The Pennsylvania State University
The Graduate School

MEAN SHRINKAGE ESTIMATORS IN HILBERT SPACES

A Thesis in
Statistics
by
Nikolas Siapoutis

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2019

The thesis of Nikolas Siapoutis was reviewed and approved* by the following:

Bharath K. Sriperumbudur
Assistant Professor of Statistics
Thesis Advisor

Donald Richards
Professor of Statistics

Ephraim Hanks
Associate Professor of Statistics
Chair of Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

In many statistical algorithms and inferential methods, the estimation of the mean function in a Hilbert space plays an important role. Examples of such algorithms and methods are principal component analysis, discriminant analysis and hypothesis testing. Mean function is often estimated by the well-known empirical average. Motivated by the Stein phenomenon we propose, in this thesis, shrinkage estimators and show them to be improved versions of the empirical average by providing oracle inequalities. We also show that the rate of convergence of these shrinkage estimators is of order $n^{-1/2}$ and it is optimal in the minimax sense. Specifically, we establish a minimax optimal rate over the class of discrete and infinitely differentiable probability measures.

Table of Contents

Acknowledgments

I would like to express my deep gratitude to my advisor Professor Bharath K. Sriperumbudur. His valuable feedback, guidance, patience and support have been crucial in completing this thesis. I sincerely appreciate all the time and effort you spent. Thank you! I would also like to thank Professor Donald Richards for his insightful and useful comments as well as the chair of graduate studies, Professor Ephraim Hanks, for his help .

Last but not least, I am truly grateful to my family and my friends for their encouragement, support and understanding. Thank you for being part of this beautiful journey!

Dedication

To my family.

Chapter 1

Introduction

Estimating the mean function in a Hilbert space is considered a central problem in statistics due to the fact that in practice many inferential methods and statistical algorithms heavily depend on the mean function. The estimation of the mean is critical in some very classical inferential methods, like constructing confident intervals and hypothesis testing. The estimation of the mean also plays a crucial role in some statistical machine learning algorithms such as principal component analysis (PCA) (Jolliffe, 1986) and factor analysis (Cattell, 1952 and Fruchter, 1954), two popular statistical methodologies for dimensionality reduction. In some classification methods, such as linear and quadratic discriminant analysis (McLachlan, 1992), the decision rule depends on the mean and covariance functions. The mean and covariance functions values are not known in real world scenarios, so we need to estimate them. In clustering methods, such as the k-means algorithm (MacQueen, 1967, Hartigan and Wong, 1979), again we need to calculate the mean of the centroids of the cluster. In all these cases, mean estimators are of great importance.

Let $\mathcal{X}$ denote a measurable space and $\mathcal{H}$ a separable Hilbert space. Suppose $X_1, X_2, \ldots, X_n$ is a sample of independent and identical distributed variables from a probability measure $\mathbb{P}$ defined over $\mathcal{X}$ and $r : \mathcal{X} \to \mathcal{H}$ is a continuous function such that $\int_{\mathcal{X}} \|r(x)\|_{\mathcal{H}} \, d\mathbb{P}(x) < \infty$. Our goal is to estimate the mean function given by the following Bochner integral (Dinculeanu, 2000, Diestel and Uhl, 1977)

$$\mu = \int_{\mathcal{X}} r(x) \, d\mathbb{P}(x) \in \mathcal{H},$$

based on the sample $X_1, X_2, \ldots, X_n$. Note that $r : \mathcal{X} \to \mathcal{H}$ is an $\mathcal{H}$-valued measurable function since r is a continuous function and $\mathcal{H}$ is a separable Hilbert space (Steinwart and Christmann, 2008). A commonly-used estimator of μ is the empirical mean, given by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} r(X_i). \tag{1.1}$$

Despite the nice properties and simplicity of the empirical average, we aim, in this thesis, to improve upon the empirical estimator in (1.1), i.e., to show that there exist estimators that can improve upon $\hat{\mu}$.

Our proposed mean estimator is motivated by an idea of Stein (1955) the Stein phenomenon or paradox (Efron and Morris, 1975). Stein showed that in the case of the multivariate normal distribution with dimension $d \geq 3$, there exist a class of minimax estimators that dominate the standard empirical estimator under squared error loss. A common approach to improving the estimation of the mean is by shrinking. Specifically, we consider an estimation of the mean by using a convex combination of the standard estimator and an arbitrary function. This certain type of estimators belong to a general class of estimators, the shrinkage estimators (Gruber, 1998). Special case of the shrinkage estimators is the well-known James-Stein estimator (James and Stein, 1961). A lot of work has been done in parametric cases, i.e., a parametric form is assumed for the probability distribution $\mathbb{P}$, e.g., multivariate normal and non-normal distributions (Brown 1965, 1966, Kiefer 1957, Brandwein and Strawderman 1990, 2000). However, little research has been done in the non-parametric setting, where no form for the probability distribution $\mathbb{P}$ is assumed. In this thesis, we propose some shrinkage estimators, based on the Stein phenomenon, for the non-parametric case.

Related research has been done for estimating the mean function and covariance operator in the special case in which $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS). Muandet et al. (2014, 2016) proposed shrinkage estimators of the mean that are $\sqrt{n}$-consistent. Interestingly, these estimators outperform the empirical average and are also optimal in the minimax sense over suitable classes of probability measures. They showed that the minimax rate is independent of the smoothness of the

kernel and the density of $\mathbb{P}$. Influenced by the previous work, Yang et al. (2019) proposed a class of optimal shrinkage estimators for the covariance operator and obtained desirable theoretical properties similar to the case of the mean estimation in the RKHS. In this thesis, we propose some optimal shrinkage estimators for the mean function in the case of general Hilbert spaces.

