau(\hat{r} + r)}{n^{1/2}} \\
&\leq \sup_{s \in R, s \geq \hat{r}} \left( \|\hat{h}_{\hat{r}} - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(\hat{r} + s)}{n^{1/2}} \right) + \frac{2\tau r}{n^{1/2}} \\
&\leq \sup_{s \in R, s \geq r} \left( \|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r + s)}{n^{1/2}} \right) + \frac{2(2+\nu)\tau r}{n^{1/2}} - \frac{2(1+\nu)\tau\hat{r}}{n^{1/2}} \\
&\leq 40 I_\infty(g, r) + \frac{2(2+\nu)\tau r}{n^{1/2}}.
\end{aligned}$$

This shows

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P_n)} \leq 40 I_\infty(g, r) + \frac{2(2+\nu)\tau r}{n^{1/2}},$$

and it follows from Lemma 4.5.3 and our choice of $t$ that

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P)} \leq 40 I_\infty(g, r) + \frac{2(2+\nu)\tau r}{n^{1/2}} + \frac{97 C\tau r}{80\sigma n^{1/2}} + \frac{C\tau^2 r}{2400\|k\|^{1/2}_{\mathsf{diag}}\sigma^2 n}.$$

Hence,

$$\begin{aligned}
&\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \\
&\leq 2\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P)} + 2\|V\hat{h}_r - g\|^2_{L^2(P)} \\
&\leq 80 I_\infty(g, r) + \frac{4(2+\nu)\tau r}{n^{1/2}} + \frac{97 C\tau r}{40\sigma n^{1/2}} + \frac{C\tau^2 r}{1200\|k\|^{1/2}_{\mathsf{diag}}\sigma^2 n} + 2\|V\hat{h}_r - g\|^2_{L^2(P)}.
\end{aligned}$$

Now suppose instead that $\hat{r} \geq r$. Since (4.11.2) holds simultaneously for all $s, r \in R$ such that $s \geq r \geq 0$, we have

$$\|\hat{h}_{\hat{r}} - \hat{h}_r\|^2_{L^2(P_n)} \leq \frac{\tau(r + \hat{r})}{n^{1/2}} + 40 I_\infty(g, r).$$

This shows

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P_n)} \le \frac{\tau r}{n^{1/2}} + 40I_\infty(g,r) + \frac{\tau \hat{r}}{n^{1/2}},$$

and it follows from Lemma 4.5.3 that

$$\begin{aligned}
&\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P)} \\
&\le \frac{\tau r}{n^{1/2}} + 40I_\infty(g,r) + \frac{\tau \hat{r}}{n^{1/2}} + \frac{97C\tau\hat{r}}{80\sigma n^{1/2}} + \frac{C\tau^2\hat{r}}{2400\|k\|^{1/2}_{\mathsf{diag}}\sigma^2 n} \\
&= \frac{\tau r}{n^{1/2}} + 40I_\infty(g,r) + \left(\frac{1}{2\nu} + \frac{97C}{160\sigma\nu} + \frac{C\tau}{4800\|k\|^{1/2}_{\mathsf{diag}}\sigma^2\nu n^{1/2}}\right)\frac{2\nu\tau\hat{r}}{n^{1/2}}.
\end{aligned}$$

By (4.6.2), the definition of $\hat{r}$ in (4.6.1) and (4.11.2), we have

$$\begin{aligned}
\frac{2\nu\tau\hat{r}}{n^{1/2}} &\le \sup_{s\in R, s\ge\hat{r}}\left(\|\hat{h}_{\hat{r}} - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(\hat{r}+s)}{n^{1/2}}\right) + \frac{2(1+\nu)\tau\hat{r}}{n^{1/2}} \\
&\le \sup_{s\in R, s\ge r}\left(\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r+s)}{n^{1/2}}\right) + \frac{2(1+\nu)\tau r}{n^{1/2}} \\
&\le 40I_\infty(g,r) + \frac{2(1+\nu)\tau r}{n^{1/2}}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \\
&\le 2\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P)} + 2\|V\hat{h}_r - g\|^2_{L^2(P)} \\
&\le \frac{2\tau r}{n^{1/2}} + 80I_\infty(g,r) + \left(\frac{1}{\nu} + \frac{97C}{80\sigma\nu} + \frac{C\tau}{2400\|k\|^{1/2}_{\mathsf{diag}}\sigma^2\nu n^{1/2}}\right)\left(40I_\infty(g,r) + \frac{2(1+\nu)\tau r}{n^{1/2}}\right) \\
&+ 2\|V\hat{h}_r - g\|^2_{L^2(P)}.
\end{aligned}$$

The result follows. ■

We assume $(g1)$ to bound the distance between $V\hat{h}_{\hat{r}}$ and $g$ in the $L^2(P)$ norm and prove Theorem 4.6.4.

**Proof of Theorem 4.6.4** By Theorem 4.6.3, we have

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \leq \inf_{r \in R} \left( (1 + D_4 \tau n^{-1/2})(D_5 \tau r n^{-1/2} + D_6 I_\infty(g, r)) + 2\|V\hat{h}_r - g\|^2_{L^2(P)} \right)$$

for some constants $D_4, D_5, D_6 > 0$ not depending on $\tau$, $r$ or $n$. By Theorem 4.5.5, we have

$$\begin{aligned}\|V\hat{h}_r - g\|^2_{L^2(P)} &\leq \frac{(97C + 20\sigma)\tau r}{40\sigma n^{1/2}} + \frac{C\tau^2 r}{1200\|k\|^{1/2}_{\mathsf{diag}}\sigma^2 n} + 10 I_\infty(g, r) \\ &\leq D_7 \tau r n^{-1/2} + D_8 \tau^2 r n^{-1} + 10 I_\infty(g, r).\end{aligned}$$

for all $r \in R$, for some constants $D_7, D_8 > 0$ not depending on $\tau$, $r$ or $n$. This gives

$$\begin{aligned}\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \leq \inf_{r \in R} \big( &(1 + D_4 \tau n^{-1/2})(D_5 \tau r n^{-1/2} + D_6 I_\infty(g, r)) \\ &+ 2D_7 \tau r n^{-1/2} + 2D_8 \tau^2 r n^{-1} + 20 I_\infty(g, r) \big) .\end{aligned}$$

Hence, the result follows with

$$D_1 = \frac{D_4 D_5 + 2D_8}{D_5 + 2D_7},\ D_2 = D_5 + 2D_7,\ D_3 = D_6 + 20.$$

■

We assume $(g2)$ to prove Theorem 4.6.5.

**Proof of Theorem 4.6.5** If we assume $(R1)$, then $r = a n^{(1-\beta)/(2(1+\beta))} \in R$ and

$$\begin{aligned}\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} &\leq (1 + D_3 \tau n^{-1/2})(D_4 \tau r n^{-1/2} + D_5 I_\infty(g, r)) \\ &\leq (1 + D_3 \tau n^{-1/2}) \left( D_4 \tau a n^{-\beta/(1+\beta)} + \frac{D_5 B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)} n^{\beta/(1+\beta)}} \right)\end{aligned}$$

for some constants $D_3, D_4, D_5 > 0$ not depending on $n$ or $\tau$ by Theorem 4.6.4 and

(4.6.3). If we assume $(R2)$, then there is at least one $r \in R$ such that

$$an^{(1-\beta)/(2(1+\beta))} \leq r < an^{(1-\beta)/(2(1+\beta))} + b$$

and

$$\begin{aligned}
&\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \\
&\leq (1 + D_3\tau n^{-1/2})(D_4\tau r n^{-1/2} + D_5 I_\infty(g, r)) \\
&\leq (1 + D_3\tau n^{-1/2})\left(D_4\tau(an^{(1-\beta)/(2(1+\beta))} + b)n^{-1/2} + \frac{D_5 B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)} n^{\beta/(1+\beta)}}\right)
\end{aligned}$$

by Theorem 4.6.4 and (4.6.3). In either case,

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \leq D_1\tau n^{-\beta/(1+\beta)} + D_2\tau^2 n^{-(1+3\beta)/(2(1+\beta))}$$

for some constants $D_1, D_2 > 0$ not depending on $n$ or $\tau$. ■

## 4.12 Proof of the Regression Results for a Collection of RKHSs

We prove Lemma 4.7.2.

**Proof of Lemma 4.7.2** Let $K$ be the $n \times n$ symmetric matrix with $K_{i,j} = k(X_i, X_j)$ for $k \in \mathcal{K}$. Then $K$ is a continuous function of $k$ and $X$, hence it is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega \times \mathcal{K}, \mathcal{F} \otimes \mathcal{B}(\mathcal{K}))$, where $k$ varies in $\mathcal{K}$. By Lemma 4.16.1, there exist an orthogonal matrix $A$ and a diagonal matrix $D$ which are both $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrices on $(\Omega \times \mathcal{K}, \mathcal{F} \otimes \mathcal{B}(\mathcal{K}))$ such that $K = ADA^\mathsf{T}$. Since $K$ is non-negative definite, the diagonal entries of $D$ are non-negative, and we may assume that they are non-increasing. Let $m = \operatorname{rk} K$, which is measurable on $(\Omega \times \mathcal{K}, \mathcal{F} \otimes \mathcal{B}(\mathcal{K}))$. For $r > 0$, if

$$r^2 < \sum_{i=1}^{m} D_{i,i}^{-1} (A^\mathsf{T} Y)_i^2,$$

then define $\mu(r) > 0$ by

$$\sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu(r))^2} (A^\mathsf{T} Y)_i^2 = r^2.$$

Otherwise, let $\mu(r) = 0$. Let $a \in \mathbb{R}^n$ be defined by

$$(A^\mathsf{T} a)_i = (D_{i,i} + n\mu(r))^{-1} (A^\mathsf{T} Y)_i$$

for $1 \leq i \leq m$ and $(A^\mathsf{T} a)_i = 0$ for $m + 1 \leq i \leq n$, noting that $A^\mathsf{T}$ has the inverse $A$ since it is an orthogonal matrix. By Lemma 3.6.1,

$$\hat{h}_{k,r} = \sum_{i=1}^{n} a_i k_{X_i}$$

for $r > 0$ and $\hat{h}_{k,0} = 0$ for $k \in \mathcal{K}$.

Since $\mu(r) > 0$ is strictly decreasing for

$$r^2 < \sum_{i=1}^{m} D_{i,i}^{-1} (A^\mathsf{T} Y)_i^2$$

and $\mu(r) = 0$ otherwise, we find

$$\{\mu(r) \leq \mu\} = \left\{ \sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu)^2} (A^\mathsf{T} Y)_i^2 \leq r^2 \right\}$$

for $\mu \in [0, \infty)$. Therefore, $\mu(r)$ is measurable on $(\Omega \times \mathcal{K} \times [0, \infty), \mathcal{F} \otimes \mathcal{B}(\mathcal{K}) \otimes \mathcal{B}([0, \infty)))$, where $k$ varies in $\mathcal{K}$ and $r$ varies in $[0,\infty)$. Hence, the $a$ above with $\mu = \mu(r)$ for $r > 0$ is measurable on $(\Omega \times \mathcal{K} \times [0,\infty), \mathcal{F} \otimes \mathcal{B}(\mathcal{K}) \otimes \mathcal{B}([0,\infty)))$. By Lemma 4.29 of Steinwart and Christmann (2008), $\Phi_k : S \to H_k$ by $\Phi_k(x) = k_x$ is continuous for all $k \in \mathcal{K}$. Hence, $\Phi : \mathcal{K} \times S \to L^\infty$ by $\Phi(k,x) = k_x$ is continuous and $k_{X_i}$ for $1 \le i \le n$ are $(L^\infty, \mathcal{B}(L^\infty))$-valued measurable functions on $(\Omega \times \mathcal{K}, \mathcal{F} \otimes \mathcal{B}(\mathcal{K}))$. Together, these show that $\hat{h}_{k,r}$ is an $(L^\infty, \mathcal{B}(L^\infty))$-valued measurable function on $(\Omega \times \mathcal{K} \times [0,\infty), \mathcal{F} \otimes \mathcal{B}(\mathcal{K}) \otimes \mathcal{B}([0,\infty)))$, where $k$ varies in $\mathcal{K}$ and $r$ varies in $[0,\infty)$, recalling that $\hat{h}_{k,0} = 0$. ■

Let $\psi_1(x) = \exp(|x|) - 1$ for $x \in \mathbb{R}$ and

$$\|Z\|_{\psi_1} = \inf\{a \in (0,\infty) : \mathbb{E}(\psi_1(Z/a)) \le 1\}$$

for any random variable $Z$ on $(\Omega, \mathcal{F})$. Note that this infimum is attained by the monotone convergence theorem, and $\|Z\|_{\psi_1}$ increases as $|Z|$ increases pointwise. Let $L^{\psi_1}$ be the set of random variables $Z$ on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\|Z\|_{\psi_1} < \infty$. We have that $(L^{\psi_1}, \|\cdot\|_{\psi_1})$ is a Banach space known as an Orlicz space (see Rao and Ren, 1991).

**Lemma 4.12.1** *Let $Z \in L^{\psi_1}$. We have*

$$\mathbb{E}(|Z|) \le (\log 2)\|Z\|_{\psi_1}.$$

*Let $t \ge 0$. We have*

$$|Z| \le \|Z\|_{\psi_1}(\log 2 + t)$$

*with probability at least $1 - e^{-t}$.*

**Proof** We have $\mathbb{E}(\exp(|Z|/\|Z\|_{\psi_1})) \le 2$. The first result follows from Jensen's inequality. The second result follows from Chernoff bounding. ■

For $m \times n$ matrices $U$ and $V$, define $U \circ V$ to be the $m \times n$ matrix with

$$(U \circ V)_{i,j} = U_{i,j} V_{i,j}.$$

Recall that

$$\begin{aligned} \mathcal{L} &= \{k/\|k\|_{\mathsf{diag}} : k \in \mathcal{K}\} \cup \{0\}, \\ D &= \sup_{f_1, f_2 \in \mathcal{L}} \|f_1 - f_2\|_\infty \leq 2, \\ J &= \left(162 \int_0^{D/2} \log(2N(a, \mathcal{L}, \|\cdot\|_\infty)) da + 1\right)^{1/2}. \end{aligned}$$

The following lemma is useful for proving Lemma 4.7.2.

**Lemma 4.12.2** *Assume ($\mathcal{K}$1). Let the $\varepsilon_i$ for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $(X_i, \varepsilon_i)$ are i.i.d. and $\varepsilon_i$ is $\sigma^2$-subgaussian given $X_i$. Let*

$$W(f) = \frac{1}{n^2} \sum_{i,j=1}^n \varepsilon_i \varepsilon_j f(X_i, X_j)$$

*for $f \in \mathcal{L}$. We have*

$$\left\| \sup_{f \in \mathcal{L}} W(f) \right\|_{\psi_1} \leq \frac{4J^2\sigma^2}{n}.$$

**Proof** Let $F$ be the $n \times n$ matrix with $F_{i,j} = f(X_i, X_j)$, where $F$ varies with $f \in \mathcal{L}$. Note that $F$ is an $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$. Then $W(f) = n^{-2} \varepsilon^\mathsf{T} F \varepsilon$. Let $Z(f) = n^{-2} \varepsilon^\mathsf{T} (F - I \circ F) \varepsilon$ for $f \in \mathcal{L}$. Note that $Z$ is continuous in $f$. We have

$$\|Z(f_1) - Z(f_2)\|_{\psi_1} \leq 36\sigma^2 n^{-1} \|f_1 - f_2\|_\infty$$

for $f_1, f_2 \in \mathcal{L}$ by Lemma 4.16.3. Let $d(f_1, f_2) = 36\sigma^2 n^{-1}\|f_1 - f_2\|_\infty$ for $f_1, f_2 \in \mathcal{L}$ and

$$D_d = \sup_{f_1, f_2 \in \mathcal{L}} d(f_1, f_2).$$

By Lemma 4.15.2 with $M = \mathcal{L}$ and $s_0 = 0$, we find

$$\begin{aligned} \left\| \sup_{f \in \mathcal{L}} |Z(f)| \right\|_{\psi_1} &\leq 18 \int_0^{D_d/2} \log(2N(a, \mathcal{L}, d)) da \\ &= \frac{648\sigma^2}{n} \int_0^{D/2} \log(2N(a, \mathcal{L}, \|\cdot\|_\infty)) da. \end{aligned}$$

Hence,

$$\left\| \sup_{f \in \mathcal{L}} W(f) \right\|_{\psi_1} \leq \left\| n^{-2} \sup_{f \in \mathcal{L}} \varepsilon^\mathsf{T} (I \circ F) \varepsilon \right\|_{\psi_1} + \frac{648\sigma^2}{n} \int_0^{D/2} \log(2N(a, \mathcal{L}, \|\cdot\|_\infty)) da.$$

We have

$$n^{-2} \sup_{f \in \mathcal{L}} \varepsilon^\mathsf{T} (I \circ F) \varepsilon \leq n^{-2} \varepsilon^\mathsf{T} \varepsilon,$$

noting that $F_{i,i} \in [0, 1]$ for $1 \leq i \leq n$ and $f \in \mathcal{L}$. Let $\delta_i$ for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent of each other and the $\varepsilon_i$, with $\delta_i \sim \mathrm{N}(0, \sigma^2)$. Lemma 3.16.1 shows

$$\begin{aligned} \mathbb{E}\left(\exp\left(n^{-2} t \sup_{f \in \mathcal{L}} \varepsilon^\mathsf{T} (I \circ F) \varepsilon\right)\right) &\leq \mathbb{E}\left(\exp\left(n^{-2} t \varepsilon^\mathsf{T} \varepsilon\right)\right) \\ &\leq \mathbb{E}\left(\exp\left(n^{-2} t \delta^\mathsf{T} \delta\right)\right) \\ &= \prod_{i=1}^n \left(1 - 2\sigma^2 n^{-2} t\right)^{-1/2} \end{aligned}$$

for $0 \leq 2\sigma^2 n^{-2} t < 1$ by computing the moment generating function of the $\delta_i^2$. We have that $(1-x)^{-1/2} \leq \exp(x)$ for $x \in [0, 1/2]$, so

$$\mathbb{E}\left(\exp\left(n^{-2}t\sup_{f\in\mathcal{L}}\varepsilon^{\mathsf{T}}(I\circ F)\varepsilon\right)\right) \leq \prod_{i=1}^{n}\exp\left(2\sigma^2 n^{-2}t\right) = \exp\left(2\sigma^2 n^{-1}t\right)$$

for $0 \leq 4\sigma^2 n^{-2} t \leq 1$. This bound is at most 2 and valid for

$$t \leq \min\left(\frac{n^2}{4\sigma^2}, \frac{(\log 2)n}{2\sigma^2}\right).$$

Hence,

$$\left\|n^{-2}\sup_{f\in\mathcal{L}}\varepsilon^{\mathsf{T}}(I\circ F)\varepsilon\right\|_{\psi_1} \leq \max\left(\frac{4\sigma^2}{n^2}, \frac{2\sigma^2}{(\log 2)n}\right) \leq \frac{4\sigma^2}{n}$$

and

$$\left\|\sup_{f\in\mathcal{L}} W(f)\right\|_{\psi_1} \leq \frac{648\sigma^2}{n}\int_0^{D/2}\log(2N(a,\mathcal{L},\|\cdot\|_\infty))da + \frac{4\sigma^2}{n}.$$

The result follows. ■

We bound the distance between $\hat{h}_{k,r}$ and $h_{k,r}$ in the $L^2(P_n)$ norm for $k \in \mathcal{K}$, $r \geq 0$ and $h_{k,r} \in rB_k$ to prove Lemma 4.7.2.

**Proof of Lemma 4.7.2** The result is trivial for $r = 0$. By Lemma 3.5.1, we have

$$\|\hat{h}_{k,r} - h_{k,r}\|_{L^2(P_n)}^2 \leq \frac{4}{n}\sum_{i=1}^{n}(Y_i - g(X_i))(\hat{h}_{k,r}(X_i) - h_{k,r}(X_i)) + 4\|h_{k,r} - g\|_{L^2(P_n)}^2$$

for all $k \in \mathcal{K}$, all $r > 0$ and all $h_{k,r} \in rB_k$. We now bound the right-hand side. We have

$$\|h_{k,r} - g\|_{L^2(P_n)}^2 \leq \|h_{k,r} - g\|_\infty^2.$$

Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))(\hat{h}_{k,r}(X_i) - h_{k,r}(X_i))$$

$$\begin{aligned}
&\leq \sup_{f \in 2rB_k} \left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) f(X_i) \right| \\
&= \sup_{f \in 2rB_k} \left| \left\langle \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) k_{X_i}, f \right\rangle_k \right| \\
&= 2r \left\| \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) k_{X_i} \right\|_k \\
&= 2r \left( \frac{1}{n^2} \sum_{i,j=1}^{n} (Y_i - g(X_i))(Y_j - g(X_j)) k(X_i, X_j) \right)^{1/2}
\end{aligned}$$

by the reproducing kernel property and the Cauchy–Schwarz inequality. Let

$$Z = \sup_{k \in \mathcal{K}} \left( \frac{1}{\|k\|_{\mathsf{diag}} n^2} \sum_{i,j=1}^{n} (Y_i - g(X_i))(Y_j - g(X_j)) k(X_i, X_j) \right).$$

By Lemma 4.12.2 with $\varepsilon_i = Y_i - g(X_i)$, we have $\|Z\|_{\psi_1} \leq 4J^2\sigma^2 n^{-1}$. By Lemma 4.12.1, we have $Z \leq 4J^2\sigma^2(\log 2 + t)n^{-1}$ with probability at least $1 - e^{-t}$. The result follows. ■

The following lemma is useful for proving Lemma 4.7.3.

**Lemma 4.12.3** *Let*

$$A = \left\{ \|k\|_{\mathsf{diag}}^{-1/4} r^{-1/2} V f_1 - \|k\|_{\mathsf{diag}}^{-1/4} r^{-1/2} V f_2 : k \in \mathcal{K}, r > 0 \textit{ and } f_1, f_2 \in rB_k \right\}.$$

*Then $A$ is separable as a subset of $L^\infty$.*

**Proof** By Theorem 4.21 of Steinwart and Christmann (2008), we have that

$$\left\{ \sum_{i=1}^{m} a_i k_{s_i} : m \geq 1 \text{ and } a_i \in \mathbb{R}, s_i \in S \text{ for } 1 \leq i \leq m \right\}$$

is dense in $H_k$ for $k \in \mathcal{K}$. Hence,

$$\left\{\sum_{i=1}^{m} a_i k_{s_i} : m \geq 1 \text{ and } a_i \in \mathbb{R}, s_i \in S \text{ for } 1 \leq i \leq m \text{ with } \sum_{i,j=1}^{m} a_i a_j k(s_i, s_j) \leq r^2\right\}$$

is dense in $rB_k \subseteq H_k$ for $k \in \mathcal{K}$ and $r > 0$. Since $S$ is separable, it has a countable dense subset $S_0$. Let $D_{k,r}$ be

$$\left\{\sum_{i=1}^{m} a_i k_{s_i} : m \geq 1 \text{ and } a_i \in \mathbb{Q}, s_i \in S_0 \text{ for } 1 \leq i \leq m \text{ with } \sum_{i,j=1}^{m} a_i a_j k(s_i, s_j) \leq r^2\right\}$$

for $k \in \mathcal{K}$ and $r > 0$. Since the function $\Phi_k : S \to H_k$ by $\Phi_k(x) = k_x$ is continuous by Lemma 4.29 of Steinwart and Christmann (2008), we have that $D_{k,r}$ is dense in $rB_k \subseteq H_k$ by suitable choices for $a_i \in \mathbb{Q}$ for $1 \leq i \leq m$. Since $k$ is bounded for all $k \in \mathcal{K}$, as subsets of $L^\infty$ we have that $D_{k,r}$ is dense in $rB_k$ and

$$A = \operatorname{cl}\left(\left\{\|k\|_{\mathsf{diag}}^{-1/4} r^{-1/2}(Vf_1 - Vf_2) : k \in \mathcal{K}, r > 0 \text{ and } f_1, f_2 \in D_{k,r}\right\}\right).$$

Since $(\mathcal{K}, \|\cdot\|_\infty)$ is separable, it has a countable dense subset $\mathcal{K}_0$. Hence,

$$A = \operatorname{cl}\left(\left\{\|k\|_{\mathsf{diag}}^{-1/4} r^{-1/2}(Vf_1 - Vf_2) : k \in \mathcal{K}_0, r \in (0,\infty) \cap \mathbb{Q} \text{ and } f_1, f_2 \in D_{k,r}\right\}\right)$$

by suitable choices for $r \in (0, \infty) \cap \mathbb{Q}$. The result follows. ■

We bound the supremum of the difference in the $L^2(P_n)$ norm and the $L^2(P)$ norm over $rB_k$ for $k \in \mathcal{K}$ and $r \geq 0$ to prove Lemma 4.7.3.

**Proof of Lemma 4.7.3** The result is trivial for $r = 0$. Let

$$Z = \sup_{k \in \mathcal{K}} \sup_{r > 0} \sup_{f_1, f_2 \in rB_k} \|k\|_{\mathsf{diag}}^{-1/2} r^{-1} \left| \|Vf_1 - Vf_2\|_{L^2(P_n)}^2 - \|Vf_1 - Vf_2\|_{L^2(P)}^2 \right|.$$

We have that $Z$ is a random variable by Lemma 4.12.3. Furthermore, let the $\varepsilon_i$ for

$1 \leq i \leq n$ be i.i.d. Rademacher random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, independent of the $X_i$. Lemma 2.3.1 of van der Vaart and Wellner (1996) shows

$$\mathbb{E}(Z) \leq 2\,\mathbb{E}\left(\sup_{k \in \mathcal{K}} \sup_{r>0} \sup_{f_1, f_2 \in rB_k} \|k\|_{\mathsf{diag}}^{-1/2} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (r^{-1/2} V f_1(X_i) - r^{-1/2} V f_2(X_i))^2 \right| \right)$$

by symmetrisation. Since

$$|Vf_1(X_i) - Vf_2(X_i)| \leq 2C$$

for all $k \in \mathcal{K}$, all $r > 0$ and all $f_1, f_2 \in rB_k$, we find

$$\frac{(r^{-1/2} V f_1(X_i) - r^{-1/2} V f_2(X_i))^2}{4C}$$

is a contraction vanishing at 0 as a function of $r^{-1} V f_1(X_i) - r^{-1} V f_2(X_i)$ for all $1 \leq i \leq n$. By Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}\left(\sup_{k \in \mathcal{K}} \sup_{r>0} \sup_{f_1, f_2 \in rB_k} \|k\|_{\mathsf{diag}}^{-1/2} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \frac{(r^{-1/2} V f_1(X_i) - r^{-1/2} V f_2(X_i))^2}{4C} \right| \,\middle|\, X \right)$$

is at most

$$2\,\mathbb{E}\left(\sup_{k \in \mathcal{K}} \sup_{r>0} \sup_{f_1, f_2 \in rB_k} \|k\|_{\mathsf{diag}}^{-1/2} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (r^{-1} V f_1(X_i) - r^{-1} V f_2(X_i)) \right| \,\middle|\, X \right)$$

almost surely. Therefore,

$$\begin{aligned}\mathbb{E}(Z) &\leq 16C\,\mathbb{E}\left(\sup_{k \in \mathcal{K}} \sup_{r>0} \sup_{f_1, f_2 \in rB_k} \|k\|_{\mathsf{diag}}^{-1/2} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (r^{-1} V f_1(X_i) - r^{-1} V f_2(X_i)) \right| \right) \\ &\leq 32C\,\mathbb{E}\left(\sup_{k \in \mathcal{K}} \sup_{r>0} \sup_{f \in rB_k} \|k\|_{\mathsf{diag}}^{-1/2} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i r^{-1} V f(X_i) \right| \right)\end{aligned}$$

by the triangle inequality. Again, by Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}(Z) \leq 64C\,\mathbb{E}\left(\sup_{k\in\mathcal{K}}\sup_{r>0}\sup_{f\in rB_k}\|k\|_{\mathsf{diag}}^{-1/2}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i r^{-1}f(X_i)\right|\right)$$

since $V$ is a contraction vanishing at 0. We have

$$\begin{aligned}
&\sup_{k\in\mathcal{K}}\sup_{r>0}\sup_{f\in rB_k}\|k\|_{\mathsf{diag}}^{-1/2}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i r^{-1}f(X_i)\right| \\
&= \sup_{k\in\mathcal{K}}\sup_{r>0}\sup_{f\in rB_k}\|k\|_{\mathsf{diag}}^{-1/2}\left|\left\langle\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i k_{X_i}, r^{-1}f\right\rangle_k\right| \\
&= \sup_{k\in\mathcal{K}}\|k\|_{\mathsf{diag}}^{-1/2}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i k_{X_i}\right\|_k \\
&= \sup_{k\in\mathcal{K}}\|k\|_{\mathsf{diag}}^{-1/2}\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\varepsilon_i\varepsilon_j k(X_i,X_j)\right)^{1/2}
\end{aligned}$$

by the reproducing kernel property and the Cauchy–Schwarz inequality. By Lemma 4.12.2 with $\sigma^2 = 1$, Lemma 4.12.1 and Jensen's inequality, we have $\mathbb{E}(Z) \leq 107JCn^{-1/2}$.