The rest of the thesis is organized as follows. In Chapter 2 we present some definitions and notations that we use throughout the thesis. In Chapter 3, we construct a large class of mean estimators that achieve lower risk than the empirical estimator. Also, we provide an alternative approach to developing shrinkage estimators based on a regression point of view. Finally, in Chapter 4 we establish the minimax rates for the estimators over certain classes of probability measures.

Chapter 2

Definitions and Notations

We begin by reviewing some basic Banach and Hilbert space theory (Kreyszig, 1989 and Rudin, 1987).

Definition 2.1 (Norm) Let $\mathcal{H}$ be a real vector space. A function $\|\cdot\|_{\mathcal{H}} : \mathcal{H} \to [0, \infty]$ is said to be a *norm* on $\mathcal{H}$ if

1. $\|f\|_{\mathcal{H}} = 0$ if and only if $f = 0$,
2. $\|\lambda f\|_{\mathcal{H}} = |\lambda| \, \|f\|_{\mathcal{H}}$, $\forall \lambda \in \mathbb{R}$, $\forall f \in \mathcal{H}$,
3. $\|f + g\|_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} + \|g\|_{\mathcal{H}}$, $\forall f, g \in \mathcal{H}$.

The above definition induces a distance between two elements in the normed vector space $\mathcal{H}$ and thus it is reasonable to formulate the following definitions.

Definition 2.2 (Convergent Sequence) A sequence of elements $\{f_n\}_{n=1}^{\infty}$ in the normed vector space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ is said to *converge* to $f \in \mathcal{H}$ if for every $\epsilon > 0$, there exists $n_\epsilon \in \mathbb{N}$, such that for all $n \geq n_\epsilon$, $\|f_n - f\|_{\mathcal{H}} < \epsilon$.

Definition 2.3 (Cauchy Sequence) A sequence of elements $\{f_n\}_{n=1}^{\infty}$ in the normed vector space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ is said to be a *Cauchy sequence* if for every $\epsilon > 0$, there

exists $n_\epsilon \in \mathbb{N}$, such that for all $n, k \geq n_\epsilon$, $\|f_n - f_k\|_{\mathcal{H}} < \epsilon$.

It is clear that every convergent sequence is a Cauchy sequence. However, not every Cauchy sequence converges. Therefore, we say that a space $\mathcal{H}$ is *complete* if every Cauchy sequence in $\mathcal{H}$ converges, i.e., it has a limit and this limit is in $\mathcal{H}$.

Definition 2.4 (Banach Space) *Banach space* is a complete normed space.

Below we define the notion of inner product which gives an additional structure on a Banach space.

Definition 2.5 (Inner Product) Let $\mathcal{H}$ be a real vector space. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to be an *inner product* on $\mathcal{H}$ if

1. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$, $\forall f \in \mathcal{H}$,
2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$, $\forall f, g \in \mathcal{H}$,
3. $\langle \alpha f_1 + \beta f_2 \rangle_{\mathcal{H}} = \alpha \langle f_1, g \rangle_{\mathcal{H}} + \beta \langle f_2, g \rangle_{\mathcal{H}}$, $\forall \alpha, \beta \in \mathbb{R}$, $\forall f_1, f_2, g \in \mathcal{H}$.

A vector space with an inner product is said to be an *inner product space*. Also, every inner product determines a norm by $\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}^{1/2}$, $f \in \mathcal{H}$.

Definition 2.6 (Hilbert Space) *Hilbert space* is a complete inner product space, i.e., a Banach space with an inner product.

Examples of Hilbert space include $L^2(\mathbb{R})$, i.e., the space of square-integrable functions with respect to the Lebesgue measure on the real line and $l^2(\mathbb{N})$, i.e., the space of square-summable sequence.

Furthermore, a Hilbert space $\mathcal{H}$ is said to be *separable* if it contains a countable and dense subset.

Let $M_b(\mathbb{R}^d)$ denote the set of all finite Borel measures on $\mathbb{R}^d$. For $\mu \in M_b(\mathbb{R}^d)$, $L^r(\mathbb{R}^d, \mu)$ denotes the Hilbert space of r-power ($r \geq 1$) μ-integrable functions and we use $L^r(\mathbb{R}^d)$ when μ is a Lebesgue measure on $\mathbb{R}^d$.

Definition 2.7 (Fourier Transform) The *Fourier transforms* of $f \in L^1(\mathbb{R}^d)$ and $\mu \in M_b(\mathbb{R}^d)$ are defined as

$$f^\wedge(y) := \mathcal{F}[f](y) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{-i\langle y,x\rangle}\, dx, \;\; y \in \mathbb{R},$$

$$\mu^\wedge(y) := \mathcal{F}[\mu](y) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\langle y,x\rangle}\, d\mu(x), \;\; y \in \mathbb{R},$$

where i denotes the imaginary unit $\sqrt{-1}$.

Finally, we present some standard notations for comparing the limiting behavior of different sequences. Let $\alpha_1, \alpha_2, \ldots$ and $\beta_1, \beta_2, \ldots$ be sequences of real numbers. We write $\alpha_n = o(\beta_n)$ as $n \to \infty$ if $\alpha_n/\beta_n \to 0$ as $n \to \infty$. Also, we write $\alpha_n = O(\beta_n)$ as $n \to \infty$ if there exist $M, N > 0$ such that $|\alpha_n/\beta_n| < M$ for all $n > N$.

There are probabilistic analogues of the o and O notation that are relevant to sequences of random variables. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be sequences of random variables. We write $X_n = o_{\mathbb{P}}(Y_n)$ as $n \to \infty$ if $X_n/Y_n \xrightarrow{P} 0$ as $n \to \infty$, i.e., $\forall \epsilon > 0$, $P(|X_n/Y_n| < \epsilon) \to 1$ as $n \to \infty$. Further, we write $X_n = O_{\mathbb{P}}(Y_n)$ as $n \to \infty$ if for every $\epsilon > 0$ there exist $M, N > 0$ such that $P(|X_n/Y_n| < M) > 1 - \epsilon$ for all $n > N$.

Chapter 3

Shrinkage Estimators

3.1 Mean Shrinkage Estimation

Let $\hat{\theta}$ be an estimator of a parameter $\theta \in \mathcal{H}$. The quality of an estimator $\hat{\theta}$ is measured by a real-value *loss function* $L : \mathcal{H} \times \mathcal{H} \to \mathbb{R}^+$. The *risk* of an estimator $\hat{\theta}$ is given by $R(\theta, \hat{\theta}) = E_\theta(L(\theta, \hat{\theta}))$. Now, let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of a parameter $\theta \in \mathcal{H}$. An estimator $\hat{\theta}_1$ is said to be *better than* $\hat{\theta}_2$ if $R(\theta, \hat{\theta}_1) \leq R(\theta, \hat{\theta}_2)$, $\forall\ \theta$ and $\exists\ \theta$ such that $R(\theta, \hat{\theta}_1) < R(\theta, \hat{\theta}_2)$. An estimator $\hat{\theta}_2$ is said to be *inadmissible* if there exists a better estimator $\hat{\theta}_1$.

Let $\alpha \geq 0$ and f^* be an arbitrary fixed function in $\mathcal{H}$, which is independent from the sample. We define the following mean shrinkage estimator

$$\hat{\mu}_\alpha = \alpha f^* + (1 - \alpha)\hat{\mu}, \tag{3.1}$$

When $\alpha = 0$, the shrinkage estimator becomes the empirical average estimator. We denote with $\Delta = R(\mu, \hat{\mu}) = E\|\hat{\mu} - \mu\|^2_{\mathcal{H}}$ and $\Delta_\alpha = R(\mu, \hat{\mu}_\alpha) = E\|\hat{\mu_\alpha} - \mu\|^2_{\mathcal{H}}$.

The risk of $\hat{\mu}$ can be expressed as

$$\begin{aligned}
\Delta &= R(\mu, \hat{\mu}) \\
&= E\|\hat{\mu} - \mu\|^2_{\mathcal{H}} \\
&= E[\|\hat{\mu}\|^2_{\mathcal{H}} + \|\mu\|^2_{\mathcal{H}} - 2\langle\, \hat{\mu}, \mu\rangle_{\mathcal{H}}]
\end{aligned}$$

By adding and subtracting the term $\|\mu\|^2_{\mathcal{H}}$, we obtain

$$\begin{aligned}\Delta &= E[\|\hat{\mu}\|^2_{\mathcal{H}} + \|\mu\|^2_{\mathcal{H}} - \|\mu\|^2_{\mathcal{H}} + \|\mu\|^2_{\mathcal{H}} - 2\langle\, \hat{\mu}, \mu\rangle_{\mathcal{H}}] \\ &= E[\|\hat{\mu}\|^2_{\mathcal{H}} - \|\mu\|^2_{\mathcal{H}}] + E[2\|\mu\|^2_{\mathcal{H}} - 2\langle\, \hat{\mu}, \mu\rangle_{\mathcal{H}}] \\ &= E\|\hat{\mu}\|^2_{\mathcal{H}} - \|\mu\|^2_{\mathcal{H}}\end{aligned}$$

Defining $\tilde{k}(X, \tilde{X}) = \langle\, r(X), r(\tilde{X})\rangle_{\mathcal{H}}$, we have that

$$\begin{aligned}\Delta &= E\left[\frac{1}{n^2}\sum_{i,j=1}^{n} \tilde{k}(X_i, X_j)\right] - \|\mu\|^2_{\mathcal{H}} \\ &= E\left[\frac{1}{n^2}\sum_{i=1}^{n} \tilde{k}(X_i, X_j) + \frac{1}{n^2}\sum_{i\neq j}^{n} \tilde{k}(X_i, X_j)\right] - \|\mu\|^2_{\mathcal{H}} \\ &= \frac{1}{n}\big[E_X \tilde{k}(X, X) - E_{X,\tilde{X}} \tilde{k}(X, \tilde{X})\big],\end{aligned}$$

The theorem below shows that the estimator $\hat{\mu}$ is inadmissible.

Theorem 3.1 Let $\mathcal{X}$ be a separable topological space. Suppose

$$\alpha_* = \frac{\Delta}{\Delta + \|f^* - \mu\|^2_{\mathcal{H}}}. \tag{3.2}$$

Then, for all probability distributions $\mathbb{P}$ and measurable continuous functions r satisfying the condition $\int_{\mathcal{X}} \|r(x)\|_{\mathcal{H}}\, d\mathbb{P}(x) < \infty$, we have $\Delta_\alpha < \Delta$ if and only if

$$0 < \alpha < 2\alpha_*.$$

In particular, $\arg\min_{\alpha\in\mathbb{R}}(\Delta_\alpha - \Delta)$ is unique and is given by α_*.