Let

$$A = \left\{\|k\|_{\mathsf{diag}}^{-1/4}r^{-1/2}Vf_1 - \|k\|_{\mathsf{diag}}^{-1/4}r^{-1/2}Vf_2 : k \in \mathcal{K}, r > 0 \text{ and } f_1, f_2 \in rB_k\right\}.$$

We have that $A \subseteq L^\infty$ is separable by Lemma 4.12.3. Furthermore,

$$\begin{aligned}
\left\|\|k\|_{\mathsf{diag}}^{-1/4}r^{-1/2}Vf_1 - \|k\|_{\mathsf{diag}}^{-1/4}r^{-1/2}Vf_2\right\|_\infty &\leq \min\left(2C\|k\|_{\mathsf{diag}}^{-1/4}r^{-1/2}, 2\|k\|_{\mathsf{diag}}^{1/4}r^{1/2}\right) \\
&\leq 2C^{1/2}
\end{aligned}$$

for all $k \in \mathcal{K}$, all $r > 0$ and all $f_1, f_2 \in rB_k$. By Lemma 4.10.1, we have

$$Z \leq \mathbb{E}(Z) + \left(\frac{32C^2t}{n} + \frac{16C\,\mathbb{E}(Z)t}{n}\right)^{1/2} + \frac{8Ct}{3n}$$

with probability at least $1 - e^{-t}$. We have $\mathbb{E}(Z) \leq 107JCn^{-1/2}$ from above. The result follows. ■

We move the bound on the distance between $V\hat{h}_{k,r}$ and $Vh_{k,r}$ from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $k \in \mathcal{K}$, $r \geq 0$ and $h_{k,r} \in rB_k$.

**Corollary 4.12.4** *Assume (Y) and ($\mathcal{K}$1). Let $t \geq 1$ and recall the definitions of $A_{3,t}$ and $A_{4,t}$ from Lemmas 4.7.2 and 4.7.3. On the set $A_{3,t} \cap A_{4,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{3,t} \cap A_{4,t}) \geq 1 - 2e^{-t}$, we have*

$$\|V\hat{h}_{k,r} - Vh_{k,r}\|^2_{L^2(P)} \leq \frac{J\|k\|^{1/2}_{\mathsf{diag}}(151C + 21\sigma)rt^{1/2}}{n^{1/2}} + \frac{8\|k\|^{1/2}_{\mathsf{diag}}Crt}{3n} + 4\|h_{k,r} - g\|^2_\infty$$

*simultaneously for all $k \in \mathcal{K}$, all $r \geq 0$ and all $h_{k,r} \in rB_k$.*

**Proof** By Lemma 4.7.2, we have

$$\|\hat{h}_{k,r} - h_{k,r}\|^2_{L^2(P_n)} \leq \frac{21J\|k\|^{1/2}_{\mathsf{diag}}\sigma rt^{1/2}}{n^{1/2}} + 4\|h_{k,r} - g\|^2_\infty$$

for all $k \in \mathcal{K}$, all $r \geq 0$ and all $h_{k,r} \in rB_k$, so

$$\|V\hat{h}_{k,r} - Vh_{k,r}\|^2_{L^2(P_n)} \leq \frac{21J\|k\|^{1/2}_{\mathsf{diag}}\sigma rt^{1/2}}{n^{1/2}} + 4\|h_{k,r} - g\|^2_\infty.$$

Since $\hat{h}_{k,r}, h_{k,r} \in rB_k$, by Lemma 4.7.3 we have

$$\begin{aligned}
&\|V\hat{h}_{k,r} - Vh_{k,r}\|^2_{L^2(P)} - \|V\hat{h}_{k,r} - Vh_{k,r}\|^2_{L^2(P_n)} \\
\leq &\sup_{f_1, f_2 \in rB_k} \left| \|Vf_1 - Vf_2\|^2_{L^2(P_n)} - \|Vf_1 - Vf_2\|^2_{L^2(P)} \right| \\
\leq &\frac{151J\|k\|^{1/2}_{\mathsf{diag}}Crt^{1/2}}{n^{1/2}} + \frac{8\|k\|^{1/2}_{\mathsf{diag}}Crt}{3n}.
\end{aligned}$$

The result follows. ■

We assume $(g1)$ to bound the distance between $V\hat{h}_{k,r}$ and $g$ in the $L^2(P)$ norm for $k \in \mathcal{K}$ and $r \geq 0$ and prove Theorem 4.7.4.

**Proof of Theorem 4.7.4** Note that $Vg = g$. We have

$$\begin{aligned}\|V\hat{h}_{k,r} - g\|^2_{L^2(P)} &\leq \Big(\|V\hat{h}_{k,r} - Vh_{k,r}\|_{L^2(P)} + \|Vh_{k,r} - g\|_{L^2(P)}\Big)^2 \\ &\leq 2\|V\hat{h}_{k,r} - Vh_{k,r}\|^2_{L^2(P)} + 2\|Vh_{k,r} - g\|^2_{L^2(P)} \\ &\leq 2\|V\hat{h}_{k,r} - Vh_{k,r}\|^2_{L^2(P)} + 2\|h_{k,r} - g\|^2_{L^2(P)}\end{aligned}$$

for all $k \in \mathcal{K}$, all $r \geq 0$ and all $h_{k,r} \in rB_k$. By Corollary 4.12.4, we have

$$\|V\hat{h}_{k,r} - Vh_{k,r}\|^2_{L^2(P)} \leq \frac{J\|k\|^{1/2}_{\mathsf{diag}}(151C + 21\sigma)rt^{1/2}}{n^{1/2}} + \frac{8\|k\|^{1/2}_{\mathsf{diag}}Crt}{3n} + 4\|h_{k,r} - g\|^2_{\infty}.$$

Hence,

$$\|V\hat{h}_{k,r} - g\|^2_{L^2(P)} \leq \frac{2J\|k\|^{1/2}_{\mathsf{diag}}(151C + 21\sigma)rt^{1/2}}{n^{1/2}} + \frac{16\|k\|^{1/2}_{\mathsf{diag}}Crt}{3n} + 10\|h_{k,r} - g\|^2_{\infty}.$$

Taking an infimum over $h_{k,r} \in rB_k$ proves the result. ■

## 4.13 Proof of the Goldenshluger–Lepski Method for a Collection of RKHSs with Gaussian Kernels

We bound the distance between $\hat{h}_{\gamma,r}$ and $\hat{h}_{\eta,s}$ in the $L^2(P_n)$ norm for $\gamma, \eta \in \Gamma$ with $\eta \leq \gamma$ and $s \geq r \geq 0$ to prove Lemma 4.8.4.

**Proof of Lemma 4.8.4** By Lemma 4.7.2, we have

$$\begin{aligned}\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L^2(P_n)} &\leq 4\|\hat{h}_{\gamma,r} - h_{\gamma,r}\|^2_{L^2(P_n)} + 4\|h_{\gamma,r} - g\|^2_{L^2(P_n)} \\ &\quad + 4\|g - h_{\eta,s}\|^2_{L^2(P_n)} + 4\|h_{\eta,s} - \hat{h}_{\eta,s}\|^2_{L^2(P_n)} \\ &\leq \frac{84J\sigma(\gamma^{-d/2}r + \eta^{-d/2}s)t^{1/2}}{n^{1/2}} + 20\|h_{\gamma,r} - g\|^2_\infty + 20\|h_{\eta,s} - g\|^2_\infty\end{aligned}$$

for all $\gamma, \eta \in \Gamma$, all $r, s \geq 0$ and all $h_{\gamma,r} \in rB_\gamma, h_{\eta,s} \in sB_\eta$. Taking an infimum over $h_{\gamma,r} \in rB_\gamma$ and $h_{\eta,s} \in sB_\eta$ gives

$$\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L^2(P_n)} \leq \frac{84J\sigma(\gamma^{-d/2}r + \eta^{-d/2}s)t^{1/2}}{n^{1/2}} + 20I_\infty(g,\gamma,r) + 20I_\infty(g,\eta,s).$$

The result follows from Lemma 4.8.1. ■

We use the Goldenshluger–Lepski method to prove Theorem 4.8.5.

**Proof of Theorem 4.8.5** Since we assume $(Y)$ and $(\mathcal{K}2)$, which implies $(\mathcal{K}1)$, we find that Lemma 4.7.2 holds, which implies that Lemma 4.8.4 holds. By our choice of $t$, we have

$$\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L^2(P_n)} \leq \frac{\tau(\gamma^{-d/2}r + \eta^{-d/2}s)}{n^{1/2}} + 40I_\infty(g,\gamma,r) \tag{4.13.1}$$

simultaneously for all $\gamma, \eta \in \Gamma$ and all $r, s \in R$ such that $\eta \leq \gamma$ and $s \geq r$. Fix $\gamma \in \Gamma$ and $r \in R$. Then

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\gamma,r}\|^2_{L_2(P)} \leq 2\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P)} + 2\|Vh_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r} - V\hat{h}_{\gamma,r}\|^2_{L_2(P)}.$$

We now bound the right-hand side. By $\Gamma \subseteq [u,v]$, the definition of $(\hat{\gamma},\hat{r})$ in (4.8.1) and (4.13.1), we have

$$\begin{aligned}
&\|\hat{h}_{\hat{\gamma},\hat{r}} - \hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P_n)} \\
&= \|\hat{h}_{\hat{\gamma},\hat{r}} - \hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P_n)} - \frac{\tau(\hat{\gamma}^{-d/2}\hat{r} + (\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r))}{n^{1/2}} \\
&+ \frac{\tau(\hat{\gamma}^{-d/2}\hat{r} + (\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}+r))}{n^{1/2}} \\
&\leq \sup_{\eta\in\Gamma,\eta\leq\hat{\gamma}} \sup_{s\in R,s\geq\hat{r}} \left( \|\hat{h}_{\hat{\gamma},\hat{r}} - \hat{h}_{\eta,s}\|^2_{L_2(P_n)} - \frac{\tau\left(\hat{\gamma}^{-d/2}\hat{r} + \eta^{-d/2}s\right)}{n^{1/2}} \right) \\
&+ \frac{\tau(\hat{\gamma}^{-d/2}\hat{r} + (v/u)^{d/2}(\hat{\gamma}^{-d/2}\hat{r} + \gamma^{-d/2}r))}{n^{1/2}} \\
&\leq \sup_{\eta\in\Gamma,\eta\leq\gamma} \sup_{s\in R,s\geq r} \left( \|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L_2(P_n)} - \frac{\tau(\gamma^{-d/2}r + \eta^{-d/2}s)}{n^{1/2}} \right) \\
&+ \frac{2(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}} - \frac{2(1+\nu)\tau\hat{\gamma}^{-d/2}\hat{r}}{n^{1/2}} + \frac{v^{d/2}\tau(2\hat{\gamma}^{-d/2}\hat{r} + \gamma^{-d/2}r)}{u^{d/2}n^{1/2}} \\
&\leq 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2\nu)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{2v^{d/2}\tau\hat{\gamma}^{-d/2}\hat{r}}{u^{d/2}n^{1/2}}.
\end{aligned}$$

This shows

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P_n)} \leq 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2\nu)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{2v^{d/2}\tau\hat{\gamma}^{-d/2}\hat{r}}{u^{d/2}n^{1/2}},$$

and it follows from Lemma 4.7.3, our choice of $t$ and $\Gamma \subseteq [u,v]$ that

$$\begin{aligned}
&\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P)} \\
&\leq 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2\nu)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{2v^{d/2}\tau\hat{\gamma}^{-d/2}\hat{r}}{u^{d/2}n^{1/2}} \\
&+ \left( \frac{151C\tau}{84\sigma n^{1/2}} + \frac{C\tau^2}{2646J^2\sigma^2 n} \right) (\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r) \\
&\leq 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2\nu)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{2v^{d/2}\tau\hat{\gamma}^{-d/2}\hat{r}}{u^{d/2}n^{1/2}} \\
&+ \left( \frac{151C\tau}{84\sigma n^{1/2}} + \frac{C\tau^2}{2646J^2\sigma^2 n} \right) (v/u)^{d/2}(\hat{\gamma}^{-d/2}\hat{r} + \gamma^{-d/2}r) \\
&= 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2\nu)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{151Cv^{d/2}\tau\gamma^{-d/2}r}{84u^{d/2}\sigma n^{1/2}} + \frac{Cv^{d/2}\tau^2\gamma^{-d/2}r}{2646J^2u^{d/2}\sigma^2 n} \\
&+ \left( \frac{v^{d/2}}{u^{d/2}\nu} + \frac{151Cv^{d/2}}{168u^{d/2}\sigma\nu} + \frac{Cv^{d/2}\tau}{5292J^2u^{d/2}\sigma^2\nu n^{1/2}} \right) \frac{2\nu\tau\hat{\gamma}^{-d/2}\hat{r}}{n^{1/2}}.
\end{aligned}$$

By (4.8.2), the definition of $(\hat{\gamma}, \hat{r})$ in (4.8.1) and (4.13.1), we have

$$\begin{aligned}
&\frac{2\nu\tau\hat{\gamma}^{-d/2}\hat{r}}{n^{1/2}} \\
&\leq \sup_{\eta\in\Gamma,\eta\leq\hat{\gamma}} \sup_{s\in R,s\geq\hat{r}} \left( \|\hat{h}_{\hat{\gamma},\hat{r}} - \hat{h}_{\eta,s}\|^2_{L_2(P_n)} - \frac{\tau(\hat{\gamma}^{-d/2}\hat{r} + \eta^{-d/2}s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau\hat{\gamma}^{-d/2}\hat{r}}{n^{1/2}} \\
&\leq \sup_{\eta\in\Gamma,\eta\leq\gamma} \sup_{s\in R,s\geq r} \left( \|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L_2(P_n)} - \frac{\tau(\gamma^{-d/2}r + \eta^{-d/2}s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}} \\
&\leq 40I_\infty(g,\gamma,r) + \frac{2(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}}. \qquad (4.13.2)
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P)} \\
&\leq 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2\nu)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{151Cv^{d/2}\tau\gamma^{-d/2}r}{84u^{d/2}\sigma n^{1/2}} + \frac{Cv^{d/2}\tau^2\gamma^{-d/2}r}{2646J^2u^{d/2}\sigma^2 n} \\
&+ \left( \frac{2v^{d/2}}{u^{d/2}\nu} + \frac{151Cv^{d/2}}{84u^{d/2}\sigma\nu} + \frac{Cv^{d/2}\tau}{2646J^2u^{d/2}\sigma^2\nu n^{1/2}} \right) \left( 20I_\infty(g,\gamma,r) + \frac{(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}} \right).
\end{aligned}$$

Since (4.13.1) holds simultaneously for all $\gamma, \eta \in \Gamma$ and all $r, s \in R$ such that $\eta \leq \gamma$ and $s \geq r$, we have

$$\|\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r} - \hat{h}_{\gamma,r}\|^2_{L_2(P_n)} \leq 40I_\infty(g,\gamma,r) + \frac{\tau(\gamma^{-d/2}r + (\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r))}{n^{1/2}}.$$

This shows

$$\|V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r} - V\hat{h}_{\gamma,r}\|^2_{L_2(P_n)} \leq 40I_\infty(g,\gamma,r) + \frac{\tau(\gamma^{-d/2}r + (\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r))}{n^{1/2}},$$

and it follows from Lemma 4.7.3, our choice of $t$ and (4.13.2) that

$$\begin{aligned}
&\|V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r} - V\hat{h}_{\gamma,r}\|^2_{L_2(P)} \\
&\leq 40I_\infty(g,\gamma,r) + \frac{\tau(\gamma^{-d/2}r + (\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r))}{n^{1/2}}
\end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{151C\tau}{84\sigma n^{1/2}}+\frac{C\tau^2}{2646J^2\sigma^2 n}\right)(\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r)\\
&=40I_\infty(g,\gamma,r)+\frac{\tau\gamma^{-d/2}r}{n^{1/2}}+\left(\frac{\tau}{n^{1/2}}+\frac{151C\tau}{84\sigma n^{1/2}}+\frac{C\tau^2}{2646J^2\sigma^2 n}\right)(\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r)\\
&\le 40I_\infty(g,\gamma,r)+\frac{\tau\gamma^{-d/2}r}{n^{1/2}}\\
&+\left(\frac{\tau}{n^{1/2}}+\frac{151C\tau}{84\sigma n^{1/2}}+\frac{C\tau^2}{2646J^2\sigma^2 n}\right)(v/u)^{d/2}(\hat{\gamma}^{-d/2}\hat{r}+\gamma^{-d/2}r)\\
&\le 40I_\infty(g,\gamma,r)+\frac{2v^{d/2}\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}}+\frac{151Cv^{d/2}\tau\gamma^{-d/2}r}{84u^{d/2}\sigma n^{1/2}}+\frac{Cv^{d/2}\tau^2\gamma^{-d/2}r}{2646J^2u^{d/2}\sigma^2 n}\\
&+\left(\frac{v^{d/2}}{2u^{d/2}\nu}+\frac{151Cv^{d/2}}{168u^{d/2}\sigma\nu}+\frac{Cv^{d/2}\tau}{5292J^2u^{d/2}\sigma^2\nu n^{1/2}}\right)\frac{2\nu\tau\hat{\gamma}^{-d/2}\hat{r}}{n^{1/2}}\\
&\le 40I_\infty(g,\gamma,r)+\frac{2v^{d/2}\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}}+\frac{151Cv^{d/2}\tau\gamma^{-d/2}r}{84u^{d/2}\sigma n^{1/2}}+\frac{Cv^{d/2}\tau^2\gamma^{-d/2}r}{2646J^2u^{d/2}\sigma^2 n}\\
&+\left(\frac{v^{d/2}}{u^{d/2}\nu}+\frac{151Cv^{d/2}}{84u^{d/2}\sigma\nu}+\frac{Cv^{d/2}\tau}{2646J^2u^{d/2}\sigma^2\nu n^{1/2}}\right)\left(20I_\infty(g,\gamma,r)+\frac{(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}}\right).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\|V\hat{h}_{\hat{\gamma},\hat{r}}-V\hat{h}_{\gamma,r}\|^2_{L_2(P)}\\
&\le 2\|V\hat{h}_{\hat{\gamma},\hat{r}}-V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P)}+2\|Vh_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}-V\hat{h}_{\gamma,r}\|^2_{L_2(P)}\\
&\le 160I_\infty(g,\gamma,r)+\frac{2v^{d/2}(5+2\nu)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}}+\frac{151Cv^{d/2}\tau\gamma^{-d/2}r}{21u^{d/2}\sigma n^{1/2}}+\frac{2Cv^{d/2}\tau^2\gamma^{-d/2}r}{1323J^2u^{d/2}\sigma^2 n}\\
&+\left(\frac{6v^{d/2}}{u^{d/2}\nu}+\frac{151Cv^{d/2}}{21u^{d/2}\sigma\nu}+\frac{2Cv^{d/2}\tau}{1323J^2u^{d/2}\sigma^2\nu n^{1/2}}\right)\left(20I_\infty(g,\gamma,r)+\frac{(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}}\right).
\end{aligned}
$$

We have

$$
\|V\hat{h}_{\hat{\gamma},\hat{r}}-g\|^2_{L_2(P)}\le 2\|V\hat{h}_{\hat{\gamma},\hat{r}}-V\hat{h}_{\gamma,r}\|^2_{L_2(P)}+2\|V\hat{h}_{\gamma,r}-g\|^2_{L_2(P)}
$$

and the result follows. ■

We assume $(g1)$ to bound the distance between $V\hat{h}_{\hat{\gamma},\hat{r}}$ and $g$ in the $L^2(P)$ norm and prove Theorem 4.8.6.

**Proof of Theorem 4.8.6** By Theorem 4.8.5, we have

$$\begin{aligned}
&\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|_{L^2(P)}^2 \\
&\leq \inf_{\gamma\in\Gamma} \inf_{r\in R} \left((1 + D_4\tau n^{-1/2})(D_5\tau\gamma^{-d/2}rn^{-1/2} + D_6 I_\infty(g,\gamma,r)) + 2\|V\hat{h}_{\gamma,r} - g\|_{L^2(P)}^2\right)
\end{aligned}$$

for some constants $D_4, D_5, D_6 > 0$ not depending on $\tau$, $\gamma$, $r$ or $n$. By Theorem 4.7.4, we have

$$\begin{aligned}
\|V\hat{h}_{\gamma,r} - g\|_{L^2(P)}^2 &\leq \frac{(151C + 21\sigma)\tau\gamma^{-d/2}r}{42\sigma n^{1/2}} + \frac{C\tau^2\gamma^{-d/2}r}{1323J^2\sigma^2 n} + 10I_\infty(g,\gamma,r) \\
&\leq D_7\tau\gamma^{-d/2}rn^{-1/2} + D_8\tau^2\gamma^{-d/2}rn^{-1} + 10I_\infty(g,\gamma,r)
\end{aligned}$$

for all $\gamma \in \Gamma$ and all $r \in R$, for some constants $D_7, D_8 > 0$ not depending on $\tau$, $\gamma$, $r$ or $n$. This gives

$$\begin{aligned}
\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|_{L^2(P)}^2 \leq \inf_{\gamma\in\Gamma} \inf_{r\in R} \big(&(1 + D_4\tau n^{-1/2})(D_5\tau\gamma^{-d/2}rn^{-1/2} + D_6 I_\infty(g,\gamma,r)) \\
&+ 2D_7\tau\gamma^{-d/2}rn^{-1/2} + 2D_8\tau^2\gamma^{-d/2}rn^{-1} + 20I_\infty(g,\gamma,r)\big) .
\end{aligned}$$

Hence, the result follows with

$$D_1 = \frac{D_4D_5 + 2D_8}{D_5 + 2D_7},\ D_2 = D_5 + 2D_7,\ D_3 = D_6 + 20.$$

■

We assume $(g3)$ to prove Theorem 4.8.7.

**Proof of Theorem 4.8.7** If we assume $(R1)$ and $(\Gamma 1)$, then $\alpha \in \Gamma$ and $r = an^{(1-\beta)/(2(1+\beta))} \in R$, so

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|_{L^2(P)}^2 \leq (1 + D_3\tau n^{-1/2})(D_4\tau\alpha^{-d/2}rn^{-1/2} + D_5 I_\infty(g,\alpha,r))$$

$$\leq (1 + D_3\tau n^{-1/2})\left(D_4\tau\alpha^{-d/2}an^{-\beta/(1+\beta)} + \frac{D_5B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}\right)$$

for some constants $D_3, D_4, D_5 > 0$ not depending on $n$ or $\tau$ by Theorem 4.8.6 and (4.8.3). If we assume ($R2$) and ($\Gamma2$), then there is at least one $\gamma \in \Gamma$ such that $\alpha/c < \gamma \leq \alpha$ and at least one $r \in R$ such that

$$an^{(1-\beta)/(2(1+\beta))} \leq r < an^{(1-\beta)/(2(1+\beta))} + b.$$

By Theorem 4.8.6, Lemma 4.8.1 and (4.8.3), we have

$$\begin{aligned}
&\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L^2(P)} \\
&\leq (1 + D_3\tau n^{-1/2})(D_4\tau\gamma^{-d/2}rn^{-1/2} + D_5I_\infty(g,\gamma,r)) \\
&\leq (1 + D_3\tau n^{-1/2})\left(D_4\tau c^{d/2}\alpha^{-d/2}(an^{(1-\beta)/(2(1+\beta))} + b)n^{-1/2} + \frac{D_5B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}\right).
\end{aligned}$$

In either case,

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L^2(P)} \leq D_1\tau n^{-\beta/(1+\beta)} + D_2\tau^2 n^{-(1+3\beta)/(2(1+\beta))}$$

for some constants $D_1, D_2 > 0$ not depending on $n$ or $\tau$. ■

## 4.14 Covering Numbers for Gaussian Kernels

Recall that

$$\mathcal{L} = \left\{f_\gamma(x_1, x_2) = \exp\left(-\|x_1 - x_2\|_2^2/\gamma^2\right) : \gamma \in \Gamma \text{ and } x_1, x_2 \in S\right\} \cup \{0\}.$$

for $\Gamma \subseteq [u, v]$ non-empty for $v \geq u > 0$. We prove a continuity result about the function $F : \Gamma \to \mathcal{L} \setminus \{0\}$ by $F(\gamma) = f_\gamma$. We also bound the covering numbers of $\mathcal{L}$.

**Lemma 4.14.1** *Assume ($\mathcal{K}$2). Let $\gamma, \eta \in \Gamma$. We have*

$$\|f_\gamma - f_\eta\|_\infty \leq \frac{(\gamma^2 - \eta^2)^{1/2}}{\gamma \vee \eta}.$$

*For $a \in (0, 1)$, we have $N(a, \mathcal{L}, \|\cdot\|_\infty) \leq \log(v/u)a^{-2} + 2$. For $a \geq 1$, we have $N(a, \mathcal{L}, \|\cdot\|_\infty) = 1$.*

**Proof** Let $\gamma \geq \eta$ and $x_1, x_2 \in S$. We have

$$\begin{aligned} |f_\gamma(x_1, x_2) - f_\eta(x_1, x_2)| &= f_\gamma(x_1, x_2) - f_\eta(x_1, x_2) \\ &\leq \exp\left(-\|x_1 - x_2\|_2^2/\gamma^2\right). \end{aligned}$$

This is at most $a \in (0, 1)$ whenever $\|x_1 - x_2\|_2 > \gamma \log(1/a)^{1/2}$. Suppose $\|x_1 - x_2\|_2 \leq \gamma \log(1/a)^{1/2}$. We have

$$\begin{aligned} |f_\gamma(x_1, x_2) - f_\eta(x_1, x_2)| &= f_\gamma(x_1, x_2) - f_\eta(x_1, x_2) \\ &\leq \exp\left(\|x_1 - x_2\|_2^2/\eta^2\right)\left(f_\gamma(x_1, x_2) - f_\eta(x_1, x_2)\right) \\ &= \exp\left(\|x_1 - x_2\|_2^2\left(\eta^{-2} - \gamma^{-2}\right)\right) - 1 \\ &\leq \exp\left(\log(1/a)\left((\gamma/\eta)^2 - 1\right)\right) - 1. \end{aligned}$$

This is at most $a$ whenever

$$\gamma \leq \left(1 + \frac{\log(1 + a)}{\log(1/a)}\right)^{1/2} \eta. \tag{4.14.1}$$

Since $x/(1+x) \leq \log(1+x) \leq x$ for $x \geq 0$, we have

$$\begin{aligned}\left(1+\frac{\log(1+a)}{\log(1/a)}\right)^{1/2} &= \left(1+\frac{\log(1+a)}{\log(1+(1-a)/a)}\right)^{1/2} \\ &\geq \left(1+\frac{a/(1+a)}{(1-a)/a}\right)^{1/2} \\ &= \left(1+\frac{a^2}{1-a^2}\right)^{1/2}.\end{aligned}$$

Hence, (4.14.1) holds whenever

$$\gamma \leq \left(1+\frac{a^2}{1-a^2}\right)^{1/2}\eta,$$

or

$$\log(\gamma) \leq \frac{1}{2}\log\left(1+\frac{a^2}{1-a^2}\right)+\log(\eta).$$

The first result follows by rearranging for $a$.

Since

$$\log\left(1+\frac{a^2}{1-a^2}\right) \geq \frac{a^2/(1-a^2)}{1+a^2/(1-a^2)} = a^2,$$

(4.14.1) holds whenever $\log(\gamma) \leq a^2/2+\log(\eta)$. Hence, for any $\gamma,\eta \in \Gamma$, we find $\|f_\gamma - f_\eta\|_\infty \leq a$ whenever $|\log(\gamma)-\log(\eta)| \leq a^2/2$. Let $b \geq 1$ and $\gamma_i \in \Gamma$ for $1 \leq i \leq b$. Recall that $\Gamma \subseteq [u,v]$. If we let

$$\log(\gamma_i) = \log(u) + a^2(2i-1)/2$$

and let $b$ be such that

$$\log(v) - \big(\log(u)+a^2(2b-1)/2\big) \leq a^2/2,$$

then we find the $f_{\gamma_i}$ for $1 \leq i \leq b$ form an $a$ cover of $(\mathcal{L}\setminus\{0\}, \|\cdot\|_\infty)$. Rearranging the above shows that we can choose

$$b = \left\lceil \frac{\log(v/u)}{a^2} \right\rceil$$

and the second result follows by adding $\{0\}$ to the cover. The third result follows from the fact that $f_\gamma(x_1, x_2) \in (0, 1]$ for all $\gamma \in \Gamma$ and all $x_1, x_2 \in S$. ■

We calculate an integral of these covering numbers.