Proof. The risk of $\hat{\mu}_\alpha$ can be expressed as follows:

$$\begin{aligned}\Delta_\alpha = R(\mu, \hat{\mu}_\alpha) &= E\|\hat{\mu}_\alpha - \mu\|^2_{\mathcal{H}} = E\|\hat{\mu}_\alpha - \mu + E(\hat{\mu}_\alpha) - E(\hat{\mu}_\alpha)\|^2_{\mathcal{H}} \\ &= \|E(\hat{\mu}_\alpha) - \mu\|^2_{\mathcal{H}} + E\|\hat{\mu}_\alpha - E(\hat{\mu}_\alpha)\|^2_{\mathcal{H}} + 2\langle\, E(\hat{\mu}_\alpha) - \mu, \hat{\mu}_\alpha - E(\hat{\mu}_\alpha)\rangle_{\mathcal{H}} \\ &= \|Bias(\hat{\mu}_\alpha)\|^2_{\mathcal{H}} + Var(\hat{\mu}_\alpha),\end{aligned}$$

where

$$Bias(\hat{\mu}_\alpha) = E(\hat{\mu}_\alpha) - \mu = E[\alpha f^* + (1-\alpha)\hat{\mu}] - \mu = \alpha(f^* - \mu)$$

and

$$Var(\hat{\mu}_\alpha) = (1-\alpha)^2 E\|\hat{\mu} - \mu\|^2_{\mathcal{H}} = (1-\alpha)^2\Delta.$$

Therefore,

$$\Delta_\alpha = \alpha^2\|f^* - \mu\|^2_{\mathcal{H}} + (1-\alpha)^2\Delta,$$

which shows that the theorem holds. ■

Thus, there exists a large class of mean shrinkage estimators that achieve lower risk than the empirical estimator for a certain choices of α. Technically speaking, these are not estimators as they depend on the unknown mean μ. In the following, we obtain estimators of the form (3.1), where α is determined by the random sample $X_1, \ldots, X_n$.

3.2 Shrinkage Parameter Estimation

As we can see from the previous discussion, to estimate the shrinkage estimators we need to choose a value for the parameter α as well as for the fixed function f^*. In this section, we present estimators that do not depend on any prior information about the data distribution. Without loss of generality, we assume that $f^* = 0$. Therefore, our proposed estimators of μ are of the form

$$\hat{\mu}_{\tilde{\alpha}} = (1-\tilde{\alpha})\hat{\mu},$$

where $\tilde{\alpha}$ is an estimator of α_*.

Suppose $\hat{E}\tilde{k}(X,X) = \frac{1}{n}\sum_{i=1}^n \tilde{k}(X_i,X_i)$ and $\hat{E}\tilde{k}(X,\tilde{X}) = \frac{1}{n(n-1)}\sum_{i\neq j}^n \tilde{k}(X_i,X_j)$.

Define $\|\hat{\mu}\|_{\mathcal{H}}^2 = \frac{1}{n^2}\sum_{i,j=1}^{n}\tilde{k}(X_i, X_j)$ and $\hat{\Delta} = \frac{\hat{E}\tilde{k}(X,X) - \hat{E}\tilde{k}(X,\tilde{X})}{n}$.

Since the maximal differences between Δ_α and Δ occurs at α_*, we examine an estimator of α_* that is given by

$$\tilde{\alpha}_1 = \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{\mu}\|_{\mathcal{H}}^2}. \tag{3.3}$$

Furthermore, we examine the following estimator of α_*

$$\tilde{\alpha}_2 = \frac{\hat{\Delta}}{\|\hat{\mu}\|_{\mathcal{H}}^2}, \tag{3.4}$$

which can be obtained as the minimizer of an unbiased estimator of the risk Δ_α with respect to α.

Lets define an estimator of the risk Δ_α, UE, given by

$$\mathrm{UE}(\alpha) = \alpha^2\big[\|\hat{\mu}\|_{\mathcal{H}}^2 - \hat{\Delta}\big] + (1-\alpha)^2\hat{\Delta}.$$

Note that $\mathrm{UE}(\alpha)$ is an unbiased estimator of the risk Δ_α since

$$\begin{aligned} E[\mathrm{UE}(\alpha)] &= E\big[\alpha^2\big[\|\hat{\mu}\|_{\mathcal{H}}^2 - \hat{\Delta}\big] + (1-\alpha)^2\hat{\Delta}\big] \\ &= \alpha^2\big[E\|\hat{\mu}\|_{\mathcal{H}}^2 - E\hat{\Delta}\big] + (1-\alpha)^2 E\hat{\Delta}. \end{aligned}$$

The term $E\|\hat{\mu}\|_{\mathcal{H}}^2$ can be expressed as

$$\begin{aligned} E\|\hat{\mu}\|_{\mathcal{H}}^2 &= E\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\tilde{k}(X_i, X_j)\right) \\ &= E\left(\frac{1}{n^2}\sum_{i=1}^{n}\tilde{k}(X_i, X_i) + \frac{1}{n^2}\sum_{i\neq j}^{n}\tilde{k}(X_i, X_j)\right) \\ &= \frac{1}{n}E_X\tilde{k}(X,X) + \frac{n-1}{n}E_{X,\tilde{X}}\tilde{k}(X,\tilde{X}). \end{aligned}$$

Also, the term $E\hat{\Delta}$ can be expressed as

$$\begin{aligned} E\hat{\Delta} &= E\bigg(\frac{1}{n^2}\sum_{i=1}^{n}\tilde{k}(X_i,X_i) - \frac{1}{n^2(n-1)}\sum_{i\neq j}^{n}\tilde{k}(X_i,X_j)\bigg) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}E\tilde{k}(X_i,X_i) - \frac{1}{n^2(n-1)}\sum_{i\neq j}^{n}E\tilde{k}(X_i,X_j) \\ &= \frac{1}{n}E_X\tilde{k}(X,X) - \frac{1}{n}E_{X,\tilde{X}}\tilde{k}(X,\tilde{X}). \end{aligned}$$

Combining the above expressions, we obtain

$$\begin{aligned} E[\mathrm{UE}(\alpha)] &= \alpha^2\bigg[\frac{1}{n}E_X\tilde{k}(X,X) + \frac{n-1}{n}E_{X,\tilde{X}}\tilde{k}(X,\tilde{X}) - \frac{1}{n}E_X\tilde{k}(X,X) \\ &\quad + \frac{1}{n}E_{X,\tilde{X}}\tilde{k}(X,\tilde{X})\bigg] + (1-\alpha)^2\bigg[\frac{1}{n}E_X\tilde{k}(X,X) - \frac{1}{n}E_{X,\tilde{X}}\tilde{k}(X,\tilde{X})\bigg] \\ &= \alpha^2 E_{X,\tilde{X}}\tilde{k}(X,\tilde{X}) + (1-\alpha)^2\bigg[\frac{1}{n}E_X\tilde{k}(X,X) - \frac{1}{n}E_{X,\tilde{X}}\tilde{k}(X,\tilde{X})\bigg] \\ &= \alpha^2\big[\|\mu\|_{\mathcal{H}}^2 - \Delta\big] + (1-\alpha)^2\Delta = \Delta_\alpha, \end{aligned}$$

which shows that the estimator $\mathrm{UE}(\alpha)$ is an unbiased estimator of the risk Δ_α.