**Lemma 4.14.2** *Assume ($\mathcal{K}2$). We have*

$$\int_0^{1/2} \log N(a, \mathcal{L}, \|\cdot\|_\infty) da \leq \frac{\log(2 + 4\log(v/u))}{2} + 1.$$

**Proof** We have

$$\int_0^{1/2} \log N(a, \mathcal{L}, \|\cdot\|_\infty) da \leq \int_0^{1/2} \log\left(2 + \log(v/u)a^{-2}\right) da$$

by Lemma 4.14.1. Changing variables to $b = 2a$ gives

$$\begin{aligned} \frac{1}{2}\int_0^1 \log\left(2 + 4\log(v/u)b^{-2}\right) db &\leq \frac{1}{2}\int_0^1 \log\left((2 + 4\log(v/u))b^{-2}\right) db \\ &= \frac{\log(2 + 4\log(v/u))}{2} + \int_0^1 \log(b^{-1}) db. \end{aligned}$$

Changing variables to $s = \log(b^{-1})$ shows

$$\int_0^1 \log\left(b^{-1}\right) db = \int_0^\infty s\exp(-s) ds = 1$$

since the last integral is the mean of an Exponential(1) random variable. ■

## 4.15 The Orlicz Space $L^{\psi_1}$

Recall that $\psi_1(x) = \exp(|x|) - 1$ for $x \in \mathbb{R}$,

$$\|Z\|_{\psi_1} = \inf\{a \in (0,\infty) : \mathbb{E}(\psi_1(Z/a)) \leq 1\}$$

for any random variable $Z$ on $(\Omega, \mathcal{F})$ and $L^{\psi_1}$ is the set of random variables $Z$ on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\|Z\|_{\psi_1} < \infty$. We have that $(L^{\psi_1}, \|\cdot\|_{\psi_1})$ is a Banach space known as an Orlicz space (see Rao and Ren, 1991). For $t \geq 0$, also recall that

$$\mathbb{E}(|Z|) \leq (\log 2)\|Z\|_{\psi_1} \text{ and } |Z| \leq \|Z\|_{\psi_1}(\log 2 + t)$$

with probability at least $1 - e^{-t}$ by Lemma 4.12.1. We prove a maximal inequality in $L^{\psi_1}$ using the same method as Lemma 2.3.3 of Giné and Nickl (2016).

**Lemma 4.15.1** *Let $Z_i \in L^{\psi_1}$ for $1 \leq i \leq I$. Then*

$$\left\| \max_{1 \leq i \leq I} |Z_i| \right\|_{\psi_1} \leq \frac{\log(2I)}{\log(5/4)} \max_{1 \leq i \leq I} \|Z_i\|_{\psi_1}.$$

**Proof** Let $M = \max_{1 \leq i \leq I} \|Z_i\|_{\psi_1}$. Also, let $C \geq 1$ and $a \in (0,\infty)$. By Lemma 4.12.1, we have

$$\begin{aligned}
\mathbb{E}\left(\exp\left(\max_{1 \leq i \leq I} |Z_i|/a\right)\right) &= \int_0^\infty \mathbb{P}\left(\max_{1 \leq i \leq I} |Z_i| > a \log t\right) dt \\
&\leq C + \int_C^\infty \mathbb{P}\left(\max_{1 \leq i \leq I} |Z_i| > a \log t\right) dt \\
&\leq C + \sum_{i=1}^I \int_C^\infty \mathbb{P}\left(|Z_i| > a \log t\right) dt \\
&\leq C + I \int_C^\infty 2t^{-a/M} dt.
\end{aligned}$$

Differentiating this bound with respect to $C$ gives $1 - 2IC^{-a/M}$, so the bound is minimised by $C = (2I)^{M/a}$. For $a > M$, the bound becomes

$$\begin{aligned} C + 2I\frac{M}{a-M}C^{-(a-M)/M} &= (2I)^{M/a} + \frac{M}{a-M}(2I)^{1-(a-M)/a} \\ &= \frac{a}{a-M}(2I)^{M/a}. \end{aligned}$$

Let

$$a = \frac{M\log(2I)}{\log b}$$

for $b > 1$. We have

$$\mathbb{E}\left(\exp\left(\max_{1\le i\le I}|Z_i|/a\right)\right) \le 2$$

if $b^2 2^b \le 4$, the hardest case being $I = 1$. This holds for $b = 5/4$ and the result follows. ■

We perform chaining in $L^{\psi_1}$ using the same method as Theorem 2.3.6 of Giné and Nickl (2016). Recall that $N(a, M, d)$ is the minimum size of an $a > 0$ cover of a metric space $(M, d)$.

**Lemma 4.15.2** *Let $Z$ be a stochastic process on $(\Omega, \mathcal{F})$ indexed by a separable metric space $(M, d)$ on which $Z$ is almost-surely continuous with $\|Z(s) - Z(t)\|_{\psi_1} \le d(s,t)$ for all $s, t \in M$. Let $D = \sup_{s,t\in M} d(s,t)$. Fix $s_0 \in M$. Then*

$$\left\|\sup_{s\in M}|Z(s) - Z(s_0)|\right\|_{\psi_1} \le \frac{4}{\log(5/4)}\int_0^{D/2}\log(2N(a,M,d))da.$$

**Proof** Since $(M, d)$ is separable, it has a countable dense subset $M_0$. We have

$$\left\|\sup_{s\in M}|Z(s) - Z(s_0)|\right\|_{\psi_1} = \left\|\sup_{s\in M_0}|Z(s) - Z(s_0)|\right\|_{\psi_1}$$

because $Z$ is almost-surely continuous on $M$. Since $M_0$ is countable, there exists a sequence of increasing finite subsets $F_n \subseteq M$ for $n \geq 1$ whose union is $M_0$. We have

$$\left\| \sup_{s \in M} |Z(s) - Z(s_0)| \right\|_{\psi_1} = \lim_{n \to \infty} \left\| \max_{s \in F_n} |Z(s) - Z(s_0)| \right\|_{\psi_1}$$

by the monotone convergence theorem. Fix $n \geq 1$ and let $F = F_n$. Let $\delta_j = 2^{-j} D$ for $j \geq 0$. Since $F$ is finite, there exists a minimum $J \geq 0$ such that

$$\{t \in F : d(s,t) \leq \delta_J\} = \{s\}$$

for all $s \in F$. Let $A_j$ for $0 \leq j \leq J-1$ be a $\delta_j$ cover of $(M,d)$ of size $N(\delta_j, M, d)$, where we let $A_0 = \{s_0\}$. We define the chain $C : F \times \{0, \ldots, J\} \to M$ as follows. Let $C(s,J) = s$ for all $s \in F$. For $1 \leq j \leq J$, given $C(s,j)$, let $C(s,j-1)$ be some closest point in $A_{j-1}$ to $C(s,j)$, depending on $s$ only through $C(s,j)$. We have

$$Z(s) - Z(s_0) = \sum_{j=1}^{J} Z(C(s,j)) - Z(C(s,j-1))$$

for $s \in F$. Hence,

$$\max_{s \in F} |Z(s) - Z(s_0)| \leq \sum_{j=1}^{J} \max_{s \in F} |Z(C(s,j)) - Z(C(s,j-1))|.$$

By Lemma 4.15.1, we have

$$\begin{aligned}
\left\| \max_{s \in F} |Z(s) - Z(s_0)| \right\|_{\psi_1} &\leq \sum_{j=1}^{J} \left\| \max_{s \in F} |Z(C(s,j)) - Z(C(s,j-1))| \right\|_{\psi_1} \\
&\leq \sum_{j=1}^{J} \frac{\log(2N(\delta_j, M, d))\delta_{j-1}}{\log(5/4)} \\
&= \frac{4}{\log(5/4)} \sum_{j=1}^{J} (\delta_j - \delta_{j+1}) \log(2N(\delta_j, M, d)) \\
&\leq \frac{4}{\log(5/4)} \int_{\delta_{J+1}}^{\delta_1} \log(2N(a, M, d)) da
\end{aligned}$$

$$\leq \frac{4}{\log(5/4)} \int_0^{D/2} \log(2N(a, M, d))da.$$

The result follows. ■

## 4.16 Subgaussian Random Variables and Symmetric Matrices

Recall Lemma 3.15.1, which is essentially Theorem 2.1 of Quintana and Rodríguez (2014).

**Lemma 4.16.1** *Let $M$ be a non-negative-definite matrix which is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$. There exist an orthogonal matrix $A$ and a diagonal matrix $D$ which are both $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrices on $(\Omega, \mathcal{F})$ such that $M = ADA^\mathsf{T}$.*

Recall that for $m \times n$ matrices $U$ and $V$, we define $U \circ V$ to be the $m \times n$ matrix with

$$(U \circ V)_{i,j} = U_{i,j} V_{i,j}.$$

The following lemma is a conditional version of Theorem 1.1 of Rudelson and Vershynin (2013), but with explicit values for the constants derived here.

**Lemma 4.16.2** *Let $\varepsilon_i$ for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent conditional on some sub-$\sigma$-algebra $\mathcal{G} \subseteq \mathcal{F}$ and let*

$$\mathbb{E}(\exp(t\varepsilon_i)|\mathcal{G}) \leq \exp(\sigma^2 t^2/2)$$

*almost surely for $t$ a random variable on $(\Omega, \mathcal{G})$. Let $M$ be an $n \times n$ symmetric matrix which is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{G})$. We have*

$$\mathbb{E}\left(\exp\left(t\varepsilon^{\mathsf{T}}(M - I \circ M)\varepsilon\right)\middle|\mathcal{G}\right) \le \exp\left(16\sigma^4 \operatorname{tr}(M^2)t^2\right)$$

*almost surely for $t$ a random variable on $(\Omega, \mathcal{G})$ such that $32\sigma^4 \operatorname{tr}(M^2)t^2 \le 1$.*

**Proof** We follow the proof of Theorem 1.1 of Rudelson and Vershynin (2013). Let

$$Z = \varepsilon^{\mathsf{T}}(M - I \circ M)\varepsilon = \sum_{i \ne j} M_{i,j}\varepsilon_i\varepsilon_j.$$

Also, let $\phi_i$ for $1 \le i \le n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent of each other, the $\varepsilon_i$ and $\mathcal{G}$, with $\phi_i \sim \text{Bernoulli}(1/2)$. Furthermore, let

$$W = \sum_{i \ne j} \phi_i(1 - \phi_j)M_{i,j}\varepsilon_i\varepsilon_j.$$

We have $Z = 4\,\mathbb{E}(W|\mathcal{G}, \varepsilon)$ almost surely, which gives

$$\exp(tZ) \le \mathbb{E}(\exp(4tW)|\mathcal{G}, \varepsilon)$$

almost surely for $t$ a random variable on $(\Omega, \mathcal{G})$ by Jensen's inequality. Let

$$S = \{1 \le i \le n : \phi_i = 1\}.$$

We can write

$$W = \sum_{i \in S, j \in S^{\mathsf{C}}} M_{i,j}\varepsilon_i\varepsilon_j.$$

Since the $\varepsilon_j$ are independent, we have

$$\mathbb{E}(\exp(tZ)|\mathcal{G}) \le \mathbb{E}(\exp(4tW)|\mathcal{G})$$

$$
\begin{aligned}
&= \mathbb{E}\left(\prod_{j\in S^{\mathsf{C}}} \mathbb{E}\left(\exp\left(4t\sum_{i\in S} M_{i,j}\varepsilon_i\varepsilon_j\right)\middle|\mathcal{G},\phi\right)\middle|\mathcal{G}\right)\\
&\leq \mathbb{E}\left(\prod_{j\in S^{\mathsf{C}}} \exp\left(8t^2\sigma^2\left(\sum_{i\in S} M_{i,j}\varepsilon_i\right)^2\right)\middle|\mathcal{G}\right)\\
&= \mathbb{E}\left(\exp\left(8t^2\sigma^2\sum_{j\in S^{\mathsf{C}}}\left(\sum_{i\in S} M_{i,j}\varepsilon_i\right)^2\right)\middle|\mathcal{G}\right)
\end{aligned}
$$

almost surely. Let $\delta_i$ for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent of each other, the $\varepsilon_i$, the $\phi_i$ and $\mathcal{G}$, with $\delta_i \sim \mathrm{N}(0, \sigma^2)$. Since the $\varepsilon_i$ are independent, we have

$$
\begin{aligned}
\mathbb{E}(\exp(tZ)|\mathcal{G}) &\leq \mathbb{E}\left(\exp\left(4t\sum_{j\in S^{\mathsf{C}}}\sum_{i\in S} M_{i,j}\varepsilon_i\delta_j\right)\middle|\mathcal{G}\right)\\
&= \mathbb{E}\left(\prod_{i\in S} \mathbb{E}\left(\exp\left(4t\sum_{j\in S^{\mathsf{C}}} M_{i,j}\delta_j\varepsilon_i\right)\middle|\mathcal{G},\phi\right)\middle|\mathcal{G}\right)\\
&\leq \mathbb{E}\left(\prod_{i\in S} \exp\left(8t^2\sigma^2\left(\sum_{j\in S^{\mathsf{C}}} M_{i,j}\delta_j\right)^2\right)\middle|\mathcal{G}\right)\\
&= \mathbb{E}\left(\exp\left(8t^2\sigma^2\sum_{i\in S}\left(\sum_{j\in S^{\mathsf{C}}} M_{i,j}\delta_j\right)^2\right)\middle|\mathcal{G}\right)
\end{aligned}
$$

almost surely. Let $F$ be the $n \times n$ matrix with $F_{i,j} = 1$ if $i = j \in S$ and 0 otherwise. Note that $F$ is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \sigma(\phi))$. Then

$$
\mathbb{E}(\exp(tZ)|\mathcal{G}) \leq \mathbb{E}\left(\exp\left(8t^2\sigma^2\delta^{\mathsf{T}}(I-F)MFM(I-F)\delta\right)\middle|\mathcal{G}\right)
$$

almost surely. By Lemma 4.16.1, there exist an orthogonal matrix $A$ and a diagonal matrix $D$ which are both $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrices on $(\Omega, \sigma(\mathcal{G}, \phi))$ such that

$$
(I-F)MFM(I-F) = ADA^{\mathsf{T}},
$$

which is non-negative definite. Since $A^\mathsf{T}\delta$ and $\delta$ have the same distribution given $\mathcal{G}$, we have

$$\begin{aligned}\mathbb{E}(\exp(tZ)|\mathcal{G}) &\leq \mathbb{E}\left(\exp\left(8t^2\sigma^2\delta^\mathsf{T} D\delta\right)\middle|\mathcal{G}\right)\\ &= \mathbb{E}\left(\prod_{i=1}^{n}\mathbb{E}\left(\exp\left(8t^2\sigma^2 D_{i,i}\delta_i^2\right)\middle|\mathcal{G},\phi\right)\middle|\mathcal{G}\right)\\ &= \mathbb{E}\left(\prod_{i=1}^{n}\left(1-16\sigma^4 D_{i,i}t^2\right)^{-1/2}\middle|\mathcal{G}\right)\end{aligned}$$

almost surely for $16\sigma^4(\max_{1\leq i\leq n} D_{i,i})t^2 < 1$ by computing the moment generating function of the $\delta_i^2$. We have that $(1-x)^{-1/2} \leq \exp(x)$ for $x \in [0,1/2]$, so

$$\mathbb{E}(\exp(tZ)|\mathcal{G}) \leq \mathbb{E}\left(\prod_{i=1}^{n}\exp\left(16\sigma^4 D_{i,i}t^2\right)\middle|\mathcal{G}\right) = \mathbb{E}\left(\exp\left(16\sigma^4\operatorname{tr}(D)t^2\right)\middle|\mathcal{G}\right)$$

almost surely for $32\sigma^4(\max_{1\leq i\leq n} D_{i,i})t^2 \leq 1$. We have

$$\operatorname{tr}(D) = \operatorname{tr}((I-F)MFM(I-F)) = \sum_{i\in S}\sum_{j\in S^\mathsf{C}} M_{i,j}^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n} M_{i,j}^2 = \operatorname{tr}(M^2)$$

and

$$\max_{1\leq i\leq n} D_{i,i} \leq \operatorname{tr}(D) \leq \operatorname{tr}(M^2).$$

The result follows. ■

We move the bound on the conditional moment generating function of $\varepsilon^\mathsf{T}(M - I\circ M)\varepsilon$ to that of $|\varepsilon^\mathsf{T}(M - I\circ M)\varepsilon|$.

**Lemma 4.16.3** *Let $\varepsilon_i$ for $1 \leq i \leq n$ be random variables on $(\Omega,\mathcal{F},\mathbb{P})$ which are independent conditional on some sub-$\sigma$-algebra $\mathcal{G} \subseteq \mathcal{F}$ and let*

$$\mathbb{E}(\exp(t\varepsilon_i)|\mathcal{G}) \leq \exp(\sigma^2t^2/2)$$

*almost surely for $t$ a random variable on $(\Omega, \mathcal{G})$. Let $M$ be an $n \times n$ symmetric matrix which is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{G})$. We have*

$$\mathbb{E}\left(\exp\left(t|\varepsilon^{\mathsf{T}}(M - I \circ M)\varepsilon|\right)\middle|\mathcal{G}\right) \leq \frac{1}{1 - 2^{7/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}t} 2^{2^{7/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}t}$$

*almost surely for $t \geq 0$, a random variable on $(\Omega, \mathcal{G})$, such that $2^{7/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}t < 1$. Hence,*

$$\mathbb{E}\left(\frac{\left|\varepsilon^{\mathsf{T}}(M - I \circ M)\varepsilon\right|}{2^{7/2}(\log 2)\sigma^2 \operatorname{tr}(M^2)^{1/2}/\log(5/4)}\middle|\mathcal{G}\right) \leq 2.$$

**Proof** Let

$$Z = \varepsilon^{\mathsf{T}}(M - I \circ M)\varepsilon.$$

By Lemma 4.16.2, we have

$$\mathbb{E}(\exp(tZ)|\mathcal{G}) \leq \exp\left(16\sigma^4 \operatorname{tr}(M^2)t^2\right)$$

almost surely for $t$ a random variable on $(\Omega, \mathcal{G})$ such that $32\sigma^4 \operatorname{tr}(M^2)t^2 \leq 1$. By Chernoff bounding, we have

$$\mathbb{P}(Z \geq z|\mathcal{G}) \leq \exp\left(-tz + 16\sigma^4 \operatorname{tr}(M^2)t^2\right)$$

almost surely for $z \geq 0$, a random variable on $(\Omega, \mathcal{G})$, $t \geq 0$ and $32\sigma^4 \operatorname{tr}(M^2)t^2 \leq 1$. Minimising over $t$ gives

$$\mathbb{P}(Z \geq z|\mathcal{G}) \leq \exp\left(-\min\left(\frac{z^2}{2^6\sigma^4 \operatorname{tr}(M^2)}, \frac{z}{2^{7/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}}\right)\right)$$

almost surely. The first term in the minimum is attained by $t = 2^{-5}\sigma^{-4} \operatorname{tr}(M^2)^{-1}z$ when $z < 2^{5/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}$, and the second term is attained by $t = 2^{-5/2}\sigma^{-2} \operatorname{tr}(M^2)^{-1/2}$

when $z \geq 2^{5/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}$. In the second case, note that

$$16\sigma^4 \operatorname{tr}(M^2)t^2 = \frac{1}{2} \leq \frac{z}{2^{7/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}}.$$

The same result holds if we replace $Z$ with $-Z$ by replacing $M$ with $-M$. For $C \geq 1$ and $t \geq 0$, random variables on $(\Omega, \mathcal{G})$, we have

$$\begin{aligned}\mathbb{E}(\exp(t|Z|)|\mathcal{G}) &= \int_0^\infty \mathbb{P}(|Z| \geq (\log s)/t|\mathcal{G})ds \\ &\leq C + \int_C^\infty \mathbb{P}(|Z| \geq (\log s)/t|\mathcal{G})ds \\ &\leq C + \int_C^\infty \mathbb{P}(Z \geq (\log s)/t|\mathcal{G})ds + \int_C^\infty \mathbb{P}(-Z \geq (\log s)/t|\mathcal{G})ds \\ &\leq C + 2\int_C^\infty \exp\left(-\min\left(\frac{(\log s)^2}{2^6\sigma^4 \operatorname{tr}(M^2)t^2}, \frac{\log s}{2^{7/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}t}\right)\right) ds\end{aligned}$$

almost surely. By letting $C \geq \exp(2^{5/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}t)$, the bound becomes

$$C + 2\int_C^\infty s^{-(2^{7/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}t)^{-1}} ds.$$

Let $u = 2^{7/2}\sigma^2 \operatorname{tr}(M^2)^{1/2}t$, a random variable on $(\Omega, \mathcal{G})$. Differentiating this bound with respect to $C$ gives $1 - 2C^{-u^{-1}}$, so the bound is minimised by $C = 2^u$. This satisfies the condition on $C$ above as

$$e^{2^{5/2}} \leq 3^6 \leq 2^{10} \leq 2^{2^{7/2}}.$$

For $u < 1$, the bound becomes

$$\begin{aligned}C + 2\frac{u}{1-u}C^{-(1-u)/u} &= 2^u + \frac{u}{1-u}2^{1-(1-u)} \\ &= \frac{1}{1-u}2^u.\end{aligned}$$

The first result follows. Let

$$u = \frac{\log b}{\log 2}$$

for $b > 1$. We have

$$\mathbb{E}(\exp(t|Z|)|\mathcal{G}) \leq 2$$

almost surely if $b^2 2^b \leq 4$. This holds for $b = 5/4$ and the second result follows. ■

# Chapter 5

# Extreme Points of Wasserstein Balls

There are many scenarios in which it is useful to be able to define a concept of distance between probability measures. For example, in statistics, we may be interested in investigating the effects of a perturbation of the distribution of our data. In this case, we need a concept of distance in order to understand the size of the perturbation. We could then analyse how robust an estimator is to such changes. The Wasserstein distance is a natural choice for comparing probability measures, as it is determined by a cost function on the underlying space. Hence, if two probability measures are concentrated around two points between which there is a small cost, the Wasserstein distance between the measures is small. This is in contrast to, for example, the Kullback–Leibler divergence, which is very large if the probability measures are sufficiently concentrated. An example in which this property is particularly important is when the cost function is equal to some metric on the space. The Wasserstein distance has been used extensively in statistical applications. For example, it has been used for goodness-of-fit tests (del Barrio, Cuesta-Albertos, Matrán, and Rodríguez-Rodríguez,

1999) and clustering (Irpino and Verde, 2008).

Having defined a concept of distance, we can examine the properties of a collection of nearby probability measures. The easiest way to do this is by defining a ball around a fixed measure. In our statistics example, this corresponds to all sufficiently small perturbations of the distribution of the data. We expect some of the largest variation in behaviour to occur at extreme points. For example, under certain conditions, a continuous linear functional on a convex set of probability measures attains its bounds at the extreme points of the set by the Choquet–Bishop–de Leeuw theorem (Theorem 5.6 of Bishop and de Leeuw, 1959). In our statistics example, this could be the worst-case error of an estimator. This is our motivation for finding the extreme points of Wasserstein balls.

We now discuss the Wasserstein distance in more detail. The Wasserstein distance is defined using the optimal transport problem. In this problem, the aim is to find the optimal transportation of one probability measure to another with respect to a given cost function. This is done by finding a coupling between the two probability measures which minimises the transport cost. Couplings in this context are usually referred to as transport plans. The modern analysis of optimal transport began with Kantorovitch (1958), and a recent expansive book on the subject has been written by Villani (2009). For $p \in [1, \infty)$, the Wasserstein distance is usually defined as the $p^{-1}$th power of the minimum transport cost when the cost function is equal to the $p$th power of the metric on the underlying space (see Definition 6.1 of Villani, 2009). However, in this chapter we allow weaker assumptions on the cost function. We also do not raise the minimum transport cost to a power in our definition of the Wasserstein distance, as this would simply change the radius of the Wasserstein ball. This is the same as the earliest definitions of distance between probability measures using the optimal transport problem, as in Kantorovitch (1958).

An important concept in optimal transport is the dual problem. Under mild conditions on the cost function, the dual of the optimal transport problem gives an alternative representation of the minimum transport cost as the maximum over two functions of the difference between integrals of the functions. One integral is taken with respect to the first probability measure and the other integral is taken with respect to the second probability measure. The functions are restricted by the cost function of the optimal transport problem. The dual problem was first introduced by Kantorovitch (1958), and later studied by Rüschendorf and Rachev (1990) and Rüschendorf (1995). A thorough overview is given in Chapter 5 of Villani (2009).

The study of the dual problem has greatly increased the understanding of the optimal transport problem itself. For example, such study has determined properties of the solutions to the original problem (Rüschendorf and Rachev, 1990; Rüschendorf, 1995). General properties can be found in Theorem 5.10 of Villani (2009), particularly part (ii), while Theorem 5.30 of the same book gives conditions under which the optimal transport problem is solved by a unique transport plan which is induced by a transport map. A transport plan induced by a transport map is a coupling which assigns full probability to the graph of a function. The function is referred to as the transport map. If the first probability measure obeys some regularity conditions, the conditions of Theorem 5.30 of Villani (2009) are shown to be satisfied when the cost function is equal to the squared Euclidean distance on $\mathbb{R}^n$ in Theorem 9.4 of the same book.

As far as we are aware, the only investigation into the extreme points of Wasserstein balls has been in the case in which the probability measure at the centre of the ball has finite support. For example, see Theorem 2.3 of Owhadi and Scovel (2017). However, in this chapter we allow the centre of the ball to be any probability measure. We first investigate the implications for the functions solving the dual of the optimal transport problem for points on the surface of the Wasserstein ball which are not extreme. By the surface of the ball, we mean the points in the ball whose distance from the centre of the ball is equal to the radius. We then show, under very general conditions, that the only extreme points which are not on the surface of the ball are the Dirac measures. This is followed by finding conditions under which points on the surface of the ball are extreme points or not extreme points. Finally, we consider the case in which the underlying space is finite. We use the dual problem to find further conditions under which points on the surface of the ball are not extreme points.

## 5.1 Literature Review

The first modern treatment of the optimal transport problem is given by Kantorovitch (1958). In the paper, the author introduces the problem of seeking the transport plan between two measures which minimises the transport cost. The measures are not required to be probability measures, but they must have the same total mass. The minimum transport cost is represented as a function of the two measures, introducing the earliest version of the Wasserstein distance. The main result of the paper is an early form of the duality theorem, which states that the dual problem has the same solution as the original optimal transport problem under very general conditions on the cost function. The dual problem is further studied by Rüschendorf and Rachev (1990) and Rüschendorf (1995). Both papers use ideas from convex analysis and examine the implications for the solutions to the original optimal transport problem.

A recent book on the subject of optimal transport has been written by Villani (2009). The book covers all major developments in the field. This chapter is particularly concerned with the following aspects. Chapter 5 covers duality, with a detailed version of the duality theorem given by Theorem 5.10. Chapter 6 examines the Wasserstein distance for general $p \in [1, \infty)$. Theorem 5.30 gives conditions under which the optimal transport problem is solved by a unique transport plan which is induced by a transport map. Theorem 9.4 then proves that these conditions are satisfied when the cost function is equal to the squared Euclidean distance on $\mathbb{R}^n$, subject to some regularity conditions on the first probability measure.

The most recent result on the extreme points of Wasserstein balls when the probability measure at the centre of the ball has finite support is Theorem 2.4 of Owhadi and Scovel (2017). The theorem states that if the probability measure at the centre of the ball has finite support of size $n$, then the extreme points of the ball have finite support of size at most $n + 2$. The paper also provides the same result for balls of probability measures defined by distances other than the Wasserstein distance.

## 5.2 Contribution

In this chapter, we give various conditions under which probability measures in a Wasserstein ball are extreme points or not extreme points. As far as we are aware, the classification of the extreme points of Wasserstein balls has previously only been studied in the case in which the probability measure at the centre of the ball has finite support (Theorem 2.3 of Owhadi and Scovel, 2017). Our results do not make this restriction. In Section 5.4, we examine the functions solving the dual problem for points on the surface of the Wasserstein ball which are not extreme points. In Section 5.5, we show that, under very general conditions, the only extreme points which are not on the surface of the ball are the Dirac measures (Lemma 5.5.3 on page 196).

We then move on to the surface of the ball in Section 5.6. We show that if the Wasserstein distance is uniquely attained by a transport plan induced by a transport map, then we have an extreme point (Lemma 5.6.1 on page 198). Conversely, under conditions on the centre of the ball and the cost function, we show that if the Wasserstein distance is attained by two distinct transport plans induced by continuous transport maps, then we do not have an extreme point (Lemma 5.6.3 on page 201).

Finally, in Section 5.7, we consider the case in which our probability measures are defined on finite sets. We examine how the results of Section 5.6 can be applied in this setting. We also make use of the variables which solve the dual problem to prove the following results. We show that if an optimal transport plan transports mass from one atom to two atoms at the same unit cost, then we do not have an extreme point (Lemma 5.7.2 on page 215). We then show a similar result in which the mass may be transported from two different atoms under conditions on the optimal dual variables (Lemma 5.7.3 on page 216).

## 5.3 Optimal Transport

Let $(X, d_X)$ and $(Y, d_Y)$ be complete separable metric spaces. Furthermore, let $\mathcal{B}(X)$, $\mathcal{B}(Y)$ and $\mathcal{B}(X \times Y)$ be the set of Borel sets on $X$, $Y$ and $X \times Y$, and let $\mathcal{P}(X)$, $\mathcal{P}(Y)$ and $\mathcal{P}(X \times Y)$ be the set of Borel probability measures on $X$, $Y$ and $X \times Y$. We consider the problem of optimally transporting a probability measure $P \in \mathcal{P}(X)$ to $Q \in \mathcal{P}(Y)$ with respect to some Borel cost function $c : X \times Y \to [0, \infty)$.

In order to study this problem, we define the marginals of $\gamma \in \mathcal{P}(X \times Y)$. Let $\pi_1 : \mathcal{P}(X \times Y) \to \mathcal{P}(X)$ by $(\pi_1\gamma)(A) = \gamma(A \times Y)$ for all $A \in \mathcal{B}(X)$ and let $\pi_2 : \mathcal{P}(X \times Y) \to \mathcal{P}(Y)$ by $(\pi_2\gamma)(B) = \gamma(X \times B)$ for all $B \in \mathcal{B}(Y)$. The marginals of $\gamma \in \mathcal{P}(X \times Y)$ are $\pi_1\gamma \in \mathcal{P}(X)$ and $\pi_2\gamma \in \mathcal{P}(Y)$. We define

$$\Pi(P, Q) = \{\gamma \in \mathcal{P}(X \times Y) : \pi_1\gamma = P \text{ and } \pi_2\gamma = Q\}$$

for $P \in \mathcal{P}(X)$ and $Q \in \mathcal{P}(Y)$. We refer to $\gamma \in \Pi(P,Q)$ as a transport plan for our optimal transport problem. The problem itself is then to find

$$\inf_{\gamma \in \Pi(P,Q)} \int c \, d\gamma.$$

In particular, we hope that this infimum is attained by some $\gamma \in \Pi(P,Q)$. Such a $\gamma$ is referred to as an optimal transport plan. If we assume

$$(c1) \qquad c \text{ is lower semicontinuous,}$$

then an optimal transport plan exists by Theorem 4.1 of Villani (2009).

The optimal transport problem above is parametrised by $P \in \mathcal{P}(X)$ and $Q \in \mathcal{P}(Y)$. For each parametrisation, the problem has a minimum transport cost which we refer to as the Wasserstein distance

$$W_c(P,Q) = \inf \left\{ \int c \, d\gamma : \gamma \in \Pi(P,Q) \right\}.$$

This infimum is attained if we assume $(c1)$. In general, $W_c$ has very few of the properties that we associate with a distance. However, $W_c(P,Q)$ does in some sense measure how far apart $P$ and $Q$ are. We have the following convexity result.

**Lemma 5.3.1** *Let $P_1, P_2 \in \mathcal{P}(X)$, $Q_1, Q_2 \in \mathcal{P}(Y)$ and $t \in (0,1)$. Then*

$$W_c(tP_1 + (1-t)P_2, tQ_1 + (1-t)Q_2) \leq tW_c(P_1,Q_1) + (1-t)W_c(P_2,Q_2).$$

**Proof** If $\gamma_1 \in \Pi(P_1,Q_1)$ and $\gamma_2 \in \Pi(P_2,Q_2)$, then

$$t\gamma_1 + (1-t)\gamma_2 \in \Pi(tP_1 + (1-t)P_2, tQ_1 + (1-t)Q_2).$$

Hence,

$$\begin{aligned}
&W_c(tP_1 + (1-t)P_2, tQ_1 + (1-t)Q_2) \\
&= \inf\left\{\int c\, d\gamma : \gamma \in \Pi(tP_1 + (1-t)P_2, tQ_1 + (1-t)Q_2)\right\} \\
&\leq \inf\left\{\int c\, d(t\gamma_1 + (1-t)\gamma_2) : \gamma_1 \in \Pi(P_1, Q_1), \gamma_2 \in \Pi(P_2, Q_2)\right\} \\
&= \inf\left\{t\int c\, d\gamma_1 + (1-t)\int c\, d\gamma_2 : \gamma_1 \in \Pi(P_1, Q_1), \gamma_2 \in \Pi(P_2, Q_2)\right\} \\
&= \inf\left\{t\int c\, d\gamma_1 : \gamma_1 \in \Pi(P_1, Q_1)\right\} + \inf\left\{(1-t)\int c\, d\gamma_2 : \gamma_2 \in \Pi(P_2, Q_2)\right\} \\
&= tW_c(P_1, Q_1) + (1-t)W_c(P_2, Q_2).
\end{aligned}$$

■

We now define the closed Wasserstein ball

$$B_c[P, r] = \{Q \in \mathcal{P}(Y) : W_c(P, Q) \leq r\}$$

for $P \in \mathcal{P}(X)$ and $r \geq 0$. By Lemma 5.3.1 with $P_1 = P_2 = P$, we know that $B[P, r]$ is convex. We call $Q \in B_c[P, r]$ an extreme point of $B_c[P, r]$ if $Q = tQ_1 + (1-t)Q_2$ for $Q_1, Q_2 \in B_c[P, r]$ and $t \in (0, 1)$ implies $Q_1 = Q_2$, in which case $Q_1 = Q_2 = Q$. We denote the set of extreme points of $B_c[P, r]$ by $\text{ext}(B_c[P, r])$.

Some transport plans transport probability measures by mapping each point $x \in X$ to a point $y \in Y$. A transport map $T : X \to Y$ is a Borel function such that $P(T^{-1}(B)) = Q(B)$ for all $B \in \mathcal{B}(Y)$. The transport plan $\gamma \in \Pi(P, Q)$ induced by $T$ is $\gamma(C) = P(\{x \in X : (x, T(x)) \in C\})$ for $C \in \mathcal{B}(X \times Y)$. We have $\{x \in X : (x, T(x)) \in C\} \in \mathcal{B}(X)$ because the function $f : X \to X \times Y$ by $f(x) = (x, T(x))$ is Borel. Note that $\gamma(A \times B) = P(A \cap T^{-1}(B))$ for $A \in \mathcal{B}(X)$ and $B \in \mathcal{B}(Y)$. The graph $G = \{(x, y) \in X \times Y : T(x) = y\}$ of $T$ has $\gamma(G) = 1$, where $G \in \mathcal{B}(X \times Y)$ because $G = \{(x, y) \in X \times Y : d_Y(T(x), y) = 0\}$ and $f : X \times Y \to [0, \infty)$ by

$f(x, y) = d(T(x), y)$ is Borel. Therefore, if $f : X \times Y \to \mathbb{R}$ is Borel and either $\gamma$-integrable or non-negative, then

$$\int f \, d\gamma = \int f(x, T(x)) \, dP(x).$$

Before going further, we define the balls

$$\begin{aligned} B_X(x, \varepsilon) &= \{z \in X : d_X(z, x) < \varepsilon\}, \\ B_X[x, \varepsilon] &= \{z \in X : d_X(z, x) \leq \varepsilon\}, \\ B_Y(y, \varepsilon) &= \{w \in Y : d_Y(w, y) < \varepsilon\}, \\ B_Y[y, \varepsilon] &= \{w \in Y : d_Y(w, y) \leq \varepsilon\} \end{aligned}$$

for $x \in X$, $y \in Y$ and $\varepsilon \geq 0$. Note that for $\varepsilon = 0$, we have $B_X(x, 0) = \varnothing$, $B_Y(y, 0) = \varnothing$ and $B_X[x, 0] = \{x\}$, $B_Y[y, 0] = \{y\}$.

## 5.4 Dual Functions

The optimal transport problem has the dual problem

$$\sup_{\psi \in L^1(P), \phi \in L^1(Q)} \left\{ \int \phi \, dQ - \int \psi \, dP : \phi(y) - \psi(x) \leq c(x, y) \text{ for all } (x, y) \in X \times Y \right\}.$$

For $\psi \in L^1(P)$ and $\phi \in L^1(Q)$ such that $\phi(y) - \psi(x) \leq c(x, y)$ for all $(x, y) \in X \times Y$, we have

$$\begin{aligned} \int \phi \, dQ - \int \psi \, dP &= \int (\phi(y) - \psi(x)) \, d\gamma(x, y) \\ &\leq \int c(x, y) \, d\gamma(x, y) \end{aligned}$$

for all $\gamma \in \Pi(P, Q)$. By taking an infimum over $\gamma \in \Pi(P, Q)$, we find that

$$\int \phi \, dQ - \int \psi \, dP \leq W_c(P, Q).$$

Hence, the maximum value of the dual problem is always at most the minimum transport cost. We refer to $\psi$ and $\phi$ as dual functions. If we assume ($c1$) that $c$ is lower semicontinuous, then the two problems have the same optimum values by Theorem 5.10 of Villani (2009). If we assume both ($c1$) and

($c2$) $c(x, y) \leq c_X(x) + c_Y(y)$ for all $(x, y) \in X \times Y$ for some $c_X \in L^1(P)$ and $c_Y \in L^1(Q)$,

then the supremum in the dual problem is attained by some $\psi \in L^1(P)$ and $\phi \in L^1(Q)$, again by Theorem 5.10 of Villani (2009). Such functions $\psi$ and $\phi$ are referred to as optimal dual functions.

We now investigate properties of the optimal dual functions for $Q \in B_c[P, r]$ on the surface of the ball which are not extreme points. We first show that if $Q \in B_c[P, r]$ with $W_c(P, Q) = r$ is a convex combination of $Q_1, Q_2 \in B_c[P, r]$, then $W_c(P, Q_1) = r$ and $W_c(P, Q_2) = r$ are attained by the same dual functions as $W_c(P, Q)$.

**Lemma 5.4.1** *Assume (c1). Let $Q \in B_c[P, r]$ with $W_c(P, Q) = r$ and suppose that $Q = tQ_1 + (1-t)Q_2$ for $Q_1, Q_2 \in B_c[P, r]$ and $t \in (0, 1)$. Let $\psi \in L^1(P)$ and $\phi \in L^1(Q)$ satisfy $\phi(y) - \psi(x) \leq c(x, y)$ for all $(x, y) \in X \times Y$ and*

$$W_c(P, Q) = \int \phi \, dQ - \int \psi \, dP.$$

*Then $W_c(P, Q_1) = r$ and $W_c(P, Q_2) = r$. Furthermore,*

$$W_c(P, Q_1) = \int \phi \, dQ_1 - \int \psi \, dP \quad \text{and} \quad W_c(P, Q_2) = \int \phi \, dQ_2 - \int \psi \, dP.$$

**Proof** By Lemma 5.3.1 with $P_1 = P_2 = P$, we have

$$tW_c(P, Q_1) + (1 - t)W_c(P, Q_2) \geq r.$$

Since $Q_1, Q_2 \in B_c[P, r]$, we have $W_c(P, Q_1) = r$ and $W_c(P, Q_2) = r$. Therefore, we have

$$\int \phi \, dQ_1 - \int \psi \, dP \leq r \text{ and } \int \phi \, dQ_2 - \int \psi \, dP \leq r$$

from the dual problems. Furthermore, we have

$$t\left(\int \phi \, dQ_1 - \int \psi \, dP\right) + (1-t)\left(\int \phi \, dQ_2 - \int \psi \, dP\right) = \int \phi \, dQ - \int \psi \, dP = r.$$

Hence,

$$\int \phi \, dQ_1 - \int \psi \, dP = r \text{ and } \int \phi \, dQ_2 - \int \psi \, dP = r.$$

The result follows. ■

We also show that if $Q_1, Q_2 \in B_c[P, r]$ with $W_c(P, Q_1) = r$ and $W_c(P, Q_2) = r$, and $W_c(P, Q_1)$ and $W_c(P, Q_2)$ are attained by the same dual functions, then $Q = tQ_1 + (1 - t)Q_2 \in B_c[P, r]$ has $W_c(P, Q) = r$ for all $t \in (0, 1)$, and is attained by the same dual functions as $W_c(P, Q_1)$ and $W_c(P, Q_2)$.

**Lemma 5.4.2** *Assume (c1). Let $Q_1, Q_2 \in B_c[P, r]$ with $W_c(P, Q_1) = r$ and $W_c(P, Q_2) = r$. Let $\psi \in L^1(P)$ and $\phi \in L^1(Q)$ satisfy $\phi(y) - \psi(x) \leq c(x, y)$ for all $(x, y) \in X \times Y$ and*

$$W_c(P, Q_1) = \int \phi \, dQ_1 - \int \psi \, dP \quad \textit{and} \quad W_c(P, Q_2) = \int \phi \, dQ_2 - \int \psi \, dP.$$

*Then $Q = tQ_1 + (1 - t)Q_2 \in B_c[P, r]$ has $W_c(P, Q) = r$ and*

$$W_c(P, Q) = \int \phi \, dQ - \int \psi \, dP$$

*for all $t \in (0,1)$.*

**Proof** By Lemma 5.3.1 with $P_1 = P_2 = P$, we have $Q \in B_c[P,r]$. Furthermore, from the dual problem we have

$$\begin{aligned} W_c(P,Q) &\geq \int \phi \, dQ - \int \psi \, dP \\ &= \int \phi \, d(tQ_1 + (1-t)Q_2) - \int \psi \, dP \\ &= t\left(\int \phi \, dQ_1 - \int \psi \, dP\right) + (1-t)\left(\int \phi \, dQ_2 - \int \psi \, dP\right) \\ &= tr + (1-t)r \\ &= r. \end{aligned}$$

The inequality must hold with equality and the result follows. ■

## 5.5 Inside the Ball

In this section, we consider probability measures which lie inside the Wasserstein ball as opposed to being on the surface of the ball. Let $\delta_y$ be the Dirac measure at $y \in Y$. Dirac measures which lie within $B_c[P,r]$ are extreme points of $B_c[P,r]$.

**Lemma 5.5.1** *Suppose that $Q = \delta_y \in B_c[P,r]$ for some $y \in Y$. Then $Q \in \mathrm{ext}(B_c[P,r])$.*

**Proof** Let $Q = tQ_1 + (1-t)Q_2$ for $Q_1, Q_2 \in B_c[P,r]$ and $t \in (0,1)$. Suppose $Q_1(\{y\}) < 1$. Then

$$Q(\{y\}) = tQ_1(\{y\}) + (1-t)Q_2(\{y\}) < 1.$$

This is a contradiction, so $Q_1(\{y\}) = 1$. Similarly, $Q_2(\{y\}) = 1$. Therefore, $Q_1 = Q_2$. The result follows. ■

In fact, the only extreme points $Q$ of $B_c[P, r]$ with $W_c(P, Q) < r$ are the Dirac measures. Before showing this, we need the following result.

**Lemma 5.5.2** *Let $Q \in \mathcal{P}(Y)$ and suppose that there is no $y \in Y$ such that $Q = \delta_y$. Then there exists $A \in \mathcal{B}(Y)$ such that $Q(A) \in (0, 1)$.*

**Proof** Since $Y$ is separable, it has a countable dense subset $Y_0$. For all $n \geq 1$, we have

$$Y = \bigcup_{y \in Y_0} B_Y(y, 1/n).$$

Hence,

$$\sum_{y \in Y_0} Q(B_Y(y, 1/n)) \geq 1.$$

Suppose that $Q(A) \in \{0, 1\}$ for all $A \in \mathcal{B}(Y)$. Then there exists $y_n \in Y_0$ such that $Q(B_Y(y_n, 1/n)) = 1$. Let

$$A = \bigcap_{n \geq 1} B_Y(y_n, 1/n).$$

Then $Q(A) = 1$, so $A \neq \varnothing$. Let $y \in A$. Then, by the definition of $A$, $y_n \to y$ as $n \to \infty$. Since $Y$ is a metric space, limits of sequences in $Y$ are unique when they exist. Hence, $A$ is a singleton and $Q$ is a Dirac measure. This is a contradiction and the result follows. ■

Before continuing, we define the measures

$$P|_{A_1}(A_2) = P(A_1 \cap A_2),$$

$$Q|_{B_1}(B_2) = Q(B_1 \cap B_2),$$

$$\gamma|_{C_1}(C_2) = \gamma(C_1 \cap C_2)$$

for $P \in \mathcal{P}(X)$, $Q \in \mathcal{P}(Y)$, $\gamma \in \mathcal{P}(X \times Y)$ and $A_1, A_2 \in \mathcal{B}(X)$, $B_1, B_2 \in \mathcal{B}(Y)$, $C_1, C_2 \in \mathcal{B}(X \times Y)$. Note that $P|_{A_1}$, $Q|_{B_1}$ and $\gamma_{C_1}$ are not in general probability measures. For $P \in \mathcal{P}(X)$ and $Q \in \mathcal{P}(Y)$, we also define the product measure $P \otimes Q \in \mathcal{P}(X \times Y)$ by $(P \otimes Q)(A \times B) = P(A)Q(B)$ for $A \in \mathcal{B}(X)$ and $B \in \mathcal{B}(Y)$. We know that $P \otimes Q$ extends to a unique probability measure on $(X \times Y, \mathcal{B}(X \times Y))$ by the Hahn-Kolmogorov theorem.

Now that we have these definitions, we can prove the following result. Under mild conditions, any $Q \in B_c[P, r]$ such that $W_c(P, Q) < r$ which is not a Dirac measure is not an extreme point of $B_c[P, r]$.

**Lemma 5.5.3** *Assume (c1). Let $Q \in B_c[P, r]$ with $W_c(P, Q) < r$ and suppose that there is no $y \in Y$ such that $Q = \delta_y$. Assume*

$$\int c \, d(P \otimes Q) < \infty.$$

*Then $Q \notin \operatorname{ext}(B_c[P, r])$.*

**Proof** Let $A \in \mathcal{B}(Y)$ such that $Q(A) \in (0, 1)$. Such an $A$ exists by Lemma 5.5.2. Define

$$Q_1 = (1 - \varepsilon Q(A))Q + \varepsilon Q|_A$$

and

$$Q_2 = (1 - \varepsilon Q(A^{\mathsf{C}}))Q + \varepsilon Q|_{A^{\mathsf{C}}}.$$

for $\varepsilon \in (0, 1)$. Note that $Q_1, Q_2 \in \mathcal{P}(Y)$ and $Q = Q_1/2 + Q_2/2$. We have

$$Q_1(A) = Q(A) + \varepsilon Q(A)(1 - Q(A)) > Q(A) - \varepsilon Q(A)(1 - Q(A)) = Q_2(A),$$

so $Q_1 \neq Q_2$. Let $\gamma \in \Pi(P,Q)$ attain $W_c(P,Q)$ and

$$\gamma_1 = (1 - \varepsilon Q(A))\gamma + \varepsilon(P \otimes Q|_A)$$

and

$$\gamma_2 = (1 - \varepsilon Q(A^{\mathsf{C}}))\gamma + \varepsilon(P \otimes Q|_A^{\mathsf{C}}).$$

Then $\gamma_1 \in \Pi(P,Q_1)$ and $\gamma_2 \in \Pi(P,Q_2)$. Futhermore,

$$\begin{aligned}\int c\, d\gamma_1 &= (1 - \varepsilon Q(A)) \int c\, d\gamma + \varepsilon \int c\, d(P \otimes Q|_A) \\ &\leq W_c(P,Q) + \varepsilon \int c\, d(P \otimes Q)\end{aligned}$$

and

$$\begin{aligned}\int c\, d\gamma_2 &= (1 - \varepsilon Q(A^{\mathsf{C}})) \int c\, d\gamma + \varepsilon \int c\, d(P \otimes Q|_A^{\mathsf{C}}) \\ &\leq W_c(P,Q) + \varepsilon \int c\, d(P \otimes Q).\end{aligned}$$

For $\varepsilon$ sufficiently small, we have

$$W_c(P,Q) + \varepsilon \int c\, d(P \otimes Q) \leq r.$$

Hence, $Q_1, Q_2 \in B_c[P,r]$. The result follows. ■

## 5.6 Surface of the Ball

We now consider probability measures on the surface of the Wasserstein ball. Any $Q \in B_c[P,r]$ with $W_c(P,Q) = r$ such that $W_c(P,Q)$ is uniquely attained by a transport plan induced by a transport map is an extreme point of $B_c[P, r]$.

**Lemma 5.6.1** *Assume (c1). Let $Q \in B_c[P, r]$ with $W_c(P, Q) = r$. Suppose that $W_c(P, Q)$ is uniquely attained by the transport plan $\gamma \in \Pi(P, Q)$ induced by the transport map $T$. Then $Q \in \mathrm{ext}(B_c[P, r])$.*

**Proof** Let $Q = tQ_1 + (1-t)Q_2$ for $Q_1, Q_2 \in B_c[P, r]$ and $t \in (0, 1)$. Let $W_c(P, Q_1)$ and $W_c(P, Q_2)$ be attained by $\gamma_1 \in \Pi(P, Q_1)$ and $\gamma_2 \in \Pi(P, Q_2)$. We have $t\gamma_1 + (1-t)\gamma_2 \in \Pi(P, Q)$, so

$$\int c \; d(t\gamma_1 + (1-t)\gamma_2) \geq r$$

by the definition of $W_c(P, Q)$. On the other hand,

$$\int c \; d(t\gamma_1 + (1-t)\gamma_2) = t \int c \; d\gamma_1 + (1-t) \int c \; d\gamma_2 \leq r$$

because $W_c(P, Q_1) \leq r$ and $W_c(P, Q_2) \leq r$. Hence,

$$\int c \; d(t\gamma_1 + (1-t)\gamma_2) = r.$$

Since $W_c(P, Q)$ is uniquely attained by $\gamma$, we find $\gamma = t\gamma_1 + (1-t)\gamma_2$. Let $G = \{(x, y) \in X \times Y : T(x) = y\}$ the graph of $T$. Note that $G \in \mathcal{B}(X \times Y)$ because $G = \{(x, y) \in X \times Y : d_Y(T(x), y) = 0\}$ and $f : X \times Y \to [0, \infty)$ by $f(x, y) = d(T(x), y)$ is Borel. Since $\gamma(G) = 1$, we have $\gamma_1(G) = 1$ and $\gamma_2(G) = 1$. Hence,

$$\begin{aligned} Q_1(B) &= \gamma_1(X \times B) \\ &= \gamma_1((X \times B) \cap G) \\ &= \gamma_1((T^{-1}(B) \times Y) \cap G) \\ &= \gamma_1(T^{-1}(B) \times Y) \\ &= P(T^{-1}(B)) \end{aligned}$$

$$= Q(B)$$

for $B \in \mathcal{B}(Y)$, so $Q_1 = Q$. Similarly, $Q_2 = Q$. The result follows. ■

Theorem 9.4 of Villani (2009) shows that the conditions of Lemma 5.6.1 are satisfied when $X = Y = \mathbb{R}^n$ for some $n \geq 1$, $c(x, y) = \|x - y\|_2^2$, Borel sets of dimension $n - 1$ are $P$-null and

$$\int \|x\|_2^2 \, dP(x) < \infty \text{ and } \int \|y\|_2^2 \, dQ(y) < \infty.$$

In particular, under the conditions on $X$, $Y$, $c$ and $P$, all $Q \in B_c[P, r]$ with $W_c(P, Q) = r$ are extreme points of $B_c[P, r]$. Hence, in this setting, we have exactly characterised the extreme points of $B_c[P, r]$. We have

$$\text{ext}(B_c[P, r]) = \{\delta_y \in \mathcal{P}(Y) : W_c(P, \delta_y) < r\} \cup \{Q \in \mathcal{P}(Y) : W_c(P, Q) = r\}.$$

Note that Theorem 9.4 of Villani (2009) is very specific to the cost function $c(x, y) = \|x - y\|_2^2$. We continue by exploring situations in which $Q \in B_c[P, r]$ with $W_c(P, Q) = r$ is not an extreme point of $B_c[P, r]$.

**Lemma 5.6.2** *Let $Q \in B_c[P, r]$ with $W_c(P, Q) = r$. Suppose that $W_c(P, Q)$ is attained by both $\gamma_1, \gamma_2 \in \Pi(P, Q)$ with $\gamma_1 \neq \gamma_2$. If there exist $A \in \mathcal{B}(X)$ and $B \in \mathcal{B}(Y)$ such that $\gamma_1(A \times B) \neq \gamma_2(A \times B)$ and*

$$\int c \, \mathbb{1}(A \times Y) \, d\gamma_1 = \int c \, \mathbb{1}(A \times Y) \, d\gamma_2,$$

*then $Q \notin \text{ext}(B_c[P, r])$.*

**Proof** Let

$$Q_1(C) = \gamma_1(A \times C) + \gamma_2(A^{\mathsf{C}} \times C)$$

and

$$Q_2(C) = \gamma_1(A^{\mathsf{C}} \times C) + \gamma_2(A \times C)$$

for $C \in \mathcal{B}(Y)$. Note that $Q_1, Q_2 \in \mathcal{P}(Y)$ and $Q = Q_1/2 + Q_2/2$. We have

$$Q_1(B) = Q(B) + \gamma_1(A \times B) - \gamma_2(A \times B) \neq Q(B) + \gamma_2(A \times B) - \gamma_1(A \times B) = Q_2(B),$$

so $Q_1 \neq Q_2$. Let

$$\gamma_{A,1} = \gamma_1|_{A \times Y} + \gamma_2|_{A^{\mathsf{C}} \times Y}$$

and

$$\gamma_{A,2} = \gamma_1|_{A^{\mathsf{C}} \times Y} + \gamma_2|_{A \times Y}.$$

Then $\gamma_{A,1} \in \Pi(P, Q_1)$ and $\gamma_{A,2} \in \Pi(P, Q_2)$. Furthermore,

$$\begin{aligned}\int c \, d\gamma_{A,1} &= \int c \, \mathbb{1}(A \times Y) \, d\gamma_1 + \int c \, \mathbb{1}(A^{\mathsf{C}} \times Y) \, d\gamma_2 \\ &= \int c \, \mathbb{1}(A \times Y) \, d\gamma_2 + \int c \, \mathbb{1}(A^{\mathsf{C}} \times Y) \, d\gamma_2 \\ &= \int c \, d\gamma_2,\end{aligned}$$

so $Q_1 \in B_c[P, r]$. Similarly, $Q_2 \in B_c[P, r]$. The result follows. ■

Note that there are always $A \in \mathcal{B}(X)$ and $B \in \mathcal{B}(Y)$ such that $\gamma_1(A \times B) \neq \gamma_2(A \times B)$, as otherwise $\gamma_1 = \gamma_2$ by Dynkin's lemma. However, $A$ and $B$ might not satisfy the condition with respect to the cost function. The result in its general form above is difficult to apply directly, however we continue by considering a specific context in which the transport plans are attained by continuous transport maps and the cost function is continuous, along with some conditions on $P$.

Before considering this context, we need to define a new concept. We call $P \in \mathcal{P}(X)$ ball-respecting if $P(B_X(x, \varepsilon)) = P(B_X[x, \varepsilon])$ for all $x \in X$ and $\varepsilon \geq 0$. In particular, this implies $P(\{x\}) = 0$ for $\varepsilon = 0$. This is similar to the notion of $P$ being inner-regular, as we can find $P(B_X[x, \varepsilon])$ from $P(B_X(x, \varepsilon))$, where $B_X(x, \varepsilon) \subseteq B_X[x, \varepsilon]$. We have the following result.