Taking the first derivative of the unbiased estimator $\mathrm{UE}(\alpha)$ with respect to α and setting it equals to zero, we obtain $\alpha = \tilde{\alpha_2} = \frac{\hat{\Delta}}{\|\hat{\mu}\|_{\mathcal{H}}^2}$. Since the second derivative with respect to α is positive, we conclude that $\tilde{\alpha_2}$ is a unique minimizer of the risk Δ_α.

Theorem 3.2 Suppose $n \geq 2$ and $f^* = 0$. Define $\tilde{k}(X,\tilde{X}) = \langle r(X), r(\tilde{X})\rangle$, $\hat{E}\tilde{k}(X,X) = \frac{1}{n}\sum_{i=1}^{n}\tilde{k}(X_i,X_i)$ and $\hat{E}\tilde{k}(X,\tilde{X}) = \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\tilde{k}(X_i,X_j)$. Let r be a measurable continuous function on a separable topological space $\mathcal{X}$ satisfying $\int_{\mathcal{X}}\tilde{k}(x,x)\,d\mathbb{P}(x) < \infty$. Define

$$\hat{\Delta} = \frac{\hat{E}\tilde{k}(X,X) - \hat{E}\tilde{k}(X,\tilde{X})}{n} \quad \text{and} \quad \|\hat{\mu}\|_{\mathcal{H}}^2 = \frac{1}{n^2}\sum_{i,j=1}^{n}\tilde{k}(X_i,X_j)$$

Assume there exist finite constants $k_i > 0$ and $\sigma_i > 0$ for $i = 1, 2$ such that

$$E\|r(X) - \mu\|_{\mathcal{H}}^m \leq \frac{m!}{2}\sigma_1^2 k_1^{m-2}, \forall\, m \geq 2$$

and

$$E|\tilde{k}(X, X) - E_x\tilde{k}(X, X)|^m \leq \frac{m!}{2}\sigma_2^2 k_2^{m-2}, \forall\, m \geq 2.$$

Then, as $n \to \infty$, $|\tilde{\alpha}_i - \alpha_*| = O_{\mathbb{P}}(n^{-3/2}), \quad i = 1, 2$ and

$$\left| \|\hat{\mu}_{\tilde{\alpha}_i} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}} \right| = O_{\mathbb{P}}(n^{-3/2}), \quad i = 1, 2.$$

Also, as $n \to \infty$, we obtain

$$\min_{\alpha} E\|\hat{\mu}_{\alpha} - \mu\|_{\mathcal{H}}^2 \leq E\|\hat{\mu}_{\tilde{\alpha}_i} - \mu\|_{\mathcal{H}}^2 \leq \min_{\alpha} E\|\hat{\mu}_{\alpha} - \mu\|_{\mathcal{H}}^2 + O(n^{-3/2}), \quad i = 1, 2.$$

Proof. By (3.2) and (3.3),

$$\begin{aligned}
\tilde{\alpha}_1 - \alpha_* &= \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{\mu}\|_{\mathcal{H}}^2} - \frac{\Delta}{\Delta + \|\mu\|_{\mathcal{H}}^2} \\
&= \frac{\hat{\Delta}(\Delta + \|\mu\|_{\mathcal{H}}^2) - \Delta(\hat{\Delta} + \|\hat{\mu}\|_{\mathcal{H}}^2)}{(\hat{\Delta} + \|\hat{\mu}\|_{\mathcal{H}}^2)(\Delta + \|\mu\|_{\mathcal{H}}^2)} \\
&= \frac{\hat{\Delta}(\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2)}{(\hat{\Delta} + \|\hat{\mu}\|_{\mathcal{H}}^2)(\Delta + \|\mu\|_{\mathcal{H}}^2)} + \frac{(\hat{\Delta} - \Delta)\|\hat{\mu}\|_{\mathcal{H}}^2}{(\hat{\Delta} + \|\hat{\mu}\|_{\mathcal{H}}^2)(\Delta + \|\mu\|_{\mathcal{H}}^2)} \\
&= \frac{\alpha_*(\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2) + (1 - \alpha_*)(\hat{\Delta} - \Delta)}{\hat{\Delta} + \|\hat{\mu}\|_{\mathcal{H}}^2}.
\end{aligned}$$

Therefore, we obtain

$$|\tilde{\alpha}_1 - \alpha_*| \leq \frac{\alpha_*\left| \|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2 \right| + (1 + \alpha_*)|\hat{\Delta} - \Delta|}{(\Delta + \|\mu\|_{\mathcal{H}}^2) - (\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2) + (\hat{\Delta} - \Delta)}.$$

By a result of Muandet et al. (2016), we have

$$\left|\|\mu\|^2 - \|\hat{\mu}\|_{\mathcal{H}}^2\right| = O_{\mathbb{P}}(n^{-1/2}), \quad |\hat{\Delta} - \Delta| = O_{\mathbb{P}}(n^{-3/2})$$

and

$$\Delta + \|\mu\|_{\mathcal{H}}^2 = O(1), \quad \alpha_* = O(n^{-1})$$

and thus it follows that $|\tilde{\alpha}_1 - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$, as $n \to \infty$.

As regards the term $\left|\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}\right|$, it is sufficient to notice that

$$\left|\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}\right| \leq \|\hat{\mu}_{\tilde{\alpha}_1} - \hat{\mu}_{\alpha_*}\|_{\mathcal{H}}$$

and

$$\|\hat{\mu}_{\tilde{\alpha}_1} - \hat{\mu}_{\alpha_*}\|_{\mathcal{H}} \leq |\tilde{\alpha}_1 - \alpha_*|\|\hat{\mu} - \mu\|_{\mathcal{H}} + |\tilde{\alpha}_1 - \alpha_*|\|\mu\|_{\mathcal{H}}$$

and therefore $\left|\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}\right| = O_{\mathbb{P}}(n^{-3/2})$.

Now, since

$$\begin{aligned}
\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\|_{\mathcal{H}}^2 - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}^2 &\leq (\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\| - \|\hat{\mu}_{\alpha_*} - \mu\|)\left|\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}\right| \\
&\leq 2(\|\hat{\mu}\|_{\mathcal{H}} + \|\mu\|_{\mathcal{H}})\left|\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}\right| \\
&\leq 2(\|\hat{\mu} - \mu\|_{\mathcal{H}} + 2\|\mu\|_{\mathcal{H}})\left|\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}\right|,
\end{aligned}$$

we can derive an upper bound for $P(\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\|_{\mathcal{H}}^2 - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}^2 > \epsilon)$ and using the fact that

$$E\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\|_{\mathcal{H}}^2 - E\|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}} = \int_0^\infty P(\|\hat{\mu}_{\tilde{\alpha}_1} - \mu\|_{\mathcal{H}}^2 - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}^2 > \epsilon)\, d\epsilon$$

we get the requested.

Similarly, for $\tilde{\alpha}_2$. By (3.2) and (3.4),

$$\begin{aligned}
\tilde{\alpha}_2 - \alpha_* &= \frac{\hat{\Delta}}{\|\hat{\mu}\|_{\mathcal{H}}^2} - \frac{\Delta}{\Delta + \|\mu\|_{\mathcal{H}}^2} = \frac{\hat{\Delta}(\Delta + \|\mu\|_{\mathcal{H}}^2) - \Delta\|\hat{\mu}\|_{\mathcal{H}}^2}{\|\hat{\mu}\|_{\mathcal{H}}^2(\Delta + \|\mu\|_{\mathcal{H}}^2)} \\
&= \frac{\hat{\Delta}\Delta}{\|\hat{\mu}\|_{\mathcal{H}}^2(\Delta + \|\mu\|_{\mathcal{H}}^2)} + \frac{\hat{\Delta}\|\mu\|_{\mathcal{H}}^2 - \Delta\|\hat{\mu}\|_{\mathcal{H}}^2}{\|\hat{\mu}\|_{\mathcal{H}}^2(\Delta + \|\mu\|_{\mathcal{H}}^2)} \\
&= \frac{\hat{\Delta}\Delta}{\|\hat{\mu}\|_{\mathcal{H}}^2(\Delta + \|\mu\|_{\mathcal{H}}^2)} + \frac{\hat{\Delta}(\|\mu\|^2 - \|\hat{\mu}\|_{\mathcal{H}}^2)}{\|\hat{\mu}\|_{\mathcal{H}}^2(\Delta + \|\mu\|_{\mathcal{H}}^2)} + \frac{(\hat{\Delta} - \Delta)\|\hat{\mu}\|_{\mathcal{H}}^2}{\|\hat{\mu}\|_{\mathcal{H}}^2(\Delta + \|\mu\|_{\mathcal{H}}^2)} \\
&= \tilde{\alpha}_2\alpha_* + \frac{\tilde{\alpha}_2(\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2)}{(\Delta + \|\mu\|_{\mathcal{H}}^2)} + \frac{(\hat{\Delta} - \Delta)}{\Delta + \|\mu\|_{\mathcal{H}}^2}.
\end{aligned}$$

Therefore, we find that

$$|\tilde{\alpha}_2 - \alpha_*| \leq \tilde{\alpha}_2\alpha_* + \frac{\tilde{\alpha}_2\left|\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2\right|}{(\Delta + \|\mu\|_{\mathcal{H}}^2)} + \frac{|\hat{\Delta} - \Delta|}{\Delta + \|\mu\|_{\mathcal{H}}^2}.$$

As for the bound on $\tilde{\alpha}_2$, we have

$$\begin{aligned}
\tilde{\alpha}_2 &= \frac{\hat{\Delta}}{\|\hat{\mu}\|_{\mathcal{H}}^2} \\
&= \frac{\hat{\Delta}}{(\|\hat{\mu}\|_{\mathcal{H}}^2 - \|\mu\|_{\mathcal{H}}^2) + (\Delta + \|\hat{\mu}\|_{\mathcal{H}}^2) - \Delta} \\
&= \frac{\hat{\Delta} - \Delta + \Delta}{(\|\hat{\mu}\|_{\mathcal{H}}^2 - \|\mu\|_{\mathcal{H}}^2) + (\Delta + \|\hat{\mu}\|_{\mathcal{H}}^2) - \Delta} \\
&= \frac{\hat{\Delta} - \Delta}{(\|\hat{\mu}\|_{\mathcal{H}}^2 - \|\mu\|_{\mathcal{H}}^2) + (\Delta + \|\hat{\mu}\|_{\mathcal{H}}^2) - \Delta} + \frac{\Delta}{(\|\hat{\mu}\|_{\mathcal{H}}^2 - \|\mu\|_{\mathcal{H}}^2) + (\Delta + \|\hat{\mu}\|_{\mathcal{H}}^2) - \Delta} \\
&= O_{\mathbb{P}}(n^{-3/2}) + O_{\mathbb{P}}(n^{-1}) = O(n^{-1}).
\end{aligned}$$