**Lemma 5.6.3** *Let $X$ be connected,* $\operatorname{supp} P = X$ *and $P$ respect balls. Let $Q \in B_c[P, r]$ with $W_c(P, Q) = r$. Suppose that $W_c(P, Q)$ is attained by both $\gamma_1, \gamma_2 \in \Pi(P, Q)$. Suppose further that the transport plans $\gamma_1$, $\gamma_2$ are induced by the transport maps $T_1$, $T_2$ and that $T_1$ and $T_2$ are continuous. Finally, suppose that $c$ is continuous and that there exists $x \in X$ such that $c(x, T_1(x)) \neq c(x, T_2(x))$. Then $T_1 \neq T_2$ and $\gamma_1 \neq \gamma_2$, and $Q \notin \operatorname{ext}(B_c[P, r])$.*

**Proof** We have $T_1 \neq T_2$ because $c(x, T_1(x)) \neq c(x, T_2(x))$. Let

$$\nu_1(A) = \frac{1}{r} \int \mathbb{1}(x \in A) c(x, T_1(x)) \, dP(x)$$

and

$$\nu_2(A) = \frac{1}{r} \int \mathbb{1}(x \in A) c(x, T_2(x)) \, dP(x)$$

for all $A \in \mathcal{B}(X)$. Note that $\nu_1, \nu_2 \in \mathcal{P}(X)$ and there is a version of the Radon–Nikodym derivative

$$\frac{d\nu_1}{dP}(x) = \frac{c(x, T_1(x))}{r} \quad \text{and} \quad \frac{d\nu_2}{dP}(x) = \frac{c(x, T_2(x))}{r}.$$

Let

$$\begin{aligned} S_1 &= \{x \in X : c(x, T_1(x)) > c(x, T_2(x))\}, \\ S_2 &= \{x \in X : c(x, T_1(x)) < c(x, T_2(x))\}, \\ S_3 &= \{x \in X : c(x, T_1(x)) = c(x, T_2(x))\}. \end{aligned}$$

Note that $\{S_1, S_2, S_3\}$ is a partition of $X$. We have $S_1, S_2$ open and $S_3$ closed by the continuity of $T_1$, $T_2$ and $c$. Since we have $x \in X$ such that $c(x, T_1(x)) \neq c(x, T_2(x))$, we have that $S_1 \cup S_2 \neq \varnothing$. If $S_1 \neq \varnothing$ then, since $S_1$ is open, there exists $x \in S_1$ and $\varepsilon > 0$ such that $B_X(x, \varepsilon) \subseteq S_1$. Suppose $S_2 = \varnothing$. Since $\operatorname{supp} P = X$, we have

$$1 = \int \mathbb{1}(S_1)\frac{d\nu_1}{dP}\,dP + \int \mathbb{1}(S_3)\frac{d\nu_1}{dP}\,dP > \int \mathbb{1}(S_1)\frac{d\nu_2}{dP}\,dP + \int \mathbb{1}(S_3)\frac{d\nu_2}{dP}\,dP = 1.$$

This is a contradiction, so $S_2 \neq \varnothing$. Similarly, if $S_2 \neq \varnothing$ then $S_1 \neq \varnothing$. Since $S_1 \cup S_2 \neq \varnothing$, we have $S_1 \neq \varnothing$ and $S_2 \neq \varnothing$. Hence, $S_3 \neq \varnothing$ by the connectedness of $X$. Suppose that both $T_1$ and $T_2$ are constant on $S_1$. Then $S_1$ is closed by the continuity of $T_1$, $T_2$ and $c$, so $S_1$ is a non-trivial clopen set. This contradicts $X$ being connected, so either $T_1$ or $T_2$ is non-constant on $S_1$. Similarly, either $T_1$ or $T_2$ is non-constant on $S_2$.

Without loss of generality, let $T_1$ be non-constant on $S_i$ for some $i \in \{1, 2\}$ and let $j = 3 - i \in \{1, 2\}$ with $j \neq i$. Then there exist $x_i \in S_i$ and $x_j \in S_j$ such that $T_1(x_i) \neq T_2(x_j)$. Furthermore, $T_1(x_i) \neq T_2(x_i)$ because $x_i \in S_i$. Let

$$\varepsilon = \min(d_Y(T_1(x_i), T_2(x_i)), d_Y(T_1(x_i), T_2(x_j))).$$

Note that $\varepsilon > 0$. By the openness of $S_i$ and $S_j$ and the continuity of $T_1$ and $T_2$, there exists $\delta > 0$ such that

$$\begin{aligned}
B_X(x_i, \delta) &\subseteq S_i, \\
B_X(x_j, \delta) &\subseteq S_j, \\
T_1(B_X(x_i, \delta)) &\subseteq B_Y(T_1(x_i), \varepsilon/2), \\
T_2(B_X(x_i, \delta)) &\subseteq B_Y(T_2(x_i), \varepsilon/2), \\
T_2(B_X(x_j, \delta)) &\subseteq B_Y(T_2(x_j), \varepsilon/2).
\end{aligned}$$

By the definition of $\varepsilon$,

$$B_Y(T_1(x_i), \varepsilon/2) \cap B_Y(T_2(x_i), \varepsilon/2) = \varnothing,$$
$$B_Y(T_1(x_i), \varepsilon/2) \cap B_Y(T_2(x_j), \varepsilon/2) = \varnothing.$$

Let

$$\alpha_i = \nu_i(B_X(x_i, \delta)) - \nu_j(B_X(x_i, \delta)),$$
$$\alpha_j = \nu_j(B_X(x_j, \delta)) - \nu_i(B_X(x_j, \delta)).$$

Since $B_X(x_i, \delta) \subseteq S_i$, $B_X(x_j, \delta) \subseteq S_j$ and $\operatorname{supp} P = X$, we find

$$\alpha_i = \int \mathbb{1}(B_X(x_i, \delta)) \left( \frac{d\nu_i}{dP} - \frac{d\nu_j}{dP} \right) dP > 0,$$
$$\alpha_j = \int \mathbb{1}(B_X(x_j, \delta)) \left( \frac{d\nu_j}{dP} - \frac{d\nu_i}{dP} \right) dP > 0.$$

If $\alpha_i = \alpha_j$, then let $\delta_i = \delta_j = \delta$. Otherwise, we have $k \in \{1, 2\}$ and $l = 3 - k \in \{1, 2\}$ with $l \neq k$ such that $\alpha_k > \alpha_l$. Let $f : [0, \delta] \to \mathbb{R}$ by

$$f(\eta) = \nu_k(B_X(x_k, \eta)) - \nu_l(B_X(x_k, \eta)).$$

We have $f(0) = 0$ and $f(\delta) = \alpha_k$. Furthermore, we have $B_X(x_k, \eta) \subseteq B_X(x_k, \delta) \subseteq S_k$ and

$$f(\eta) = \int \mathbb{1}(B_X(x_k, \eta)) \left( \frac{d\nu_k}{dP} - \frac{d\nu_l}{dP} \right) dP.$$

Since $P$ respects balls, we have that $f$ is continuous. By the intermediate value theorem, there exists $\delta_k \in [0, \delta]$ such that $f(\delta_k) = \alpha_l$. Let $\delta_l = \delta$. Then

$$\nu_k(B_X(x_k, \delta_k)) - \nu_l(B_X(x_k, \delta_k)) = \alpha_l,$$
$$\nu_l(B_X(x_l, \delta_l)) - \nu_k(B_X(x_l, \delta_l)) = \alpha_l.$$

So

$$\nu_1(B_X(x_1,\delta_1)) - \nu_2(B_X(x_1,\delta_1)) = \nu_2(B_X(x_2,\delta_2)) - \nu_1(B_X(x_2,\delta_2)).$$

Furthermore, since $B_X(x_i,\delta_i) \subseteq B_X(x_i,\delta)$ and $B_X(x_j,\delta_j) \subseteq B_X(x_j,\delta)$, we have

$$\begin{aligned} B_X(x_i,\delta_i) &\subseteq S_i, \\ B_X(x_j,\delta_j) &\subseteq S_j, \\ T_1(B_X(x_i,\delta_i)) &\subseteq B_Y(T_1(x_i),\varepsilon/2), \\ T_2(B_X(x_i,\delta_i)) &\subseteq B_Y(T_2(x_i),\varepsilon/2), \\ T_2(B_X(x_j,\delta_j)) &\subseteq B_Y(T_2(x_j),\varepsilon/2). \end{aligned}$$

Recall that, by the definition of $\varepsilon$,

$$\begin{aligned} B_Y(T_1(x_i),\varepsilon/2) \cap B_Y(T_2(x_i),\varepsilon/2) &= \varnothing, \\ B_Y(T_1(x_i),\varepsilon/2) \cap B_Y(T_2(x_j),\varepsilon/2) &= \varnothing. \end{aligned}$$

Let $A = B_X(x_1,\delta_1) \cup B_X(x_2,\delta_2) \in \mathcal{B}(X)$ and $B = B_Y(T_1(x_i),\varepsilon/2) \in \mathcal{B}(Y)$. Then

$$\nu_1(A) = \nu_1(B_X(x_1,\delta_1)) + \nu_1(B_X(x_2,\delta_2)) = \nu_2(B_X(x_1,\delta_1)) + \nu_2(B_X(x_1,\delta_1)) = \nu_2(A),$$

so

$$\int c\,\mathbb{1}(A \times Y)\; d\gamma_1 = \int c\,\mathbb{1}(A \times Y)\; d\gamma_2.$$

Furthermore,

$$\gamma_1(A \times B) = P(A \cap T_1^{-1}(B)) \geq P(B_X(x_i,\delta_i)) > 0$$

because $\operatorname{supp} P = X$. On the other hand,

$$
\begin{aligned}
&A \cap T_2^{-1}(B) \\
&\subseteq T_2^{-1}(T_2(A) \cap B) \\
&= T_2^{-1}((T_2(B_X(x_1, \delta_1)) \cap B_Y(T_1(x_i), \varepsilon/2)) \cup (T_2(B_X(x_2, \delta_2)) \cap B_Y(T_1(x_i), \varepsilon/2))) \\
&\subseteq T_2^{-1}((B_Y(T_2(x_1), \varepsilon/2) \cap B_Y(T_1(x_i), \varepsilon/2)) \cup (B_Y(T_2(x_2), \varepsilon/2) \cap B_Y(T_1(x_i), \varepsilon/2))) \\
&= T_2^{-1}(\varnothing \cup \varnothing) \\
&= \varnothing.
\end{aligned}
$$

Therefore,

$$\gamma_2(A \times B) = P(A \cap T_2^{-1}(B)) = 0.$$

Hence, $\gamma_1(A \times B) \neq \gamma_2(A \times B)$. It follows that $Q \notin \operatorname{ext}(B_c[P, r])$ by Lemma 5.6.2. ■

Note that in the proof above there is a choice of $x_i \in S_i$ and $x_j \in S_j$ for $i, j \in \{1, 2\}$ with $i \neq j$, subject to $T_1(x_i) \neq T_2(x_j)$. Under certain conditions, each distinct pair $(x_1, x_2) \in S_1 \times S_2$ produces a distinct pair $(Q_1, Q_2) \in B_c[P, r]^2$ with $Q_1 \neq Q_2$ such that $Q = Q_1/2 + Q_2/2$. In such circumstances, if there is an uncountable set of possible $(x_1, x_2) \in S_1 \times S_2$, then $Q$ can be represented as the average of an uncountable number of pairs of elements of $B_c[P, r]$. We now investigate such a setting.

Let $X = (0, 1)$ with the Euclidean distance, which is connected, and let $P = \operatorname{Unif}(0, 1)$ with $\operatorname{supp} P = X$. Since $P$ is non-atomic, it respects balls. Let $Y = (1, 2)$ with the Euclidean distance and let $Q = \operatorname{Unif}(1, 2)$. Let $c(x, y) = |x - y|$, which is continuous. For any $\gamma \in \Pi(P, Q)$, we have

$$
\begin{aligned}
\int c \, d\gamma &= \int (y - x) \, d\gamma(x, y) \\
&= \int y \, dQ(y) - \int x \, dP(x)
\end{aligned}
$$

$$= 3/2 - 1/2$$
$$= 1.$$

Hence, $W_c(P, Q) = 1$. Let $\gamma_1, \gamma_2 \in \Pi(P, Q)$ be induced by

$$T_1(x) = 1 + x,$$
$$T_2(x) = 2 - x$$

for $x \in X$. Then $T_1$ and $T_2$ are continuous, and

$$c(1/3, T_1(1/3)) = 1 \neq 4/3 = c(1/3, T_2(1/3)).$$

Hence, $Q$ is not an extreme point of $B_c[P, r]$ by Lemma 5.6.3.

Following the proof of Lemma 5.6.3, we find

$$\frac{d\nu_1}{dP}(x) = 1 \text{ and } \frac{d\nu_2}{dP}(x) = 2 - 2x$$

for $x \in X$. It follows that $S_1 = (1/2, 1)$, $S_2 = (0, 1/2)$ and $S_3 = \{1/2\}$. We can select any $x_1 \in S_1$ and $x_2 \in S_2$, as long as $x_2 \neq 1 - x_1$. In particular, there is an uncountable set of possible $(x_1, x_2) \in S_1 \times S_2$. Each distinct possible pair $(x_1, x_2) \in S_1 \times S_2$ produces a distinct pair $(Q_1, Q_2) \in B_c[P, 1]^2$ such that $Q = Q_1/2 + Q_2/2$. This is because a distinct $A = B_X(x_1, \delta_1) \cup B_X(x_2, \delta_2) \in \mathcal{B}(X)$ is produced for some $\delta_1, \delta_2 > 0$ such that $B_X(x_1, \delta_1) \subseteq S_1$ and $B_X(x_2, \delta_2) \subseteq S_2$. The proof of Lemma 5.6.2 then defines

$$Q_1(C) = \gamma_1(A \times C) + \gamma_2(A^{\mathsf{C}} \times C)$$
$$= P(A \cap T_1^{-1}(C)) + P(A^{\mathsf{C}} \cap T_2^{-1}(C))$$

and

$$\begin{aligned} Q_2(C) &= \gamma_1(A^{\mathsf{C}} \times C) + \gamma_2(A \times C) \\ &= P(A^{\mathsf{C}} \cap T_1^{-1}(C)) + P(A \cap T_2^{-1}(C)) \end{aligned}$$

for $C \in \mathcal{B}(Y)$.

For a given choice of $x_1$ and $x_2$, we can find $A$ exactly in this setting. Let $x_1 = 3/4$ and $x_2 = 1/3$. We follow the proof of Lemma 5.6.3. We have $T_1(x_1) = 7/4$, $T_2(x_2) = 5/3$ and $T_2(x_1) = 5/4$. We then find $\varepsilon = 1/12$ and can let $\delta = 1/24$ to obtain

$$\begin{aligned} B_X(x_1, \delta) &= (17/24, 19/24), \\ B_X(x_2, \delta) &= (7/24, 9/24) \end{aligned}$$

and

$$\begin{aligned} T_1(B_X(x_1, \delta)) &= (41/24, 43/24) = B_Y(T_1(x_1), \varepsilon/2), \\ T_2(B_X(x_1, \delta)) &= (29/24, 31/24) = B_Y(T_2(x_1), \varepsilon/2), \\ T_2(B_X(x_2, \delta)) &= (39/24, 41/24) = B_Y(T_2(x_2), \varepsilon/2). \end{aligned}$$

Therefore,

$$\nu_1(B_X(x_1, \delta)) = \int_{17/24}^{19/24} 1 \, dx = 1/12$$

and

$$\nu_2(B_X(x_1, \delta)) = \int_{17/24}^{19/24} (2 - 2x) \, dx = 1/24,$$

so $\alpha_1 = 1/24$. Furthermore,

$$\nu_2(B_X(x_2, \delta)) = \int_{7/24}^{9/24} (2 - 2x) \, dx = 1/9$$

and

$$\nu_1(B_X(x_2,\delta)) = \int_{7/24}^{9/24} 1\ dx = 1/12,$$

so $\alpha_2 = 1/36$. Since $\alpha_1 > \alpha_2$, we define $f : [0,\delta] \to \mathbb{R}$ by

$$f(\eta) = \nu_1(B_X(x_1,\eta)) - \nu_2(B_X(x_1,\eta)).$$

We have

$$\nu_1(B_X(x_1,\eta)) = \int_{3/4-\eta}^{3/4+\eta} 1\ dx = 2\eta$$

and

$$\nu_2(B_X(x_1,\eta)) = \int_{3/4-\eta}^{3/4+\eta} (2-2x)\ dx = \eta,$$

so $f(\eta) = \eta$. We then define $\delta_1 = 1/36$ so that $f(\delta_1) = \alpha_2$, and $\delta_2 = \delta = 1/24$. Therefore,

$$B_X(x_1,\delta_1) = (26/36, 28/36),$$
$$B_X(x_2,\delta_2) = (7/24, 9/24)$$

and $A = (7/24, 9/24) \cup (26/36, 28/36)$. For Lemma 5.6.2, we also need $B = (41/24, 43/24)$. We have

$$\gamma_1(A \times B) = P(A \cap T_1^{-1}(B)) = P((26/36, 28/36)) = 1/18$$

and

$$\gamma_2(A \times B) = P(A \cap T_2^{-1}(B)) = P(\varnothing) = 0,$$

so $\gamma_1(A \times B) \neq \gamma_2(A \times B)$. Furthermore,

$$\int c\,\mathbb{1}(A \times Y)\ d\gamma_1 = \int c(x, T_1(x))\,\mathbb{1}(x \in A)\ dP(x) = \int \mathbb{1}(x \in A)\ dx = 5/36$$

and

$$\int c\,\mathbb{1}(A \times Y)\, d\gamma_2 = \int c(x, T_2(x))\,\mathbb{1}(x \in A)\, dP(x) = \int (2 - 2x)\,\mathbb{1}(x \in A)\, dx = 5/36.$$

Hence, the conditions of Lemma 5.6.2 are satisfied. The proof of the lemma shows that $Q = Q_1/2 + Q_2/2$ for $Q_1, Q_2 \in B_c[P, 1]$, where

$$Q_1(C) = P(A \cap T_1^{-1}(C)) + P(A^{\mathsf{C}} \cap T_2^{-1}(C))$$

and

$$Q_2(C) = P(A^{\mathsf{C}} \cap T_1^{-1}(C)) + P(A \cap T_2^{-1}(C))$$

for $C \in \mathcal{B}(Y)$.

There are more general results for $X, Y \subseteq \mathbb{R}^n$ and $c(x, y) = \|x - y\|_1$. In particular, we are interested in $P \in \mathcal{P}(X)$ and $Q \in \mathcal{P}(Y)$ such that $c$ is linear on $\operatorname{supp}(P) \times \operatorname{supp}(Q)$. This happens if, for each $1 \leq i \leq n$, we have that $x_i - y_i$ has the same sign for all $(x, y) \in \operatorname{supp}(P) \times \operatorname{supp}(Q)$. In this case, $r = W_c(P, Q)$ is attained by all $\gamma \in \Pi(P, Q)$. Let $X$ be connected, $\operatorname{supp} P = X$ and $P$ respect balls. If there are two continuous transport maps $T_1, T_2$ which induce $\gamma_1, \gamma_2 \in \Pi(P, Q)$ such that $c(x, T_1(x)) \neq c(x, T_2(x))$ for some $x \in X$, then $Q \notin \operatorname{ext}(B_c[P, r])$ by Lemma 5.6.3. This is in contrast to the situation in which $X, Y \subseteq \mathbb{R}^n$ and $c(x, y) = \|x - y\|_2^2$. The discussion after Lemma 5.6.1 showed that, under a square-integrability condition on $P$, all $Q \in B_c[P, r]$ with $W_c(P, Q) = r$ are extreme points of $B_c[P, r]$.

## 5.7 Discrete Optimal Transport

We now consider the setting in which $X$ and $Y$ are finite sets. Without loss of generality, we let $X = \{1, \ldots, m\}$ and $Y = \{1, \ldots, n\}$ for $m, n \geq 1$. Note that all subsets of $X$ are Borel for any metric on $X$, and similarly for $Y$ and $X \times Y$. We define some new notation for this section. The results relating to this new notation are restatements of those in Sections 5.3 and 5.4. Let $\mathbf{1}$ be the vector of ones, with the dimension determined by the context, and let

$$\begin{aligned} \Delta(X) &= \left\{p \in [0,1]^m : \mathbf{1}^\mathsf{T} p = 1\right\}, \\ \Delta(Y) &= \left\{q \in [0,1]^n : \mathbf{1}^\mathsf{T} q = 1\right\}, \\ \Delta(X \times Y) &= \left\{\Gamma \in [0,1]^{m \times n} : \mathbf{1}^\mathsf{T} \Gamma \mathbf{1} = 1\right\}. \end{aligned}$$

Note that there is a bijection $f : \mathcal{P}(X) \to \Delta(X)$ by $(fP)_i = P(\{i\})$ for $1 \leq i \leq m$. Similarly, there is a bijection between $\mathcal{P}(Y)$ and $\Delta(Y)$, and $\mathcal{P}(X \times Y)$ and $\Delta(X \times Y)$. We can define the equivalent of the marginals of $\Gamma \in \Delta(X \times Y)$. Let $v_1 : \Delta(X \times Y) \to \Delta(X)$ by $v_1 \Gamma = \Gamma \mathbf{1}$ and $v_2 : \Delta(X \times Y) \to \Delta(Y)$ by $v_2 \Gamma = \Gamma^\mathsf{T} \mathbf{1}$. We also define

$$V(p, q) = \{\Gamma \in \Delta(X \times Y) : v_1 \Gamma = p \text{ and } v_2 \Gamma = q\}$$

for $p \in \Delta(X)$ and $q \in \Delta(Y)$.

Note that any cost function $c$ on $X \times Y$ is bounded and continuous, so assumptions $(c1)$ and $(c2)$ are satisfied. Let $C_{i,j} = c(i, j)$ for $1 \leq i \leq m$ and $1 \leq j \leq n$ and let the equivalent of the Wasserstein distance

$$w_c(p, q) = \inf\left\{\operatorname{tr}(C^\mathsf{T} \Gamma) : \Gamma \in V(p, q)\right\}$$

for $p \in \Delta(X)$ and $q \in \Delta(Y)$. We know that $w_c(p,q)$ is attained by some $\Gamma \in V(p,q)$ since assumption $(c1)$ is satisfied. Let the equivalent of the Wasserstein ball

$$b_c[p,r] = \{q \in \Delta(Y) : w_c(p,q) \leq r\}$$

for $p \in \Delta(X)$ and $r \geq 0$. We have that $b_c[p,r]$ is convex. We can also define the equivalent of dual functions, which we refer to as dual variables. Note that there is a bijection $f : L^1(X) \to \mathbb{R}^m$ by $(f\psi)_i = \psi(i)$ for $1 \leq i \leq m$. Similarly, there is a bijection between $L^1(Y)$ and $\mathbb{R}^n$. We know that

$$w_c(p,q) = \sup_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^n} \left\{ q^{\mathsf{T}}\mu - p^{\mathsf{T}}\lambda : \mu_j - \lambda_i \leq C_{i,j} \text{ for all } 1 \leq i \leq m \text{ and } 1 \leq j \leq n \right\}$$

since assumption $(c1)$ is satisfied. In fact, we know that the supremum is attained by some $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$ since assumption $(c2)$ is satisfied.

We have already classified the extreme points $q$ of $b_c[p,r]$ such that $w_c(p,q) < r$ in Section 5.5. They are the $q$ such that $q_{j_0} = 1$ for some $1 \leq j_0 \leq n$ and $q_j = 0$ for all $1 \leq j \leq n$ with $j \neq j_0$. We now consider $q \in b_c[p,r]$ such that $w_c(p,q) = r$. In the discrete setting for a fixed value of $r$, it is rare that any transport plan is induced by a transport map, as this would require that the elements of $q$ can be created from sums of the elements of $p$. In order to apply Lemma 5.6.1, which gives conditions under which $q \in b_c[p,r]$ with $w_c(p,q) = r$ is an extreme point of $b_c[p,r]$, we require a unique optimal transport plan which is induced by a transport map. This would be very unusual, so we do not seek to apply Lemma 5.6.1 in the discrete setting. We do not seek to apply Lemma 5.6.3 for similar reasons. However, we do consider scenarios in which Lemma 5.6.2 can be applied. Under conditions on two transport maps attaining $w_c(p,q)$, Lemma 5.6.2 states that $q \in b_c[p,r]$ with $w_c(p,q) = r$ is not an extreme point of $b_c[p,r]$.

We now give an example illustrating when Lemma 5.6.2 can and cannot be applied. Suppose that $m = 2$, $n = 4$ and $p \in \Delta(X)$ with $p = (1/2, 1/2)$. We consider $b_c[p, 1/3]$, where

$$C = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

Let $q_1 \in \Delta(Y)$ with $q_1 = (1/3, 1/4, 1/12, 1/3)$. If $\Gamma \in V(p, q_1)$, then $\operatorname{tr}(C^{\mathsf{T}}\Gamma) \geq q_{1,2} + q_{1,3} = 1/3$. Let $\Gamma_1, \Gamma_2 \in V(p, q_1)$ with

$$\Gamma_1 = \begin{pmatrix} 1/3 & 1/6 & 0 & 0 \\ 0 & 1/12 & 1/12 & 1/3 \end{pmatrix} \quad \text{and} \quad \Gamma_2 = \begin{pmatrix} 1/3 & 1/12 & 1/12 & 0 \\ 0 & 1/6 & 0 & 1/3 \end{pmatrix}.$$

Then $\operatorname{tr}(C^{\mathsf{T}}\Gamma_1) = \operatorname{tr}(C^{\mathsf{T}}\Gamma_2) = 1/3$, so $\Gamma_1$ and $\Gamma_2$ attain $w_c(p, q_1) = 1/3$. Let $A = \{1\}$ and $B = \{2\}$. Furthermore, let $a \in \mathbb{R}^m$ with $a_i = 1$ if $i \in A$ and $a_i = 0$ if $i \notin A$. Also, let $b \in \mathbb{R}^n$ with $b_j = 1$ if $j \in B$ and $b_j = 0$ if $j \notin B$. Then $a^{\mathsf{T}}\Gamma_1 b = 1/6$ and $a^{\mathsf{T}}\Gamma_2 b = 1/12$, so $a^{\mathsf{T}}\Gamma_1 b \neq a^{\mathsf{T}}\Gamma_2 b$. Furthermore, $\operatorname{tr}(C^{\mathsf{T}}aa^{\mathsf{T}}\Gamma_1) = \operatorname{tr}(C^{\mathsf{T}}aa^{\mathsf{T}}\Gamma_2) = 1/6$. Hence, the conditions of Lemma 5.6.2 are satisfied, so $q_1$ is not an extreme point of $b_c[p, 1/3]$. In fact, the proof of Lemma 5.6.2 shows that $q_1 = q_2/2 + q_3/2$ for $q_2, q_3 \in b_c[p, 1/3]$ with $q_2 = (1/3, 1/3, 0, 1/3)$ and $q_3 = (1/3, 1/6, 1/6, 1/3)$.

Now consider $q_2 \in b_c[p, 1/3]$ with $q_2 = (1/3, 1/3, 0, 1/3)$. We know that $w_c(p, q_2) = 1/3$ by Lemma 5.3.1, since $q_1 = q_2/2 + q_3/2$. Let $\Gamma \in V(p, q_2)$. Then $\Gamma_{1,3} = \Gamma_{2,3} = 0$ since $v_2\Gamma = q_1$ and $q_{1,3} = 0$. Suppose further that $\Gamma$ attains $w_c(p, q_2)$. Since $C_{1,2}\Gamma_{1,2} + C_{2,2}\Gamma_{2,2} = q_{2,2} = 1/3$, we must have $\Gamma_{1,4} = 0$ and $\Gamma_{2,1} = 0$. Therefore, $\Gamma_{1,1} = 1/3$ and $\Gamma_{2,4} = 1/3$ because $v_2\Gamma = q_2$. Furthermore, $\Gamma_{1,2} = 1/6$ and $\Gamma_{2,2} = 1/6$ because $v_1\Gamma = p$. It follows that $w_c(p, q_2)$ is uniquely attained by

$$\Gamma = \begin{pmatrix} 1/3 & 1/6 & 0 & 0 \\ 0 & 1/6 & 0 & 1/3 \end{pmatrix},$$

so Lemma 5.6.2 does not apply. However, $q_2 = q_4/2 + q_5/2$ for $q_4 = (1/6, 1/3, 0, 1/2)$ and $q_5 = (1/2, 1/3, 0, 1/6)$. By letting

$$\Gamma_4 = \begin{pmatrix} 1/6 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \Gamma_5 = \begin{pmatrix} 1/2 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 1/6 \end{pmatrix},$$

we find $q_4, q_5 \in b_c[p, 1/3]$ because $\Gamma_4 \in V(p, q_4)$ and $\Gamma_5 \in V(p, q_5)$ with $\mathrm{tr}(C^\mathsf{T}\Gamma_4) = \mathrm{tr}(C^\mathsf{T}\Gamma_5) = 1/3$. Hence, $q_2$ is not an extreme point of $b_c[p, 1/3]$, even though $w_c(p, q_2)$ is uniquely attained.