Again, from Muandet et al. (2016), we have

$$\left|\|\mu\|^2 - \|\hat{\mu}\|_{\mathcal{H}}^2\right| = O_{\mathbb{P}}(n^{-1/2}), \quad |\hat{\Delta} - \Delta| = O_{\mathbb{P}}(n^{-3/2})$$

and

$$\Delta + \|\mu\|_{\mathcal{H}}^2 = O(1), \quad \alpha_* = O(n^{-1}).$$

and thus it follows that $|\tilde{\alpha}_2 - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$

Alternatively,

$$\tilde{\alpha}_2 - \alpha_* = \tilde{\alpha}_2\alpha_* + \frac{\tilde{\alpha}_2(\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2)}{(\Delta + \|\mu\|_{\mathcal{H}}^2)} + \frac{(\hat{\Delta} - \Delta)}{\Delta + \|\mu\|_{\mathcal{H}}^2}.$$

Rearranging $\tilde{\alpha}_2$,

$$\tilde{\alpha}_2 - \tilde{\alpha}_2\alpha_* - \frac{\tilde{\alpha}_2(\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2)}{(\Delta + \|\mu\|_{\mathcal{H}}^2)} = \alpha_* + \frac{(\hat{\Delta} - \Delta)}{\Delta + \|\mu\|_{\mathcal{H}}^2}.$$

We get,

$$\tilde{\alpha}_2 = \frac{\alpha_* + \dfrac{(\hat{\Delta} - \Delta)}{\Delta + \|\mu\|_{\mathcal{H}}^2}}{1 - \alpha_* - \dfrac{\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2}{\Delta + \|\mu\|_{\mathcal{H}}^2}}.$$

Thus,

$$\begin{aligned}\tilde{\alpha}_2 - \alpha_* &= \frac{\alpha_* + \dfrac{(\hat{\Delta} - \Delta)}{\Delta + \|\mu\|_{\mathcal{H}}^2}}{1 - \alpha_* - \dfrac{\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2}{\Delta + \|\mu\|_{\mathcal{H}}^2}} - \alpha_* \\ &= \frac{\alpha_*\Delta + (\hat{\Delta} - \Delta) + \alpha_*(\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2)}{(\|\hat{\mu}\|_{\mathcal{H}}^2 - \|\mu\|_{\mathcal{H}}^2) + \|\mu\|_{\mathcal{H}}^2}\end{aligned}$$

and therefore

$$\begin{aligned}|\tilde{\alpha}_2 - \alpha_*| &= \left|\frac{\alpha_*\Delta + (\hat{\Delta} - \Delta) + \alpha_*(\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2)}{\|\mu\|_{\mathcal{H}}^2 - (\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2)}\right| \\ &\leq \frac{\alpha_*\Delta + |\hat{\Delta} - \Delta| + \alpha_*\left|\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2\right|}{\left|\|\mu\|_{\mathcal{H}}^2 - \left|\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2\right|\right|}.\end{aligned}$$

Using the inequalities $|x + y| \leq |x| + |y|$ and $|x - y| \geq \big||x| - |y|\big|$, as well as the asymptotic results $\Delta = O(n^{-1})$, $\left|\|\mu\|_{\mathcal{H}}^2 - \|\hat{\mu}\|_{\mathcal{H}}^2\right| = O_{\mathbb{P}}(n^{-1/2})$, $\|\mu\|_{\mathcal{H}}^2 = O(1)$,

$\alpha_* = O(n^{-1})$ and $|\hat{\Delta} - \Delta| = O_{\mathbb{P}}(n^{-3/2})$, we obtain

$$|\tilde{\alpha}_2 - \alpha_*| = O_{\mathbb{P}}(n^{-3/2}). \quad \blacksquare$$

It is clear from the above result that $\hat{\mu}_{\tilde{\alpha}_i}$ is a $\sqrt{n}$-consistent estimator of μ for $i = 1, 2$ since

$$\begin{aligned}\|\hat{\mu}_{\tilde{\alpha}_i} - \mu\|_{\mathcal{H}} &\leq \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}} + O_{\mathbb{P}}(n^{-3/2}) \\ &\leq (1 - \alpha_*)\|\hat{\mu} - \mu\|_{\mathcal{H}} + \alpha_*\|\mu\|_{\mathcal{H}} + O_{\mathbb{P}}(n^{-3/2}) \\ &= O(1)O_{\mathbb{P}}(1/\sqrt{n}) + O(1/n) + O_{\mathbb{P}}(n^{-3/2}) = O_{\mathbb{P}}(1/\sqrt{n}), \quad i = 1, 2.\end{aligned}$$

Also, notice that the conditions of the Theorem 3.2 about the existence of constants $k_i > 0$ and $\sigma_i > 0$ for $i = 1, 2$ are satisfied for bounded $\tilde{k}$.

3.3 Shrinkage Estimation as Regression Problem

In this section, we provide an alternative approach to getting the shrinkage estimator based on a regression point of view. It can be shown that the empirical average $\hat{\mu}$ is the minimizer of the empirical risk

$$\hat{\mathcal{R}}(g) = \frac{1}{n}\sum_{i=1}^{n}\|r(X_i) - g\|_{\mathcal{H}}^2, \quad g \in \mathcal{H}.$$

Now, define $\Omega : \mathbb{R}^+ \to \mathbb{R}^+$ a monotonically increasing function given by $\Omega(t) = \lambda t^2$ and $\lambda \geq 0$. Instead of minimizing the empirical risk, one can consider minimization of the regularized empirical risk,

$$\hat{\mathcal{R}}_\lambda(g) = \hat{\mathcal{R}}(g) + \lambda\Omega(\|g\|_{\mathcal{H}}) = \frac{1}{n}\sum_{i=1}^{n}\|r(X_i) - g\|_{\mathcal{H}}^2 + \lambda\Omega(\|g\|_{\mathcal{H}}), \quad g \in \mathcal{H}.$$