Finally, consider $q_4 \in b_c[p, 1/3]$ with $q_4 = (1/6, 1/3, 0, 1/2)$. We know that $w_c(p, q_4) = 1/3$ by Lemma 5.3.1, since $q_2 = q_4/2 + q_5/2$. Suppose that $q_4 = tq_6 + (1 - t)q_7$ for $q_6, q_7 \in b_c[p, 1/3]$ and $t \in (0, 1)$. We have that $w_c(p, q_6) = w_c(p, q_7) = 1/3$ by Lemma 5.3.1. Suppose that $\Gamma_6$ attains $w_c(p, q_6)$ and $\Gamma_7$ attains $w_c(p, q_7)$. We have that $q_{6,3} = q_{7,3} = 0$, so $\Gamma_{6,1,3} = \Gamma_{6,2,3} = 0$ and $\Gamma_{7,1,3} = \Gamma_{7,2,3} = 0$. Furthermore, $tq_{6,2} + (1-t)q_{7,2} = 1/3$. However, $q_{6,2} \leq w_c(p, q_6) = 1/3$ and $q_{7,2} \leq w_c(p, q_7) = 1/3$, so $q_{6,2} = q_{7,2} = 1/3$. Since $C_{1,2}\Gamma_{6,1,2} + C_{2,2}\Gamma_{6,2,2} = q_{6,2} = 1/3$, we must have $\Gamma_{6,1,4} = 0$ and $\Gamma_{6,2,1} = 0$. Similarly, $\Gamma_{7,1,4} = 0$ and $\Gamma_{7,2,1} = 0$. It follows that $t\Gamma_{6,2,4} + (1-t)\Gamma_{7,2,4} = 1/2$. However, $\Gamma_{6,2,4} \leq 1/2$ and $\Gamma_{7,2,4} \leq 1/2$ because $v_1\Gamma_6 = v_1\Gamma_7 = p$, so $\Gamma_{6,2,4} = \Gamma_{7,2,4} = 1/2$. It follows that $\Gamma_{6,2,2} = \Gamma_{7,2,2} = 0$ because $v_1\Gamma_6 = v_1\Gamma_7 = p$. Furthermore, $\Gamma_{6,1,2} = \Gamma_{7,1,2} = 1/3$ because $q_{6,2} = q_{7,2} = 1/3$. Finally, $\Gamma_{6,1,1} = \Gamma_{7,1,1} = 1/6$ because $v_1\Gamma_6 = v_1\Gamma_7 = p$. Hence, $\Gamma_6 = \Gamma_7 = \Gamma_4$ and $q_6 = q_7 = q_4$. Therefore, $q_4$ is an extreme point of $b_c[p, 1/3]$.

For discrete optimal transport problems, we can greatly exploit the dual variables $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$ when finding conditions under which $q \in b_c[p, r]$ with $w_c(p, q) = r$ is not an extreme point of $b_c[p, r]$. Our aim is to use Lemma 5.4.1 as follows. Suppose that $q \in b_c[p, r]$ is not an extreme point of $b_c[p, r]$. Then there exist $q_1, q_2 \in b_c[p, r]$ and $t \in (0, 1)$ such that $q = tq_1 + (1-t)q_2$. We then know from Lemma 5.4.1 that dual variables attaining $w_c(p, q_1) = r$ and $w_c(p, q_2) = r$ are the same as those attaining $w_c(p, q) = r$. This allows us to restrict our search for $q_1$ and $q_2$. We first show the following auxiliary result. It is well-known, even when generalised to non-discrete settings (see Theorem 5.10 of Villani, 2009). We include the result with its proof for completeness.

**Lemma 5.7.1** *Let $\Gamma \in V(p, q)$ and let $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$ satisfy $\mu_j - \lambda_i \leq C_{i,j}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$. The following are equivalent.*

*(i)* $\operatorname{tr}(C^\mathsf{T}\Gamma) = q^\mathsf{T}\mu - p^\mathsf{T}\lambda$.

*(ii) Either $\Gamma_{i,j} = 0$ or $\mu_j - \lambda_i = C_{i,j}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$.*

*When (i) and (ii) hold, $\Gamma$ attains $w_c(p, q)$ and $w_c(p, q) = q^\mathsf{T}\mu - p^\mathsf{T}\lambda$. Such $\Gamma$, $\mu$ and $\lambda$ exist.*

**Proof** Suppose (i). Since $p = \Gamma\mathbf{1}$ and $q = \Gamma^\mathsf{T}\mathbf{1}$, we have

$$\begin{aligned}
\operatorname{tr}(C^\mathsf{T}\Gamma) &= q^\mathsf{T}\mu - p^\mathsf{T}\lambda \\
&= \mathbf{1}^\mathsf{T}\Gamma\mu - \mathbf{1}^\mathsf{T}\Gamma^\mathsf{T}\lambda \\
&= \mathbf{1}^\mathsf{T}\Gamma\mu - \lambda^\mathsf{T}\Gamma\mathbf{1} \\
&= \operatorname{tr}(\mu\mathbf{1}^\mathsf{T}\Gamma - \mathbf{1}\lambda^\mathsf{T}\Gamma) \\
&= \operatorname{tr}((\mathbf{1}\mu^\mathsf{T} - \lambda\mathbf{1}^\mathsf{T})^\mathsf{T}\Gamma).
\end{aligned}$$

Hence, $\operatorname{tr}((C - \mathbf{1}\mu^\mathsf{T} + \lambda\mathbf{1}^\mathsf{T})^\mathsf{T}\Gamma) = 0$. Furthermore, $\mu_j - \lambda_i \leq C_{i,j}$ and $\Gamma_{i,j} \geq 0$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$. Therefore, we have that (ii) holds.

Now suppose (ii) instead. We have that $\Gamma_{i,j}\mu_j - \Gamma_{i,j}\lambda_i = \Gamma_{i,j}C_{i,j}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$. Summing over $i$ and $j$ shows that (i) holds.

If (i) holds, we know that $\Gamma$ is optimal from Section 5.4. Hence, $\Gamma$ attains $w_c(p, q)$

and $w_c(p,q) = q^\mathsf{T}\mu - p^\mathsf{T}\lambda$. We know that $\Gamma$, $\lambda$ and $\mu$ exist from Section 5.3, since assumptions $(c1)$ and $(c2)$ are satisfied. ■

We now give conditions under which $q \in b_c[p,r]$ is not an extreme point of $b_c[p,r]$. The next result shows that if an optimal transport plan transports mass from one atom to two atoms at the same unit cost, then we do not have an extreme point.

**Lemma 5.7.2** *Let* $q \in b_c[p,r]$ *and let* $\Gamma \in V(p,q)$ *attain* $w_c(p,q)$*. Suppose that there exist* $1 \leq i_1 \leq m$ *and* $1 \leq j_1, j_2 \leq n$ *with* $j_1 \neq j_2$ *such that* $\Gamma_{i_1,j_1}, \Gamma_{i_1,j_2} > 0$ *and* $C_{i_1,j_1} = C_{i_1,j_2}$*. Then* $q \notin \operatorname{ext}(b_c[p,r])$*.*

**Proof** Let $\varepsilon = \min(\Gamma_{i_1,j_1}, \Gamma_{i_1,j_2})$ and $q_1, q_2 \in \Delta(Y)$ with

$$q_{1,j} = \begin{cases} q_{j_1} + \varepsilon & \text{if } j = j_1 \\ q_{j_2} - \varepsilon & \text{if } j = j_2 \\ q_j & \text{if } j \notin \{j_1, j_2\} \end{cases}$$

and

$$q_{2,j} = \begin{cases} q_{j_1} - \varepsilon & \text{if } j = j_1 \\ q_{j_2} + \varepsilon & \text{if } j = j_2 \\ q_j & \text{if } j \notin \{j_1, j_2\}. \end{cases}$$

Note that $q = q_1/2 + q_2/2$, and $q_1 \neq q_2$ since $\varepsilon > 0$. Let $\Gamma_1 \in V(p,q_1)$ and $\Gamma_2 \in V(p,q_2)$ with

$$\Gamma_{1,i,j} = \begin{cases} \Gamma_{i_1,j_1} + \varepsilon & \text{if } i = i_1 \text{ and } j = j_1 \\ \Gamma_{i_1,j_2} - \varepsilon & \text{if } i = i_1 \text{ and } j = j_2 \\ \Gamma_{i,j} & \text{if } i \neq i_1 \text{ or } j \notin \{j_1, j_2\} \end{cases}$$

and

$$\Gamma_{2,i,j} = \begin{cases} \Gamma_{i_1,j_1} - \varepsilon & \text{if } i = i_1 \text{ and } j = j_1 \\ \Gamma_{i_1,j_2} + \varepsilon & \text{if } i = i_1 \text{ and } j = j_2 \\ \Gamma_{i,j} & \text{if } i \neq i_1 \text{ or } j \notin \{j_1, j_2\}. \end{cases}$$

Let $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$ satisfy $\mu_j - \lambda_i \leq C_{i,j}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$, and $w_c(p,q) = q^\mathsf{T}\mu - p^\mathsf{T}\lambda$. From Lemma 5.7.1, we have that $\Gamma_{i,j} = 0$ or $\mu_j - \lambda_i = C_{i,j}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$. If $\Gamma_{1,i,j} > 0$, then $\Gamma_{i,j} > 0$ and $\mu_j - \lambda_i = C_{i,j}$. Hence, $\Gamma_1$ attains $w_c(p, q_1)$ by Lemma 5.7.1. Similarly, $\Gamma_2$ attains $w_c(p, q_2)$. Furthermore, we have $C_{i_1,j_1} = C_{i_1,j_2}$, so $\operatorname{tr}(C^\mathsf{T}\Gamma_1) = \operatorname{tr}(C^\mathsf{T}\Gamma_2) = \operatorname{tr}(C^\mathsf{T}\Gamma)$. Therefore, $q_1, q_2 \in b_c[p,r]$. The result follows. ■

Note that this result provides us with another way of showing that $q_1$ is not an extreme point of $b_c[p, 1/3]$ in the example above. We apply the result with $\Gamma = \Gamma_1$, $i_1 = 2$, $j_1 = 2$ and $j_2 = 3$. However, it cannot be used to show that $q_2$ is not an extreme point of $b_c[p, 1/3]$. We also have the following result. It is similar to the previous lemma, except that we may consider two different atoms from which mass is transported, under conditions on the optimal dual variables.

**Lemma 5.7.3** *Let $q \in b_c[p,r]$ and let $\Gamma \in V(p,q)$ attain $w_c(p,q)$. Also let $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$ satisfy $\mu_j - \lambda_i \leq C_{i,j}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$, and $w_c(p,q) = q^\mathsf{T}\mu - p^\mathsf{T}\lambda$. Suppose that there exist $1 \leq i_1, i_2 \leq m$ with $i_1 \neq i_2$ and $1 \leq j_1, j_2 \leq n$ with $j_1 \neq j_2$ such that $\Gamma_{i_1,j_1}, \Gamma_{i_2,j_2} > 0$, $\mu_{j_1} = \mu_{j_2}$, $C_{i_1,j_1} = C_{i_1,j_2}$ and $C_{i_2,j_1} = C_{i_2,j_2}$. Then $q \notin \operatorname{ext}(b_c[p,r])$.*

**Proof** Let $\varepsilon = \min(\Gamma_{i_1,j_1}, \Gamma_{i_2,j_2})$ and $q_1, q_2 \in \Delta(Y)$ with

$$q_{1,j} = \begin{cases} q_{j_1} - \varepsilon & \text{if } j = j_1 \\ q_{j_2} + \varepsilon & \text{if } j = j_2 \\ q_j & \text{if } j \notin \{j_1, j_2\} \end{cases}$$

and

$$q_{2,j} = \begin{cases} q_{j_1} + \varepsilon & \text{if } j = j_1 \\ q_{j_2} - \varepsilon & \text{if } j = j_2 \\ q_j & \text{if } j \notin \{j_1, j_2\}. \end{cases}$$

Note that $q = q_1/2 + q_2/2$, and $q_1 \neq q_2$ since $\varepsilon > 0$. Let $\Gamma_1 \in V(p, q_1)$ and $\Gamma_2 \in V(p, q_2)$ with

$$\Gamma_{1,i,j} = \begin{cases} \Gamma_{i_1,j_1} - \varepsilon & \text{if } i = i_1 \text{ and } j = j_1 \\ \Gamma_{i_1,j_2} + \varepsilon & \text{if } i = i_1 \text{ and } j = j_2 \\ \Gamma_{i,j} & \text{if } i \neq i_1 \text{ or } j \notin \{j_1, j_2\} \end{cases}$$

and

$$\Gamma_{2,i,j} = \begin{cases} \Gamma_{i_2,j_1} + \varepsilon & \text{if } i = i_2 \text{ and } j = j_1 \\ \Gamma_{i_2,j_2} - \varepsilon & \text{if } i = i_2 \text{ and } j = j_2 \\ \Gamma_{i,j} & \text{if } i \neq i_2 \text{ or } j \notin \{j_1, j_2\}. \end{cases}$$

From Lemma 5.7.1, we have that $\Gamma_{i,j} = 0$ or $\mu_j - \lambda_i = C_{i,j}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$. If $\Gamma_{1,i,j} > 0$ for $i \neq i_1$ or $j \neq j_2$, then $\Gamma_{i,j} > 0$ and $\mu_j - \lambda_i = C_{i,j}$. For $i = i_1$ and $j = j_2$, since $\Gamma_{i_1,j_1} > 0$, $\mu_{j_1} = \mu_{j_2}$ and $C_{i_1,j_1} = C_{i_1,j_2}$, we have

$$\begin{aligned} \mu_{j_2} - \lambda_{i_1} &= \mu_{j_1} - \lambda_{i_1} \\ &= C_{i_1,j_1} \\ &= C_{i_1,j_2}. \end{aligned}$$

Similarly, if $\Gamma_{2,i,j} > 0$ for $i \neq i_2$ or $j \neq j_1$, then $\Gamma_{i,j} > 0$ and $\mu_j - \lambda_i = C_{i,j}$. For $i = i_2$ and $j = j_1$, since $\Gamma_{i_2,j_2} > 0$, $\mu_{j_1} = \mu_{j_2}$ and $C_{i_2,j_1} = C_{i_2,j_2}$, we have

$$\begin{aligned} \mu_{j_1} - \lambda_{i_2} &= \mu_{j_2} - \lambda_{i_2} \\ &= C_{i_2,j_2} \\ &= C_{i_2,j_1}. \end{aligned}$$

Hence, $\Gamma_1$ attains $w_c(p, q_1)$ and $\Gamma_2$ attains $w_c(p, q_2)$ by Lemma 5.7.1. Furthermore, we have $C_{i_1,j_1} = C_{i_1,j_2}$ and $C_{i_2,j_1} = C_{i_2,j_2}$, so $\operatorname{tr}(C^\mathsf{T}\Gamma_1) = \operatorname{tr}(C^\mathsf{T}\Gamma_2) = \operatorname{tr}(C^\mathsf{T}\Gamma)$. Therefore, $q_1, q_2 \in b_c[p, r]$. The result follows. ■

Again, this result can be used to show that $q_1$ is not an extreme point of $b_c[p, 1/3]$, but cannot be used to show that $q_3$ is not an extreme point of $b_c[p, 1/3]$. We may let $\lambda = (0, 0)$ and $\mu = (0, 1, 1, 0)$. Selecting $\Gamma = \Gamma_1$, we set $i_1 = 1$, $i_2 = 2$, $j_1 = 2$ and $j_2 = 3$. In general, we can use Lemma 5.7.1 to calculate the dual variables from $\Gamma \in V(p, q)$ attaining $w_c(p, q)$ by setting $\mu_j - \lambda_i = C_{i,j}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$ such that $\Gamma_{i,j} > 0$. We may also set $\lambda_1 = 0$, since adding the same constant to all elements of $\lambda$ and $\mu$ has no effect on $q^\mathsf{T}\mu - p^\mathsf{T}\lambda$.

## 5.8 Discussion

In this chapter, we study conditions under which probability measures in a Wasserstein ball are extreme points or not extreme points. We show that, under very general conditions, the only extreme points of Wasserstein balls which do not lie on the surface of the ball are Dirac measures. We then investigate which points on the surface of the ball are extreme points. We find that if the Wasserstein distance is uniquely attained by a transport plan induced by a transport map, then the point is an extreme point. On the other hand, under conditions on the centre of the ball and the cost function, if the Wasserstein distance is attained by two distinct transport plans induced by continuous transport maps, then the point is not an extreme point. Furthermore, when our probability measures are defined on finite sets, we use the solutions to the dual problem to prove conditions under which we do not have an extreme point.

Although we make use of the solutions to the dual problem in the discrete setting, it would be interesting to investigate how these solutions can be used more generally. We have some idea of their behaviour from the results in Section 5.4. However, ideally we would use properties of the solutions in order to characterise the extreme points on the surface of the ball.

# Chapter 6

# Optimal Transport for Covariate Shift in RKHS Regression

In some statistical settings, an estimator is used for prediction in a slightly different situation to that in which the original data set is collected. For example, the estimator could be applied a little later in time or in a neighbouring location. If this is the case, a new independent data point which we want to predict could have a different distribution to that of the data set used to construct the estimator. This makes it difficult to bound the error of the estimator at the new data point. Clearly it is not possible to provide guarantees for all potential distributions of the new data point. However, if we assume that the distribution of the new data point is only a slight perturbation of the distribution which generates the original data set, some assurances can be given.

In order to quantify the size of the perturbation of the distribution of the original data set, we need a concept of distance between probability measures. In this chapter, we use the Wasserstein distance. The Wasserstein distance is determined by a cost function on the underlying space, which means that information about the cost between two points is transferred to the distance between two probability measures. An important example is given by setting the cost function equal to some metric on the space. The Wasserstein distance also arises naturally in the analysis of the Ivanov-regularised least-squares estimators we consider for the regression problem below. We consider all perturbations of the distribution of the original data set up to a fixed size, which defines a ball around this distribution with respect to the Wasserstein distance.

Before discussing the regression problem considered in this chapter, we first describe the Wasserstein distance in more detail. The Wasserstein distance is defined using the optimal transport problem, which aims to find the optimal transportation of one probability measure to another with respect to a given cost function. This is done by finding a transportation plan between the two probability measures which minimises the transport cost. The modern treatment of this problem began with Kantorovitch (1958), and a more recent examination is given by Villani (2009). For $p \in [1, \infty)$, the Wasserstein distance is usually defined as the $p^{-1}$th power of the minimum transport cost when the cost function is equal to the $p$th power of the metric on the underlying space (see Definition 6.1 of Villani, 2009). However, in this chapter we allow weaker assumptions on the cost function but demand that $p = 1$. This is the same as the earliest definitions of distance between probability measures using the optimal transport problem (Kantorovitch, 1958).

In this chapter, we consider a regression problem in which we seek guarantees on our estimator with respect to distributions other than that which generates the original covariates. This allows us to bound the expected squared error of the estimator for an expectation over a new independent covariate generated by a different distribution. We refer to this situation as a covariate shift. Covariate shift problems have previously been considered for a single known perturbation of the original covariate distribution by Shimodaira (2000). They have also been studied for a single unknown perturbation by Sugiyama et al. (2008). However, we seek to control the worst-case squared $L^2$ error with respect to all perturbations of the original covariate distribution up to a fixed size. Specifically, we consider a Wasserstein ball of probability measures centred at the original covariate distribution.

We first give bounds on the worst-case squared $L^2$ error for Ivanov-regularised least-squares estimators. These estimators are defined to be the minimisers of the empirical squared error over balls of different radii in a reproducing-kernel Hilbert space (RKHS). Ivanov-regularised least-squares estimators are discussed further in Chapter 3. The Wasserstein distance arises naturally in the analysis of the estimators. We consider both unbounded and bounded regression functions. When the regression function is unbounded, we produce expectation bounds on the worst-case squared error under very general conditions. We are also able to produce expectation bounds when the regression function is bounded. If we further assume that the errors of the response variables are subgaussian, we can provide high-probability bounds on the worst-case squared $L^2$ error. The bounds we produce do not tend to 0 as the number of data points tends to infinity. This is to be expected, as in general the original covariate distribution and its perturbation can have different supports.

We then discuss the challenges which arise when attempting to define alternative estimators. The estimators we consider are based on an empirical version of the worst-case squared $L^2$ error. The original covariate distribution is replaced by the empirical distribution of the covariates. However, finding an empirical version of the regression function in this setting is more difficult. There are also problems when trying to compute such estimators. However, we do provide one result based on the Choquet–Bishop–de Leeuw theorem (Theorem 5.6 of Bishop and de Leeuw, 1959) which aides the computation. Under suitable conditions, the worst-case squared $L^2$

error of the estimator is attained at an extreme point of the Wasserstein ball of perturbations. We conclude by briefly considering the approximation properties of the regression function.

## 6.1 Literature Review

Covariate shift for parametric statistical models is discussed by Shimodaira (2000) for a single perturbation of the original covariate distribution. The author assumes that both the original covariate distribution and its perturbation are known. The ratio of the densities of the distributions at the original covariates is used to weight the log-likelihood contributions of the data points. Estimation is then performed using a maximum weighted log-likelihood estimation procedure.

For more general statistical models, covariate shift is considered by Sugiyama et al. (2008). Again, the authors investigate the case in which there is a single perturbation of the original covariate distribution. However, both the original covariate distribution and its perturbation are unknown and only samples from each are available. Similarly to Shimodaira (2000), the log-likelihood contributions of each data point are weighted so that the resulting weighted log-likelihood is more closely related to the perturbation of the original covariate distribution than the original covariate distribution itself. However, instead of a ratio of densities, the weights are modelled as a linear combination of a finite set of basis functions.

The first modern treatment of the optimal transport problem is given by Kantorovitch (1958). The author represents the minimum transport cost as a function of the two measures defining the problem, introducing the earliest version of the Wasserstein distance. The measures are not required to be probability measures, but they must have the same total mass. More recently, a book on optimal transport has been written by Villani (2009). The book covers a wide range of topics, but in particular Chapter 6 examines the Wasserstein distance for general $p \in [1, \infty)$.

## 6.2 Contribution

In this chapter, we provide bounds on the worst-case squared $L^2$ error of Ivanov-regularised least-squares estimators with respect to a Wasserstein ball of probability measures centred at the original covariate distribution. We first consider the case in which the regression function is unbounded. We provide an expectation bound when using the most natural cost function for the optimal transport problem which defines the Wasserstein ball (Theorem 6.6.2 on page 230). We also provide an expectation bound when the cost function is equal to the square of the kernel metric (Theorem 6.6.5 on page 234).

When the regression function is bounded, we provide an expectation bound for the case in which the cost function of the optimal transport problem is again equal to the square of the kernel metric (Theorem 6.6.8 on page 237). Furthermore, we provide a high-probability bound under the additional assumption that the errors of the response variables are subgaussian (Theorem 6.6.10 on page 240).

We then consider the problem of defining alternative estimators based on an empirical version of the worst-case squared $L^2$ error in Section 6.7. We discuss the problems with both the analysis and computation of such estimators. We show that under suitable conditions, the worst-case squared $L^2$ error of the estimator is attained at an extreme point of the Wasserstein ball of perturbations (a consequence of Lemma 6.7.1 on page 245). Finally, we briefly consider the approximation properties of the regression function.

## 6.3 Optimal Transport

Let $(S, d)$ be a complete separable metric space, let $\mathcal{B}(S)$ and $\mathcal{B}(S \times S)$ be the Borel sets on $S$ and $S \times S$ and let $\mathcal{P}(S)$ and $\mathcal{P}(S \times S)$ be the set of Borel probability measures on $S$ and $S \times S$. We consider the problem of optimally transporting a probability measure $P \in \mathcal{P}(S)$ to $Q \in \mathcal{P}(S)$ with respect to some Borel cost function $c : S \times S \to [0, \infty)$. Denote the marginals of $\gamma \in \mathcal{P}(S \times S)$ by $\pi_1\gamma, \pi_2\gamma \in \mathcal{P}(S)$. We define

$$\Pi(P, Q) = \{\gamma \in \mathcal{P}(S \times S) : \pi_1\gamma = P \text{ and } \pi_2\gamma = Q\}$$

for $P, Q \in \mathcal{P}(S)$ and refer to $\gamma \in \Pi(P, Q)$ as a transport plan. The optimal transport problem seeks to find

$$\inf_{\gamma \in \Pi(P,Q)} \int c \, d\gamma.$$

The value of the infimum is known as the Wasserstein distance $W_c(P, Q)$. We define the closed Wasserstein ball

$$B_c[P, W] = \{Q \in \mathcal{P}(S) : W_c(P, Q) \leq W\}$$

for $P \in \mathcal{P}(S)$ and $W \geq 0$. We have that $B_c[P, W]$ is convex by Lemma 5.3.1 with $P_1 = P_2 = P$. We define $Q \in B_c[P, W]$ to be an extreme point of $B_c[P, W]$ if $Q = tQ_1 + (1 - t)Q_2$ for $Q_1, Q_2 \in B_c[P, W]$ and $t \in (0, 1)$ implies $Q_1 = Q_2$, in which case $Q_1 = Q_2 = Q$. We denote the set of extreme points of any convex set $A$ by $\operatorname{ext}(A)$.

## 6.4 RKHSs and Their Interpolation Spaces

An RKHS $H$ on $S$ is a Hilbert space of real-valued functions on $S$ such that, for all $x \in S$, there is some $k_x \in H$ such that $h(x) = \langle h, k_x \rangle_H$ for all $h \in H$. The function $k(x_1, x_2) = \langle k_{x_1}, k_{x_2} \rangle_H$ for $x_1, x_2 \in S$ is known as the kernel and is symmetric and positive-definite.

We now define interpolation spaces between a Banach space $(Z, \|\cdot\|_Z)$ and a dense subspace $(V, \|\cdot\|_V)$ (see Bergh and Löfström, 1976). The $K$-functional of $(Z, V)$ is

$$K(z,t) = \inf_{v \in V} (\|z - v\|_Z + t\|v\|_V)$$

for $z \in Z$ and $t > 0$. We define

$$\|z\|_{\beta,q} = \left( \int_0^\infty (t^{-\beta} K(z,t))^q t^{-1} dt \right)^{1/q} \text{ and } \|z\|_{\beta,\infty} = \sup_{t>0} (t^{-\beta} K(z,t))$$

for $z \in Z$, $\beta \in (0,1)$ and $1 \leq q < \infty$. We then define the interpolation space $[Z, V]_{\beta,q}$ to be the set of $z \in Z$ such that $\|z\|_{\beta,q} < \infty$. The size of $[Z, V]_{\beta,q}$ decreases as $\beta$ increases. Recall Lemma 3.1.1, which is essentially Theorem 3.1 of Smale and Zhou (2003).

**Lemma 6.4.1** *Let $(Z, \|\cdot\|_Z)$ be a Banach space, $(V, \|\cdot\|_V)$ be a dense subspace of $Z$ and $z \in [Z, V]_{\beta,\infty}$. We have*

$$\inf\{\|v - z\|_Z : v \in V, \|v\|_V \leq r\} \leq \frac{\|z\|_{\beta,\infty}^{1/(1-\beta)}}{r^{\beta/(1-\beta)}}.$$

From the above, when $H$ is dense in $L^\infty$, we can define the interpolation spaces $[L^\infty, H]_{\beta,q}$, where $L^\infty$ is the space of bounded measurable functions on $(S, \mathcal{B}(S))$. We assume that $S$ is a topological space. We set $q = \infty$ and work with the largest space of functions for a fixed $\beta \in (0,1)$. We are then able to apply the approximation result in Lemma 6.4.1.

## 6.5 Problem Definition

We now define the regression problem. Let $(S, d)$ be a complete separable metric space and $(X_i, Y_i)$ for $1 \le i \le n$ be i.i.d. $(S \times \mathbb{R}, \mathcal{B}(S) \otimes \mathcal{B}(\mathbb{R}))$-valued random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume $X_i \sim P$ and $\mathbb{E}(Y_i^2) < \infty$, where $\mathbb{E}$ denotes integration with respect to $\mathbb{P}$. We have $\mathbb{E}(Y_i | X_i) = g(X_i)$ almost surely for some function $g$ which is measurable on $(S, \mathcal{B}(S))$ (Section A3.2 of Williams, 1991). Since $\mathbb{E}(Y_i^2) < \infty$, we have that $g \in L^2(P)$ by Jensen's inequality. We assume that

$(Y1)$ $$\operatorname{var}(Y_i | X_i) \le \sigma^2 \text{ almost surely for } 1 \le i \le n.$$

Our results depend on how well $g$ can be approximated by elements of an RKHS $H$ with kernel $k$. We make the following assumptions.

$(H)$ The RKHS $H$ with kernel $k$ has the following properties:

- The RKHS $H$ is separable.
- The kernel $k$ is bounded.
- The kernel $k$ is a measurable function on $(S \times S, \mathcal{B}(S) \otimes \mathcal{B}(S))$.

We define

$$\|k\|_\infty = \sup_{x \in S} k(x,x)^{1/2} < \infty.$$

We can guarantee that $H$ is separable by, for example, assuming that $k$ is continuous (Lemma 4.33 of Steinwart and Christmann, 2008). Since $H$ has a kernel $k$ which is measurable on $(S \times S, \mathcal{B}(S) \otimes \mathcal{B}(S))$, we have that all functions in $H$ are measurable on $(S, \mathcal{B}(S))$ (Lemma 4.24 of Steinwart and Christmann, 2008).

## 6.6 Ivanov-Regularised Least-Squares Estimators

We are interested in estimating the regression function $g$ using elements of the RKHS $H$. Let $B_H$ be the closed unit ball of $H$ and let $r > 0$. We define the Ivanov-regularised least-squares estimator constrained to lie in $rB_H$ as

$$\hat{h}_r = \operatorname*{arg\,min}_{f \in rB_H} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2.$$

This estimator is discussed in Chapter 3. Its definition is unique if we demand that $\hat{h}_r \in \operatorname{sp}\{k_{X_i} : 1 \leq i \leq n\}$. We also define $\hat{h}_0 = 0$. The estimator $\hat{h}_r$ becomes smoother as $r$ decreases, as it is constrained to lie closer to 0.

### 6.6.1 Unbounded Regression Function

We start by considering the setting in which the regression function is unbounded. For $r \geq 0$ and $h_r \in rB_H$, Corollary 3.10.4 provides expectation bounds on the squared $L^2(P)$ norm of $\hat{h}_r - h_r$. These can be transferred to bounds on the squared $L^2(Q)$ norm of $\hat{h}_r - h_r$ using optimal transport. We have the following result.

**Lemma 6.6.1** *Assume (H). Let the cost function $c : S \times S \to [0, \infty)$ by*

$$c(x_1, x_2) = \|k_{x_2} + k_{x_1}\|_H \, \|k_{x_2} - k_{x_1}\|_H.$$

*For $Q \in \mathcal{P}(S)$, we have*

$$\|\hat{h}_r - h_r\|_{L^2(Q)}^2 \leq \|\hat{h}_r - h_r\|_{L^2(P)}^2 + 4r^2 W_c(P, Q)$$

*for all $r > 0$.*

**Proof** Let $x_1, x_2 \in S$. By the reproducing kernel property and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&|(\hat{h}_r - h_r)(x_2)^2 - (\hat{h}_r - h_r)(x_1)^2| \\
&= |(\hat{h}_r - h_r)(x_2) + (\hat{h}_r - h_r)(x_1)| \, |(\hat{h}_r - h_r)(x_2) - (\hat{h}_r - h_r)(x_1)| \\
&= |\langle \hat{h}_r - h_r, k_{x_2} + k_{x_1} \rangle_H| \, |\langle \hat{h}_r - h_r, k_{x_2} - k_{x_1} \rangle_H| \\
&\leq \|\hat{h}_r - h_r\|_H^2 \, \|k_{x_2} + k_{x_1}\|_H \, \|k_{x_2} - k_{x_1}\|_H \\
&\leq 4r^2 \|k_{x_2} + k_{x_1}\|_H \, \|k_{x_2} - k_{x_1}\|_H .
\end{aligned}$$

Hence,

$$(\hat{h}_r - h_r)(x_2)^2 \leq (\hat{h}_r - h_r)(x_1)^2 + 4r^2 \|k_{x_2} + k_{x_1}\|_H \, \|k_{x_2} - k_{x_1}\|_H .$$

Integrating over $(x_1, x_2)$ with respect to $\gamma \in \Pi(P, Q)$ gives

$$\|\hat{h}_r - h_r\|_{L^2(Q)}^2 \leq \|\hat{h}_r - h_r\|_{L^2(P)}^2 + 4r^2 \int \|k_{x_2} + k_{x_1}\|_H \, \|k_{x_2} - k_{x_1}\|_H \, d\gamma(x_1, x_2).$$

The result follows by taking an infimum over $\gamma \in \Pi(P, Q)$. ■

We can use this result to provide an expectation bound on the squared $L^2(Q)$ error of $\hat{h}_r$ in the same way as the proof of Theorem 3.7.1. In order to understand how well

$g$ can be approximated by elements of $H$ in this context, we define

$$I_c(g, r, W) = \inf \left\{ \sup_{Q \in B_c[P,W]} \|h_r - g\|^2_{L^2(Q)} : h_r \in rB_H \right\}$$

for $r > 0$ and $W > 0$.

**Theorem 6.6.2** *Assume (Y1) and (H). Let the cost function* $c : S \times S \to [0, \infty)$ *by*

$$c(x_1, x_2) = \|k_{x_2} + k_{x_1}\|_H \, \|k_{x_2} - k_{x_1}\|_H.$$

*Supposing that the expectation below exists, we have*

$$\mathbb{E}\left( \sup_{Q \in B_c[P,W]} \|\hat{h}_r - g\|^2_{L^2(Q)} \right) \leq \frac{8\|k\|_\infty \sigma r}{n^{1/2}} + \frac{64\|k\|^2_\infty r^2}{n^{1/2}} + 10 I_c(g, r, W) + 8Wr^2$$

*for all* $r > 0$ *and all* $W > 0$.

**Proof** By Lemma 6.6.1, we have

$$\begin{aligned} \|\hat{h}_r - g\|^2_{L^2(Q)} &\leq 2\|\hat{h}_r - h_r\|^2_{L^2(Q)} + 2\|h_r - g\|^2_{L^2(Q)} \\ &\leq 2\|\hat{h}_r - h_r\|^2_{L^2(P)} + 8r^2 W_c(P, Q) + 2\|h_r - g\|^2_{L^2(Q)}. \end{aligned}$$

Hence,

$$\sup_{Q \in B_c[P,W]} \|\hat{h}_r - g\|^2_{L^2(Q)} \leq 2\|\hat{h}_r - h_r\|^2_{L^2(P)} + 8Wr^2 + \sup_{Q \in B_c[P,W]} 2\|h_r - g\|^2_{L^2(Q)}.$$

Therefore,

$$\mathbb{E}\left( \sup_{Q \in B_c[P,W]} \|\hat{h}_r - g\|^2_{L^2(Q)} \right) \leq \frac{8\|k\|_\infty \sigma r}{n^{1/2}} + \frac{64\|k\|^2_\infty r^2}{n^{1/2}} + 8Wr^2 + \sup_{Q \in B_c[P,W]} 10\|h_r - g\|^2_{L^2(P)}$$

by Corollary 3.10.4. The result follows by taking an infimum over $h_r \in rB_H$. ■

The first two terms on the right-hand side of the inequality in Theorem 6.6.2 make up the variance of the estimator and tend to 0 as $n$ tends to infinity. However, the last two terms do not tend to 0 as $n$ tends to infinity. The third term is the bias of the estimator, which decreases with $r$, while the fourth term is the cost of being robust against changes in the covariate distribution, which increases with $r$. Asymptotically, our bound comprises only these final two terms. If we seek to minimise these two terms over $r$, we obtain a value of $r$ which does not depend on $n$, but instead simply balances the bias and the cost of distributional robustness.

If we assume

$$(g1) \qquad g \in [L^\infty, H]_{\beta,\infty} \text{ with norm at most } B \text{ for } \beta \in (0,1) \text{ and } B > 0,$$

we find, from Lemma 6.4.1, that

$$I_c(g, r, W) \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} \tag{6.6.1}$$

for $r > 0$. If we also assume $(H)$, then we find that the regression function $g$ is bounded. Therefore, we simply assume (6.6.1) instead of $(g1)$, in which case $g$ need not be bounded. This gives us the following result.

**Theorem 6.6.3** *Assume (Y1), (H) and* (6.6.1). *Let the cost function* $c : S \times S \to [0, \infty)$ *by*

$$c(x_1, x_2) = \|k_{x_2} + k_{x_1}\|_H \, \|k_{x_2} - k_{x_1}\|_H.$$

*Let* $r > 0$ *and* $W > 0$. *Supposing that the expectation below exists, we have*

$$\mathbb{E}\left(\sup_{Q \in B_c[P,W]} \|\hat{h}_r - g\|_{L^2(Q)}^2\right) \leq \frac{8\|k\|_\infty \sigma r}{n^{1/2}} + \frac{64\|k\|_\infty^2 r^2}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + 8Wr^2.$$

*Let $D_1 > 0$. Setting*

$$r = D_1 B W^{-(1-\beta)/2}$$

*makes the right-hand side of the inequality equal to*

$$D_2 \|k\|_\infty \sigma B W^{-(1-\beta)/2} n^{-1/2} + D_3 \|k\|_\infty^2 B^2 W^{-(1-\beta)} n^{-1/2} + D_4 B^2 W^\beta$$

*for constants $D_2, D_3, D_4 > 0$ depending only on $D_1$ and $\beta$.*

**Proof** The initial bound follows from Theorem 6.6.2 and (6.6.1). The next bound follows with

$$D_2 = 8D_1,\ D_3 = 64D_1^2 \text{ and } D_4 = 10D_1^{-2\beta/(1-\beta)} + 8D_1^2.$$

■

Minimising the last two terms of the initial bound in Theorem 6.6.3 with respect to $r$ gives

$$r = \left(\frac{5\beta}{4(1-\beta)}\right)^{(1-\beta)/2} BW^{-(1-\beta)/2}.$$

In particular, $r$ is of the form in Theorem 6.6.3. Larger values of $W$ give a smaller value of $r$, especially for small $\beta$. This means that if we demand robustness against larger sets of covariate distributions, we must select a smoother estimator. Note that the last term of the later bound in Theorem 6.6.3, which does not tend to 0 as $n \to \infty$, increases as $W$ increases. The expected worst-case squared $L^2$ error increases as we demand more distributional robustness.

Although the optimal transport problem defined above is the most natural for the analysis of the covariate shift problem, we can also perform the analysis by using an optimal transport problem involving a more recognisable cost function. We have the following result.

**Lemma 6.6.4** *Assume (H). Let the cost function $c : S \times S \to [0, \infty)$ by*

$$c(x_1, x_2) = \|k_{x_2} - k_{x_1}\|_H^2.$$

*For $Q \in \mathcal{P}(S)$, we have*

$$\|\hat{h}_r - h_r\|_{L^2(Q)}^2 \leq 2\|\hat{h}_r - h_r\|_{L^2(P)}^2 + 8r^2 W_c(P, Q)$$

*for all $r > 0$.*

**Proof** Let $x_1, x_2 \in S$. By the reproducing kernel property and the Cauchy–Schwarz inequality, we have

$$\begin{aligned} |(\hat{h}_r - h_r)(x_2) - (\hat{h}_r - h_r)(x_1)| &= |\langle \hat{h}_r - h_r, k_{x_2} - k_{x_1} \rangle_H| \\ &\leq \|\hat{h}_r - h_r\|_H \; \|k_{x_2} - k_{x_1}\|_H \\ &\leq 2r\|k_{x_2} - k_{x_1}\|_H. \end{aligned}$$

Hence,

$$(\hat{h}_r - h_r)(x_2) \leq (\hat{h}_r - h_r)(x_1) + 2r\|k_{x_2} - k_{x_1}\|_H$$

and

$$(\hat{h}_r - h_r)(x_2)^2 \leq 2(\hat{h}_r - h_r)(x_1)^2 + 8r^2\|k_{x_2} - k_{x_1}\|_H^2.$$

Integrating over $(x_1, x_2)$ with respect to $\gamma \in \Pi(P, Q)$ gives

$$\|\hat{h}_r - h_r\|_{L^2(Q)}^2 \leq 2\|\hat{h}_r - h_r\|_{L^2(P)}^2 + 8r^2 \int \|k_{x_2} - k_{x_1}\|_H^2 \, d\gamma(x_1, x_2).$$

The result follows by taking an infimum over $\gamma \in \Pi(P, Q)$. ■

The cost function in Lemma 6.6.4 is the square of the kernel metric on $S$ for the kernel $k$ (see (4.20) of Steinwart and Christmann, 2008). Let $d_k(x_1, x_2) = \|k_{x_1} - k_{x_2}\|_H$ be the kernel metric on $S$ for the kernel $k$. Then for $\Phi : S \to H$ by $\Phi(x) = k_x$, we have $\|\Phi(x_2) - \Phi(x_1)\|_H = d_k(x_1, x_2)$. In particular, $\Phi$ is continuous on $S$ if we take the metric $d$ on $S$ to be $d = d_k$. In this case, it follows from Lemma 4.33 of Steinwart and Christmann (2008) that $H$ is separable, since $S$ is. Furthermore, the functions in $H$ are measurable on $\mathcal{B}(S)$ by Lemma 4.24 of Steinwart and Christmann (2008).

We can use Lemma 6.6.4 to provide an expectation bound on the squared $L^2(Q)$ error of $\hat{h}_r$.

**Theorem 6.6.5** *Assume (Y1) and (H). Let the cost function $c : S \times S \to [0, \infty)$ by*

$$c(x_1, x_2) = \|k_{x_2} - k_{x_1}\|_H^2.$$

*Supposing that the expectation below exists, we have*

$$\mathbb{E}\left(\sup_{Q \in B_c[P,W]} \|\hat{h}_r - g\|_{L^2(Q)}^2\right) \leq \frac{16\|k\|_\infty \sigma r}{n^{1/2}} + \frac{128\|k\|_\infty^2 r^2}{n^{1/2}} + 18 I_c(g, r, W) + 16 W r^2$$

*for all $r > 0$ and all $W > 0$.*

**Proof** By Lemma 6.6.4, we have

$$\begin{aligned}\|\hat{h}_r - g\|_{L^2(Q)}^2 &\leq 2\|\hat{h}_r - h_r\|_{L^2(Q)}^2 + 2\|h_r - g\|_{L^2(Q)}^2 \\ &\leq 4\|\hat{h}_r - h_r\|_{L^2(P)}^2 + 16 r^2 W_c(P, Q) + 2\|h_r - g\|_{L^2(Q)}^2.\end{aligned}$$

Hence,

$$\sup_{Q \in B_c[P,W]} \|\hat{h}_r - g\|_{L^2(Q)}^2 \leq 4\|\hat{h}_r - h_r\|_{L^2(P)}^2 + 16 W r^2 + \sup_{Q \in B_c[P,W]} 2\|h_r - g\|_{L^2(Q)}^2.$$

Therefore,

$$\mathbb{E}\left(\sup_{Q\in B_c[P,W]}\|\hat{h}_r - g\|^2_{L^2(Q)}\right) \leq \frac{16\|k\|_\infty \sigma r}{n^{1/2}} + \frac{128\|k\|^2_\infty r^2}{n^{1/2}} + 16Wr^2 + \sup_{Q\in B_c[P,W]} 10\|h_r - g\|^2_{L^2(Q)}$$

by Corollary 3.10.4. The result follows by taking an infimum over $h_r \in rB_H$. ■

We again assume (6.6.1) to obtain the following result.

**Theorem 6.6.6** *Assume (Y1), (H) and* (6.6.1)*. Let the cost function* $c : S \times S \to [0,\infty)$ *by*

$$c(x_1, x_2) = \|k_{x_2} - k_{x_1}\|^2_H.$$

*Let* $r > 0$ *and* $W > 0$*. Supposing that the expectation below exists, we have*

$$\mathbb{E}\left(\sup_{Q\in B_c[P,W]}\|\hat{h}_r - g\|^2_{L^2(Q)}\right) \leq \frac{16\|k\|_\infty \sigma r}{n^{1/2}} + \frac{128\|k\|^2_\infty r^2}{n^{1/2}} + \frac{18B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + 16Wr^2.$$

*Let* $D_1 > 0$*. Setting*

$$r = D_1 B W^{-(1-\beta)/2}$$

*makes the right-hand side of the inequality equal to*

$$D_2\|k\|_\infty \sigma B W^{-(1-\beta)/2} n^{-1/2} + D_3\|k\|^2_\infty B^2 W^{-(1-\beta)} n^{-1/2} + D_4 B^2 W^\beta$$

*for constants* $D_2, D_3, D_4 > 0$ *depending only on* $D_1$ *and* $\beta$*.*

**Proof** The initial bound follows from Theorem 6.6.5 and (6.6.1). The next bound follows with

$$D_2 = 16D_1,\ D_3 = 128D_1^2 \text{ and } D_4 = 18D_1^{-2\beta/(1-\beta)} + 16D_1^2.$$

■

Minimising the last two terms of the initial bound in Theorem 6.6.6 with respect to $r$ gives

$$r = \left( \frac{9\beta}{8(1-\beta)} \right)^{(1-\beta)/2} BW^{-(1-\beta)/2}.$$

In particular, $r$ is of the form in Theorem 6.6.6.

### 6.6.2 Bounded Regression Function

We now consider the case in which the regression function is bounded. We assume

$$(g2) \qquad \|g\|_\infty \leq C \text{ for } C > 0.$$

We can make $\hat{h}_r$ closer to $g$ by constraining it to lie in the same interval $[-C, C]$. We define the projection $V : \mathbb{R} \to [-C, C]$ by

$$V(t) = \begin{cases} -C & \text{if} \quad t < -C \\ t & \text{if} \quad |t| \leq C \\ C & \text{if} \quad t > C \end{cases}$$

for $t \in \mathbb{R}$. The analysis in this setting requires more care due to the clipping of the estimator. In fact, we are forced to use the analysis in which the cost function of the optimal transport problem is equal to the squared kernel metric.

**Lemma 6.6.7** *Assume (H). Let the cost function $c : S \times S \to [0, \infty)$ by*

$$c(x_1, x_2) = \|k_{x_2} - k_{x_1}\|_H^2.$$

*For $Q \in \mathcal{P}(S)$, we have*

$$\|V\hat{h}_r - Vh_r\|_{L^2(Q)}^2 \leq 2\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + 8r^2 W_c(P, Q)$$

*for all* $r > 0$.

**Proof** Let $x_1, x_2 \in S$. By the reproducing kernel property and the Cauchy–Schwarz inequality, we have

$$\begin{aligned} |(V\hat{h}_r - Vh_r)(x_2) - (V\hat{h}_r - Vh_r)(x_1)| &\leq |V\hat{h}_r(x_2) - V\hat{h}_r(x_1)| + |Vh_r(x_2) - Vh_r(x_1)| \\ &\leq |\hat{h}_r(x_2) - \hat{h}_r(x_1)| + |h_r(x_2) - h_r(x_1)| \\ &= |\langle \hat{h}_r, k_{x_2} - k_{x_1} \rangle_H| + |\langle h_r, k_{x_2} - k_{x_1} \rangle_H| \\ &\leq \|\hat{h}_r\|_H \, \|k_{x_2} - k_{x_1}\|_H + \|h_r\|_H \, \|k_{x_2} - k_{x_1}\|_H \\ &\leq 2r\|k_{x_2} - k_{x_1}\|_H. \end{aligned}$$

Hence,

$$(V\hat{h}_r - Vh_r)(x_2) \leq (V\hat{h}_r - Vh_r)(x_1) + 2r\|k_{x_2} - k_{x_1}\|_H$$

and

$$(V\hat{h}_r - Vh_r)(x_2)^2 \leq 2(V\hat{h}_r - Vh_r)(x_1)^2 + 8r^2\|k_{x_2} - k_{x_1}\|_H^2.$$

Integrating over $(x_1, x_2)$ with respect to $\gamma \in \Pi(P, Q)$ gives

$$\|V\hat{h}_r - Vh_r\|_{L^2(Q)}^2 \leq 2\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + 8r^2 \int \|k_{x_2} - k_{x_1}\|_H^2 \, d\gamma(x_1, x_2).$$

The result follows by taking an infimum over $\gamma \in \Pi(P, Q)$. ■

We can use Lemma 6.6.7 to provide an expectation bound on the squared $L^2(Q)$ error of $V\hat{h}_r$. In order to understand how well $g$ can be approximated by elements of $H$ for bounded regression functions, we define

$$I_\infty(g, r) = \inf \{\|h_r - g\|_\infty : h_r \in rB_H\}$$

for $r > 0$.

**Theorem 6.6.8** *Assume (Y1), (H) and (g2). Let the cost function $c : S \times S \to [0,\infty)$ by*

$$c(x_1, x_2) = \|k_{x_2} - k_{x_1}\|_H^2.$$

*Supposing that the expectation below exists, we have*

$$\mathbb{E}\left(\sup_{Q \in B_c[P,W]} \|V\hat{h}_r - g\|_{L^2(Q)}^2\right) \le \frac{16\|k\|_\infty(16C+\sigma)r}{n^{1/2}} + 18I_\infty(g,r) + 16Wr^2$$

*for all $r > 0$ and all $W > 0$.*

**Proof** By Lemma 6.6.7, we have

$$\begin{aligned}\|V\hat{h}_r - g\|_{L^2(Q)}^2 &\le 2\|V\hat{h}_r - Vh_r\|_{L^2(Q)}^2 + 2\|Vh_r - g\|_{L^2(Q)}^2 \\ &\le 4\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + 16r^2 W_c(P,Q) + 2\|Vh_r - g\|_\infty^2.\end{aligned}$$

Hence,

$$\sup_{Q \in B_c[P,W]} \|V\hat{h}_r - g\|_{L^2(Q)}^2 \le 4\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + 16Wr^2 + 2\|Vh_r - g\|_\infty^2.$$

Therefore,

$$\mathbb{E}\left(\sup_{Q \in B_c[P,W]} \|V\hat{h}_r - g\|_{L^2(Q)}^2\right) \le \frac{16\|k\|_\infty(16C+\sigma)r}{n^{1/2}} + 16Wr^2 + 18\|h_r - g\|_\infty^2$$

by Corollary 3.11.2. The result follows by taking an infimum over $h_r \in rB_H$. ∎

We now assume $(g1)$ in full. By Lemma 6.4.1, we have that

$$I_\infty(g,r) \le \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} \tag{6.6.2}$$

for $r > 0$. We have the following result.

**Theorem 6.6.9** *Assume (Y1), (H), (g1) and (g2). Let the cost function $c : S \times S \to [0, \infty)$ by*

$$c(x_1, x_2) = \|k_{x_2} - k_{x_1}\|_H^2.$$

*Let $r > 0$ and $W > 0$. Supposing that the expectation below exists, we have*

$$\mathbb{E}\left(\sup_{Q \in B_c[P,W]} \|V\hat{h}_r - g\|_{L^2(Q)}^2\right) \leq \frac{16\|k\|_\infty(16C+\sigma)r}{n^{1/2}} + \frac{18B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + 16Wr^2.$$

*Let $D_1 > 0$. Setting*

$$r = D_1 B W^{-(1-\beta)/2}$$

*makes the right-hand side of the inequality equal to*

$$D_2\|k\|_\infty(16C+\sigma)BW^{-(1-\beta)/2}n^{-1/2} + D_3B^2W^\beta$$

*for constants $D_2, D_3 > 0$ depending only on $D_1$ and $\beta$.*

**Proof** The initial bound follows from Theorem 6.6.8 and (6.6.2). The next bound follows with

$$D_2 = 16D_1 \text{ and } D_3 = 18D_1^{-2\beta/(1-\beta)} + 16D_1^2.$$

■

Minimising the last two terms of the initial bound in Theorem 6.6.9 with respect to $r$ again gives

$$r = \left(\frac{9\beta}{8(1-\beta)}\right)^{(1-\beta)/2} BW^{-(1-\beta)/2}.$$

When the regression function is bounded, we can also obtain high-probability bounds on the squared $L^2(Q)$ error of $V\hat{h}_r$. However, in order for the high-probability bounds to hold, we must make an additional assumption on the errors of the response variables. Let $U$ and $V$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say $U$ is $\sigma^2$-subgaussian if

$$\mathbb{E}(\exp(tU)) \leq \exp(\sigma^2 t^2/2)$$

for all $t \in \mathbb{R}$. We say $U$ is $\sigma^2$-subgaussian given $V$ if

$$\mathbb{E}(\exp(tU)|V) \leq \exp(\sigma^2 t^2/2)$$

almost surely for all $t \in \mathbb{R}$. We assume

$$(Y2) \qquad Y_i - g(X_i) \text{ is } \sigma^2\text{-subgaussian given } X_i \text{ for } 1 \leq i \leq n.$$

This assumption is stronger than $(Y1)$.

**Theorem 6.6.10** *Assume (Y2), (H) and (g2). Let the cost function $c : S \times S \to [0, \infty)$ by*

$$c(x_1, x_2) = \|k_{x_2} - k_{x_1}\|_H^2.$$

*Let $t > 0$. There exists a measurable set with probability at least $1 - 3e^{-t}$ on which*

$$\sup_{Q \in B_c[P,W]} \|V\hat{h}_r - g\|^2_{L^2(Q)}$$

*is at most*

$$\frac{16 \left(2C^2 + 8\|k\|_\infty^{1/2} C^{3/2} r^{1/2} + \|k\|_\infty (16C + 5\sigma) r\right) t^{1/2}}{n^{1/2}} + \frac{32C^2 t}{3n} + 18 I_\infty(g, r) + 16 W r^2$$

*for all $r > 0$ and all $W > 0$.*

**Proof** By Lemma 6.6.7, we have

$$\begin{aligned}
\|V\hat{h}_r - g\|^2_{L^2(Q)} &\leq 2\|V\hat{h}_r - Vh_r\|^2_{L^2(Q)} + 2\|Vh_r - g\|^2_{L^2(Q)} \\
&\leq 4\|V\hat{h}_r - Vh_r\|^2_{L^2(P)} + 16 r^2 W_c(P, Q) + 2\|Vh_r - g\|^2_\infty.
\end{aligned}$$

Hence,

$$\sup_{Q \in B_c[P,W]} \|V\hat{h}_r - g\|^2_{L^2(Q)} \leq 4\|V\hat{h}_r - Vh_r\|^2_{L^2(P)} + 16Wr^2 + 2\|Vh_r - g\|^2_\infty.$$

Therefore, there exists a measurable set with probability at least $1 - 3e^{-t}$ on which

$$\sup_{Q \in B_c[P,W]} \|V\hat{h}_r - g\|^2_{L^2(Q)}$$

is at most

$$\frac{16\left(2C^2 + 8\|k\|^{1/2}_\infty C^{3/2} r^{1/2} + \|k\|_\infty(16C + 5\sigma)r\right) t^{1/2}}{n^{1/2}} + \frac{32C^2 t}{3n} + 16Wr^2 + 18\|h_r - g\|^2_\infty$$

by Corollary 3.13.4. Taking a sequence of $h_{r,n} \in rB_H$ for $n \geq 1$ with

$$\|h_{r,n} - g\|^2_\infty \downarrow I_\infty(g, r)$$

as $n \to \infty$ proves the result. ■

We again assume $(g1)$ to obtain the following result.

**Theorem 6.6.11** *Assume (Y2), (H), (g1) and (g2). Let the cost function $c : S \times S \to [0, \infty)$ by*

$$c(x_1, x_2) = \|k_{x_2} - k_{x_1}\|^2_H.$$

*Let $r > 0$, $W > 0$ and $t > 0$. There exists a measurable set with probability at least $1 - 3e^{-t}$ on which*

$$\sup_{Q \in B_c[P,W]} \|V\hat{h}_r - g\|^2_{L^2(Q)}$$

*is at most*

$$\frac{16\left(2C^2 + 8\|k\|^{1/2}_\infty C^{3/2} r^{1/2} + \|k\|_\infty(16C + 5\sigma)r\right) t^{1/2}}{n^{1/2}} + \frac{32C^2 t}{3n} + \frac{18B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + 16Wr^2.$$

*Let $D_1 > 0$. Setting*

$$r = D_1 B W^{-(1-\beta)/2}$$

*makes the right-hand side of the inequality equal to*

$$\begin{aligned} &D_2 C^2 t^{1/2} n^{-1/2} + D_3 \|k\|_\infty^{1/2} C^{3/2} B^{1/2} W^{-(1-\beta)/4} t^{1/2} n^{-1/2} \\ &+ D_4 \|k\|_\infty (16C + 5\sigma) B W^{-(1-\beta)/2} t^{1/2} n^{-1/2} + D_5 C^2 t n^{-1} + D_6 B^2 W^\beta \end{aligned}$$

*for constants $D_2, D_3, D_4, D_5, D_6 > 0$ depending only on $D_1$ and $\beta$.*

**Proof** The initial bound follows from Theorem 6.6.10 and (6.6.2). The next bound follows with

$$D_2 = 32,\ D_3 = 128 D_1^{1/2},\ D_4 = 16 D_1\ D_5 = 32/3 \text{ and } D_6 = 18 D_1^{-2\beta/(1-\beta)} + 16 D_1^2.$$

■

Minimising the last two terms of the initial bound in Theorem 6.6.11 with respect to $r$ again gives

$$r = \left(\frac{9\beta}{8(1-\beta)}\right)^{(1-\beta)/2} B W^{-(1-\beta)/2}.$$

## 6.7 Alternative Estimators

In this section, we consider estimators of the regression function $g$ other than $\hat{h}_r$ for $r > 0$. Consider an estimator $\hat{g}$ of $g$. For $W > 0$, we are interested in $\hat{g}$ such that

$$\sup_{Q \in B_c[P,W]} \|\hat{g} - g\|_{L^2(Q)}^2 \tag{6.7.1}$$

is small. Hence, we consider an estimator which minimises an empirical version of this quantity. However, there are many difficulties with this approach. Firstly, our responses $Y_i$ for $1 \leq i \leq n$ are only given at a finite number of covariates $X_i$ for $1 \leq i \leq n$. However, to define an empirical version of (6.7.1) we need an empirical version of the regression function $g$ at each point $x \in S$. Suppose that we have access to a stochastic process $\tilde{Y}$ on $S$ such that $Y_i = \tilde{Y}(X_i)$ for $1 \leq i \leq n$ and that $\tilde{Y}$ has mean function $g$, by which we mean $\mathbb{E}(\tilde{Y}(x)) = g(x)$ for all $x \in S$. We can then define an empirical version of (6.7.1) by

$$\sup_{Q \in B_c[P_n, W_n]} \|\hat{g} - \tilde{Y}\|^2_{L^2(Q)} \tag{6.7.2}$$

for $W_n > 0$. Here, $P_n$ is the empirical distribution of the $X_i$ for $1 \leq i \leq n$.

As with most nonparametric estimation procedures, we need to take steps to ensure that overfitting of our estimator to our data does not occur. This can achieved, for example, by using Ivanov regularisation. In this case, we minimise (6.7.2) subject to the constraint that the estimator $\hat{g}$ lies in $rB_H$ for $r > 0$. We refer to this estimator as $\hat{g}_r$. We obtain

$$\sup_{Q \in B_c[P_n, W_n]} \|\hat{g}_r - \tilde{Y}\|^2_{L^2(Q)} \leq \sup_{Q \in B_c[P_n, W_n]} \|h_r - \tilde{Y}\|^2_{L^2(Q)}$$

for all $h_r \in rB_H$. If we define

$$Z = \sup_{f \in rB_H} \left| \sup_{Q \in B_c[P_n, W_n]} \|f - \tilde{Y}\|^2_{L^2(Q)} - \sup_{Q \in B_c[P_n, W_n]} \|f - g\|^2_{L^2(Q)} \right|, \tag{6.7.3}$$

then we find

$$\sup_{Q \in B_c[P_n, W_n]} \|\hat{g}_r - g\|^2_{L^2(Q)} \leq \sup_{Q \in B_c[P_n, W_n]} \|h_r - g\|^2_{L^2(Q)} + 2Z.$$

However, bounding $Z$ is incredibly difficult in general.

Even in situations in which bounding $Z$ is possible, we still need to change the balls of probability measures with centre $P_n$ in the above expression so that they have centre $P$. An important step in achieving this aim is to produce a bound $W_c(P_n, P) \leq \varepsilon_n$ with high probability. For example, when $S \subseteq \mathbb{R}^d$ and $c(x_1, x_2) = \|x_2 - x_1\|_2^p$ for $x_1, x_2 \in S$ and $p \in [1, \infty)$, Theorem 2 of Fournier and Guillin (2015) shows that $W_c(P_n, P)$ is of order $n^{-1/2}$ if $p > d/2$ for sufficiently concentrated probability measures $P$.

In general, we still need $W_c$ to satisfy additional properties in order to centre the balls at $P$. A sufficient condition is that $c = d^p$ the metric on $S$ to the power $p$ for $p \in [1, \infty)$. Note that $W_c$ is symmetric in this case, and $W_c^{1/p}$ satisfies the triangle inequality by Definition 6.1 of Villani (2009). The definition of the Wasserstein distance in Villani (2009) is our definition to the power $1/p$. Combining this with $W_c(P_n, P) \leq \varepsilon_n$ gives

$$B_c[P, W] \subseteq B_c[P_n, (W^{1/p} + \varepsilon_n^{1/p})^p] \subseteq B_c[P, (W^{1/p} + 2\varepsilon_n^{1/p})^p].$$

Letting $W_n = (W^{1/p} + \varepsilon_n^{1/p})^p$ shows that

$$\sup_{Q \in B_c[P,W]} \|\hat{g}_r - g\|_{L^2(Q)}^2 \leq \sup_{Q \in B_c[P,(W^{1/p}+2\varepsilon_n^{1/p})^p]} \|h_r - g\|_{L^2(Q)}^2 + 2Z.$$

This bounds (6.7.1) for the estimator $\hat{g}_r$. We could continue, for example, by replacing the $L^2(Q)$ norm on the right-hand side with the $L^\infty$ norm and taking an infimum over $h_r \in rB_H$ to obtain

$$\sup_{Q \in B_c[P,W]} \|\hat{g}_r - g\|_{L^2(Q)}^2 \leq I_\infty(g, r) + 2Z.$$

We could even assume $(g1)$ in order to bound $I_\infty(g, r)$. However, the real challenge is to bound $Z$ in (6.7.3), which there is no clear way of doing. Additionally, recall that we are assuming that we have access to the stochastic process $\tilde{Y}$ used to define (6.7.1).

### 6.7.1 Estimator Computation

As well as the difficulties in the analysis of $\hat{g}_r$, there are also challenges in its computation. In order to consider this problem, we need to define some new concepts. Let $T$ be a compact Hausdorff space. We let $(M(T), \|\cdot\|_{\mathrm{TV}})$ be the Banach space of finite signed Borel measures on $T$ equipped with the total variation norm. We also let $M_R(T)$ be the subspace of $M(T)$ consisting of the regular finite signed Borel measures on $T$ and $\mathcal{P}_R(T)$ consist of the regular Borel probability measures on $T$. Let $(C(T), \|\cdot\|_\infty)$ be the Banach space of continuous real-valued functions on $T$ equipped with the supremum norm and let $C(T)^*$ be the dual of $C(T)$. By the Riesz representation theorem, we have that $C(T)^*$ is isometrically isomorphic to $M_R(T)$. This can be seen by considering Theorem 6.19 of Rudin (1987) for functionals which take real values on real-valued functions. Hence, for a sequence $\mu_n \in M_R(T)$ for $n \geq 1$ and a point $\mu \in M_R(T)$, we have that $\mu_n \to \mu$ weak-* as $n \to \infty$ if

$$\int f d\mu_n \to \int f d\mu$$

as $n \to \infty$ for all $f \in C(T)$. This form of convergence is often referred to as weak convergence in probability theory. However, this is the name of a different form of convergence in functional analysis, so we refer to it using the name weak-* convergence from functional analysis to avoid confusion.

**Lemma 6.7.1** *Let $T$ be a compact Haussdorf space and $f \in C(T)$. Suppose that*

*$A \subseteq \mathcal{P}_R(T)$ is weak-* closed and convex. Then $L : A \to \mathbb{R}$ by*

$$L(Q) = \int f \, dQ$$

*attains its maximum value on* $\operatorname{ext}(A)$.

**Proof** By the Riesz representation theorem, we have that $C(T)^*$ is isometrically isomorphic to $M_R(T)$. Furthermore, by the Banach–Alaoglu theorem (Theorem 3.15 of Rudin, 1991), we have that the closed unit ball $B_{M_R(T)} = \{\mu \in M_R(T) : \|\mu\|_{\mathrm{TV}} \leq 1\}$ of $M_R(T)$ is weak-* compact. Let $C_+(T)$ be the subset of $C(T)$ consisting of the positive continuous functions on $T$. The subset $M_{R,+}(T)$ of $M_R(T)$ consisting of regular finite positive Borel measures on $T$ can be written as

$$M_{R,+}(T) = \left\{ \mu \in M_R(T) : \int f \, d\mu \geq 0 \text{ for all } f \in C_+(T) \right\},$$

so it is weak-* closed. Furthermore,

$$U = \left\{ \mu \in M_R(T) : \int 1 \, d\mu = 1 \right\}$$

is weak* closed, so $\mathcal{P}_R(T) = B_{M_R(T)} \cap M_{R,+}(T) \cap U$ is weak-* compact.

By assumption, $A \subseteq \mathcal{P}_R(T)$ is weak-* closed, so weak-* compact. It is also convex. Since the weak-* topology on $M_R(T)$ is induced by the collection of seminorms

$$\|\mu\|_f = \left| \int f \, d\mu \right|$$

for $f \in C(T)$, we have that $M_R(T)$ is weak-* locally convex. By the Choquet–Bishop–de Leeuw theorem (Theorem 5.6 of Bishop and de Leeuw, 1959), for all $Q \in A$ there exists a probability measure $w_Q$ on the sigma-algebra generated by the weak-* Borel sets of $A$ and $\text{ext}(A)$ such that

$$G(Q) = \int G \, dw_Q$$

for all $G$ in the weak-* dual of $M_R(S)$ and $w_Q(\text{ext}(A)) = 1$. In particular,

$$L(Q) = \int L \, dw_Q.$$

Since $L$ is weak-* continuous on $A$, which is weak-* compact, there is some $Q \in A$ at which $L$ attains its maximum. For this $Q$, we have $L(Q) - L(\tilde{Q}) \geq 0$ for all $\tilde{Q} \in A$ and

$$\int (L(Q) - L(\tilde{Q})) \, dw_Q(\tilde{Q}) = 0.$$

Hence, $L(Q) - L = 0$ $w_Q$-almost surely. Since $w_Q(\text{ext}(A)) = 1$, there is some $\tilde{Q} \in \text{ext}(A)$ for which $L(\tilde{Q}) = L(Q)$, the maximum value of $L$. ■

Recall that we assume that the covariate set $(S, d)$ is a complete separable metric space. In order to apply the above lemma, we also need to assume that $S$ is compact. In this case, all Borel probability measures are regular (Theorem 2.18 of Rudin, 1987), so $\mathcal{P}_R(S) = \mathcal{P}(S)$. We can then apply the above result with $A = B_c[P_n, W_n]$, under conditions on $B_c[P_n, W_n]$, in order to help to compute $\hat{g}_r$ in (6.7.2). Initially, we do not restrict the centre of the ball to be the empirical distribution of the $X_i$ for $1 \leq i \leq n$, so we consider $B_c[P, W]$ in place of $B_c[P_n, W_n]$. In this case, the conditions on $B_c[P, W]$ are that it must be weak-* closed and convex. We know that $B_c[P, W]$ is convex by Lemma 5.3.1, so we are only concerned about whether or not $B_c[P, W]$ is weak-* closed.

One situation in which $B_c[P, W]$ is weak-* closed is when the cost function $c$ is continuous. Suppose that $Q_n \in B_c[P, W]$ for $n \geq 1$, $Q \in \mathcal{P}(S)$ and $Q_n \to Q$ weak-* as

$n \to \infty$. Theorem 4.1 of Villani (2009) shows that there exists $\gamma_n \in \Pi(P, Q_n)$ which attains $W_c(P, Q_n) \leq W$ for $n \geq 1$. Furthermore, Theorem 5.20 of Villani (2009) shows that, for some subsequence $\gamma_{n(k)}$ for $k \geq 1$ of $\gamma_n$, we have that $\gamma_{n(k)} \to \gamma \in \Pi(P, Q)$ weak-* as $k \to \infty$ and that $\gamma$ attains $W_c(P, Q)$. Since $\gamma_{n(k)} \to \gamma$ weak-* as $k \to \infty$ and $c$ is continuous, we have that

$$\begin{aligned} W_c(P, Q) &= \int c \, d\gamma \\ &= \lim_{k \to \infty} \int c \, d\gamma_{n(k)} \\ &= \lim_{k \to \infty} W_c(P, Q_{n(k)}) \\ &\leq W. \end{aligned}$$

It follows that $Q \in B_c[P, W]$. Hence, $B_c[P, W]$ is weak-* closed.

Another case in which $B_c[P, W]$ is weak-* closed is when the cost function $c = d^p$ the metric on $S$ to the power $p$ for $p \in [1, \infty)$. Note that $W_c$ is symmetric in this case. Theorem 6.9 of Villani (2009) shows that for $Q_n \in \mathcal{P}(S)$ for $n \geq 1$ and $Q \in \mathcal{P}(S)$, we have that $Q_n \to Q$ weak-* as $n \to \infty$ if and only if $W_c(Q_n, Q) \to 0$ as $n \to \infty$. We do not need any further conditions because $S$ is compact and hence bounded, so Definition 6.4 of Villani (2009) simply defines $\mathcal{P}(S)$ and condition (iii) in Definition 6.8 of Villani (2009) is automatically satisfied. This is not quite sufficient for $B_c[P, W]$ to be weak-* closed. However, recall that $W_c^{1/p}$ satisfies the triangle inequality by Definition 6.1 of Villani (2009). Together, these two properties show that $B_c[P, W]$ is weak-* closed. Suppose that $Q_n \in B_c[P, W]$ for $n \geq 1$, $Q \in \mathcal{P}(S)$ and $Q_n \to Q$ weak-* as $n \to \infty$. Then $W_c(Q_n, Q) \to 0$ as $n \to \infty$. Furthermore,

$$\begin{aligned} W_c(Q, P)^{1/p} &\leq W_c(Q, Q_n)^{1/p} + W_c(Q_n, P)^{1/p} \\ &\leq W_c(Q, Q_n)^{1/p} + W^{1/p} \end{aligned}$$

$$\to W^{1/p}$$

as $n \to \infty$. It follows that $W_c(P,Q) \leq W$ and $Q \in B_c[P,W]$. Hence, $B_c[P,W]$ is weak-* closed.

Now that we have seen examples in which $B_c[P,W]$ is weak-* closed, we investigate the consequences of Lemma 6.7.1 with $A = B_c[P_n, W_n]$. If we assume that our estimator $\hat{g}$ and our response process $\tilde{Y}$ are continuous for each point $\omega \in \Omega$ our sample space, then

$$\sup_{Q \in B_c[P_n,W_n]} \|\hat{g} - \tilde{Y}\|^2_{L^2(Q)} = \max_{Q \in \operatorname{ext}(B_c[P_n,W_n])} \|\hat{g} - \tilde{Y}\|^2_{L^2(Q)}.$$

Let $\Delta_n(S) = \{Q \in \mathcal{P}(S) : |\operatorname{supp}(Q)| \leq n\}$. If $c$ is lower semicontinuous, then by Theorem 2.3 of Owhadi and Scovel (2017) we have that $\operatorname{ext}(B_c[P_n,W_n]) \subseteq \Delta_{n+2}(S)$. This gives us more information about the set of probability measures over which we have to maximise, but it is still a very difficult problem. Further conditions under which $Q \in B_c[P_n,W_n]$ is an extreme point of $B_c[P_n,W_n]$ are given in Chapter 5. In particular, Section 5.7 discusses the case in which $Q$ has finite support. As mentioned earlier, we need to restrict our choice of estimator $\hat{g}$ to prevent overfitting. One option is to demand that $\hat{g}$ lies in $rB_H$, for example.

### 6.7.2 Regression Function Approximation

Given the problems with trying to define and calculate an estimator using (6.7.2), another approach is to investigate the approximation properties of the regression function instead. Such properties may be useful for defining other estimators. We consider functions $f : S \to \mathbb{R}$ such that

$$\sup_{Q \in B_c[P,W]} \|f - g\|^2_{L^2(Q)}$$

is small. Suppose that $g$ is continuous and $B_c[P, W]$ is weak-* closed. If we demand that $f$ is continuous, then

$$\sup_{Q \in B_c[P,W]} \|f - g\|_{L^2(Q)}^2 = \sup_{Q \in \text{ext}(B_c[P,W])} \|f - g\|_{L^2(Q)}^2$$

by Lemma 6.7.1. We must place restrictions on $f$ so that we do not select $f = g$. For example, we could search for $h_r \in rB_H$ which minimises

$$\sup_{Q \in \text{ext}(B_c[P,W])} \|h_r - g\|_{L^2(Q)}^2.$$

Unfortunately, in general there are no useful characterisations of $\text{ext}(B_c[P, W])$, unlike $\text{ext}(B_c[P_n, W_n]) \subseteq \Delta_{n+2}(S)$. However, some conditions under which $Q \in B_c[P, W]$ is an extreme point of $B_c[P, W]$ are given in Chapter 5.

## 6.8 Discussion

In this chapter, we consider ways of bounding the worst-case squared $L^2$ error of different estimators with respect to a Wasserstein ball of probability measures centred at the original covariate distribution. We begin by providing expectation bounds on this error for Ivanov-regularised least-squares estimators when the regression function is unbounded. We also provide an expectation bound when the regression function is bounded, as well as a high-probability bound in the case in which the errors of the response variables are subgaussian. We then consider alternative estimators based on an empirical version of the worst-case squared $L^2$ error. We examine the problems with both the analysis and computation of these estimators.

Clearly more research into estimators other than the Ivanov-regularised least-squares estimators is needed. However, there are obvious issues with both the analysis and computation of the alternative estimators considered in this chapter. Once some of these obstacles have been overcome, it would be interesting to consider situations in which both the original covariate distribution and the distribution of the response variables are subject to perturbation.

# Chapter 7

# Conclusion

In this thesis, we study kernel least-squares estimators for the regression problem subject to a norm constraint. We bound the squared $L^2(P)$ error of our estimators, where $P$ is the covariate distribution. Furthermore, we provide bounds on the worst-case squared $L^2(Q)$ error over all probability measures $Q$ in a Wasserstein ball centred at $P$. This motivates us to examine the extreme points of Wasserstein balls. We now review the main content of the thesis. We also discuss some directions for further research.

## 7.1 Ivanov-Regularised Least-Squares Estimators over Large RKHSs and Their Interpolation Spaces

In Chapter 3, we show how Ivanov regularisation can be used to produce estimators which have a small squared $L^2(P)$ error. In this setting, we use Ivanov regularisation to bound the reproducing-kernel Hilbert space (RKHS) norm of the estimators. We begin by considering the case in which the regression function lies in an interpolation space between $L^2(P)$ and the RKHS $H$. We assume only that $H$ is separable with a bounded and measurable kernel.

Under the mild assumption that the response variables have bounded variance, we provide an expectation bound on the squared $L^2(P)$ error of our estimator of order $n^{-\beta/2}$. Here, $n$ is the number of data points and $\beta$ parametrises the interpolation space between $L^2(P)$ and $H$ containing the regression function. As far as we are aware, this is the first time an estimator has been analysed in this setting.

If we assume that the regression function is bounded, then we can clip the estimator so that it is closer to the regression function. Specifically, we change the values that the estimator can take so that they are not outside the range of values of the regression function. In this setting, we show that the clipped estimator has an expected squared $L^2(P)$ error of order $n^{-\beta/(1+\beta)}$. This order is the optimal power of $n$. Under the stronger assumption that the response variables have subgaussian errors and that the regression function comes from an interpolation space between $L^\infty$ and $H$, we show that the squared $L^2(P)$ error is of order $n^{-\beta/(1+\beta)}$ with high probability.

When the regression function is bounded, we use training and validation to obtain both expectation bounds and high-probability bounds of the same order of $n^{-\beta/(1+\beta)}$. Training and validation is an adaptive estimation procedure which splits the data set into a training set and a validation set. The training set is used to define a collection of estimators for a range of sizes of norm constraint, while the validation set is used to select a final estimator from this collection. This allows us to select the size of the norm constraint for our Ivanov regularisation without knowing which interpolation space contains the regression function.

Our analysis of the Ivanov-regularised estimators is performed by controlling empirical processes over balls in the RKHS. On the other hand, the analysis of Tikhonov-regularised estimators usually uses the spectral decomposition of the kernel operator. It would be illuminating to analyse our Ivanov-regularised estimators using this method.

## 7.2 The Goldenshluger–Lepski Method for Constrained Least-Squares Estimators over RKHSs

In Chapter 4, we apply a different adaptive estimation procedure called the Goldenshluger–Lepski method to our Ivanov-regularised least-squares estimators. We only consider the case in which the regression function is bounded, so we clip our estimators to make them closer to the regression function. We use all of the data to produce a collection of non-adaptive estimators for different fixed sizes of norm constraint, before performing pairwise comparisons to select a final estimator.

Since the covariate distribution $P$ and the $L^2(P)$ norm are unknown, we use the $L^2(P_n)$ norm when calculating the pairwise comparisons between the non-adaptive estimators. Here, $P_n$ is the empirical distribution of the covariates. The $L^2(P)$ norm is the natural norm in which to perform the pairwise comparisons, as this is the norm in which we seek guarantees on our estimator. However, we still attain these guarantees when using the $L^2(P_n)$ norm for the comparisons.

We create two adaptive procedures. In the first procedure, we fix an RKHS and adapt to the size of the norm constraint. This is similar to our training and validation procedure, as we adapt to the same parameter. As far as we are aware, this is the first time that the Goldenshluger–Lepski method has been applied in the context of RKHS regression. In the second procedure, we consider a collection of RKHSs with Gaussian kernels and adapt to both the size of the norm constraint in the RKHSs and the RKHS itself.

By assuming that the regression function lies in an interpolation space between $L^\infty$ and an RKHS $H$ parametrised by $\beta$, we obtain a bound on a fixed quantile of the squared $L^2(P)$ error of our adaptive estimator of order $n^{-\beta/(1+\beta)}$. This is true for both the procedure in which the RKHS is fixed to be $H$ and the procedure in which $H$ comes from a collection of RKHSs with Gaussian kernels. The order $n^{-\beta/(1+\beta)}$ for the squared $L^2(P)$ error of the adaptive estimators matches the order of the smallest bounds obtained for the non-adaptive estimators in Chapter 3.

We currently demand that the set of width parameters of the Gaussian kernels is bounded for the procedure in which we consider a collection of RKHSs. This is quite limiting. For example, we would be able to estimate a greater collection of functions is we were able to allow the width parameter to tend to 0 as $n$ tends to infinity. Further analysis of this procedure for the case in which the width parameter tends to 0 may produce estimators which can be applied in more general situations.

It would be interesting to investigate whether it is possible to extend the use of the Goldenshluger–Lepski method from the case in which we consider a collection of RKHSs with Gaussian kernels to cases in which we consider other collections of RKHSs. Our analysis of the RKHSs with Gaussian kernels relies on the fact that the closed unit ball of the RKHS generated by a Gaussian kernel increases as the width of the kernel decreases. If another collection of RKHSs also exhibited this nestedness property, then a similar analysis should be possible. If the RKHSs did not exhibit this property, then a new form of analysis would be needed.

## 7.3 Extreme Points of Wasserstein Balls

In Chapter 5, we change direction to study conditions under which probability measures in a Wasserstein ball are extreme points or not extreme points. We show that, under very mild conditions, the only extreme points of Wasserstein balls which do not lie on the surface of the ball are Dirac measures. By the surface of the ball, we mean the points in the ball whose distance from the centre of the ball is equal to the radius.

We then consider points on the surface of the ball. We find that if the Wasserstein distance is uniquely attained by a transport plan induced by a transport map, then the point is an extreme point. On the other hand, under conditions on the centre of the ball and the cost function, if the Wasserstein distance is attained by two distinct transport plans induced by continuous transport maps, then the point is not an extreme point. We then consider the case in which our probability measures are defined on finite sets. We use the solutions to the dual problem to provide conditions under which we do not have an extreme point.

Our results only make full use of the dual problem in the discrete setting. However, it would be useful to apply the dual problem in other settings as well. This would give us other ways of determining conditions under which a point in the ball is an extreme point or not an extreme point. In particular, conditions in terms of the solutions to the dual problem would be of interest.

## 7.4 Optimal Transport for Covariate Shift in RKHS Regression

In Chapter 6, we analyse the worst-case squared $L^2(Q)$ error of different estimators over a Wasserstein ball of probability measures $Q$ centred at the covariate distribution $P$. This ball comprises all perturbations of $P$ of any size up to the radius of the ball. We first provide expectation bounds on the worst-case squared $L^2(Q)$ error for our Ivanov-regularised least-squares estimators when the regression function is unbounded.

We then provide bounds on the worst-case squared $L^2(Q)$ error when the regression function is bounded. We clip the Ivanov-regularised least-squares estimators so that they are closer to the regression function. We also provide high-probability bounds in this setting under the assumption that the errors of the response variables are subgaussian. We conclude by considering problems with the analysis and computation of alternative estimators based on an empirical version of the worst-case squared $L^2(Q)$ error.

It would be interesting to obtain bounds on the worst-case squared $L^2(Q)$ error for estimators other than the Ivanov-regularised least-squares estimators. We also need to be able to compute such estimators. Neither of these two aims seem to be achievable for the alternative estimators considered in Chapter 6. We could also investigate situations in which both the covariate distribution $P$ and the distribution of the response variables are subject to perturbation.

# Bibliography

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 2312–2320, 2011.

Jöran Bergh and Jörgen Löfström. *Interpolation Spaces. An Introduction.* Springer–Verlag, Berlin–New York, 1976.

Lucien Birgé. An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics*, volume 36 of *IMS Lecture Notes Monogr. Ser.*, pages 113–133. Inst. Math. Statist., Beachwood, OH, 2001.

Errett Bishop and Karel de Leeuw. The representations of linear functionals by measures on sets of extreme points. *Ann. Inst. Fourier*, 9:305–331, 1959.

Andrea Caponnetto and Ernesto de Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

Ernesto De Vito, Sergei Pereverzyev, and Lorenzo Rosasco. Adaptive kernel methods using the balancing principle. *Found. Comput. Math.*, 10(4):455–479, 2010.

Eustasio del Barrio, Juan A Cuesta-Albertos, Carlos Matrán, and Jesús M Rodríguez-Rodríguez. Tests of goodness of fit based on the $L^2$-wasserstein distance. *Annals of Statistics*, 27(4):1230–1239, 1999.

Mona Eberts and Ingo Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Stat.*, 7, 2013.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint arXiv:1702.07254*, 2017.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738, 2015.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, New York, 2016.

Alexander Goldenshluger and Oleg Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008.

Alexander Goldenshluger and Oleg Lepski. Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields*, 143(1-2):41–71, 2009.

Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39 (3):1608–1632, 2011.

Alexander Goldenshluger and Oleg Lepski. General selection rule from a family of linear estimators. *Theory Probab. Appl.*, 57(2):209–226, 2013.

Antonio Irpino and Rosanna Verde. Dynamic clustering of interval data using a wasserstein-based distance. *Pattern Recognition Letters*, 29(11):1648–1658, 2008.

L. Kantorovitch. On the translocation of masses. *Management Sci.*, 5:1–4, 1958.

Oleg Lepski. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.*, 36(4), 1991a.

Oleg Lepski. On a problem of adaptive estimation in gaussian white noise. *Theory Probab. Appl.*, 35(3):454–466, 1991b.

Oleg Lepski. Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.*, 37(3):433–448, 1993.

Shuai Lu, Peter Mathé, and Sergei V Pereverzev. Balancing principle in supervised learning for a general regularization scheme. *Appl. Comput. Harmon. Anal.*, 2018.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *Ann. Statist.*, 38(1):526–565, 2010.

Luca Oneto, Sandro Ridella, and Davide Anguita. Tikhonov, Ivanov and Morozov regularization for support vector machine learning. *Mach. Learn.*, 103(1):103–136, 2016.

Houman Owhadi and Clint Scovel. Extreme points of a ball about a measure with finite support. *Commun. Math. Sci.*, 15(1):77–96, 2017.

Yamilet Quintana and José M. Rodríguez. Measurable diagonalization of positive definite matrices. *J. Approx. Theory*, 185:91–97, 2014.

Malempati Madhusudana Rao and Zhong Dao Ren. *Theory of Orlicz Spaces*. Marcel Dekker, New York, 1991.

Mark Rudelson and Roman Vershynin. Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.*, 18, 2013.

Walter Rudin. *Real and Complex Analysis.* McGraw–Hill Book Company, New York, third edition, 1987.

Walter Rudin. *Functional Analysis.* McGraw–Hill, Inc., New York, second edition, 1991.

L. Rüschendorf. Optimal solutions of multivariate coupling problems. *Appl. Math. (Warsaw)*, 23(3):325–338, 1995.

L. Rüschendorf and S. T. Rachev. A characterization of random variables with minimum $L^2$-distance. *J. Multivariate Anal.*, 32(1):48–54, 1990.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference*, 90(2):227–244, 2000.

Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Anal. Appl. (Singap.)*, 1(1):17–41, 2003.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

Bharath Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines.* Springer–Verlag, New York, 2008.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constr. Approx.*, 35(3):363–417, 2012.

Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *The 22nd Conference on Learning Theory*, 2009.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Ann. Inst. Statist. Math.*, 60(4):699–746, 2008.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer–Verlag, New York, 1996.

Cédric Villani. *Optimal Transport. Old and New*. Springer–Verlag, Berlin–New York, 2009.

David Williams. *Probability with Martingales.* Cambridge University Press, Cambridge, 1991.