The regularized empirical risk $\hat{\mathcal{R}}_\lambda(g)$ can be expressed as

$$\begin{aligned}\hat{\mathcal{R}}_\lambda(g) &= \|g\|_{\mathcal{H}}^2(1+\lambda) - 2\langle g, \frac{1}{n}\sum_{i=1}^n r(X_i)\rangle + c \\ &= \|g\|_{\mathcal{H}}^2(1+\lambda) - 2\langle g, \frac{\hat{\mu}}{1+\lambda}(1+\lambda)\rangle + c \\ &= \|g\|_{\mathcal{H}}^2 - 2\langle g, \frac{\hat{\mu}}{1+\lambda}\rangle + \left\|\frac{\hat{\mu}}{1+\lambda}\right\|_{\mathcal{H}}^2 - \left\|\frac{\hat{\mu}}{1+\lambda}\right\|_{\mathcal{H}}^2 \\ &= \left\|g - \frac{\hat{\mu}}{1+\lambda}\right\|_{\mathcal{H}}^2 - \left\|\frac{\hat{\mu}}{1+\lambda}\right\|_{\mathcal{H}}^2,\end{aligned}$$

where c is a constant that does not depend on λ and g. Clearly, the minimum is given by

$$g = \hat{\mu}_\lambda = \frac{\hat{\mu}}{1+\lambda},$$

which is a shrinkage estimator with $\alpha = \frac{\lambda}{1+\lambda}$ and $f^* = 0$. Note that $\hat{\mu}_\lambda$ is a consistent estimator of μ since

$$\|\hat{\mu}_\lambda - \mu\|_{\mathcal{H}} \leq \frac{1}{1+\lambda}\|\hat{\mu} - \mu\|_{\mathcal{H}} + \frac{\lambda}{1+\lambda}\|\mu\|_{\mathcal{H}}.$$

By choosing $\lambda = \frac{\hat{\Delta}}{\|\hat{\mu}\|_{\mathcal{H}}^2}$, we get similar results as before. Alternatively, we can have a choice of λ from leave-one-out cross validation.

Proposition 3.3 Suppose $n \geq 2$, $\rho = \frac{1}{n^2}\sum_{i,j=1}^n \tilde{k}(X_i, X_j)$, $\tau = \frac{1}{n}\sum_i^n \tilde{k}(X_i, X_i)$ and $n\rho > \tau$. Denote by $\hat{\mu}_\lambda^{(-i)}$ the mean estimator from $X_1, X_2, \ldots X_{i-1}, X_{i+1}, \ldots X_n$. Define

$$\mathcal{R}_{cv}(\lambda) = \frac{1}{n}\sum_{i=1}^n \left\|r(X_i) - \hat{\mu}_\lambda^{(-i)}\right\|_{\mathcal{H}}^2.$$

Then, the unique minimizer of $\mathcal{R}_{cv}(\lambda)$ is given by

$$\lambda_r = \frac{n(\tau - \rho)}{(n-1)(n\rho - \tau)}.$$

Proof. Define $\alpha = \frac{\lambda}{\lambda+1}$. We have

$$\begin{aligned}
\mathcal{R}_{cv}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\left\|r(X_i) - \hat{\mu}_\lambda^{(-i)}\right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{1-\alpha}{n-1}\sum_{j\neq i} r(X_j) - r(X_i)\right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{n(1-\alpha)}{n-1}\hat{\mu} - \frac{1-\alpha}{n-1}r(X_i) - r(X_i)\right\|_{\mathcal{H}}^2 \\
&= \left\|\frac{n(1-\alpha)}{n-1}\hat{\mu}\right\|_{\mathcal{H}}^2 - \frac{2}{n}\Big\langle\sum_{i=1}^{n}\frac{n-\alpha}{n-1}r(X_i), \frac{n(1-\alpha)}{n-1}\hat{\mu}\Big\rangle_{\mathcal{H}} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{n-\alpha}{n-1}r(X_i)\right\|_{\mathcal{H}}^2 \\
&= \frac{\alpha^2(n^2\rho - 2n\rho + \tau) + 2n\alpha(\rho-\tau) + n^2(\tau-\rho)}{(n-1)^2}.
\end{aligned}$$

Taking the first derivative with respect to λ and setting it equal to zero, we obtain $\lambda = \lambda_r = \frac{n(\tau-\rho)}{(n-1)(n\rho-\tau)}$. Since the second derivative with respect to λ is positive, we conclude that λ_r is a unique minimizer of the $\mathcal{R}_{cv}(\lambda)$. ■

The following result shows that $\hat{\mu}_\lambda$ is a $\sqrt{n}$-consistent estimator of μ and that $\Delta_{\lambda_r} \leq \Delta + O(n^{-3/2})$.

Theorem 3.4 Under the assumptions of Theorem 3.2 and Proposition 3.3, we have that as $n \to \infty$,

$$\left|\|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}\right| = O_{\mathbb{P}}(n^{-3/2}).$$

Also, as $n \to \infty$,

$$\min_\alpha E\|\hat{\mu}_\alpha - \mu\|_{\mathcal{H}}^2 \leq E\|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}}^2 \leq \min_\alpha E\|\hat{\mu}_\alpha - \mu\|_{\mathcal{H}}^2 + O(n^{-3/2}).$$

Proof. Define

$$\alpha_r = \frac{\lambda_r}{\lambda_r + 1} = \frac{n(\rho-\tau)}{n(n-2)\rho + \tau} \tag{3.5}$$

and

$$\gamma = \frac{\hat{E}\tilde{k}(X, \tilde{X})}{\hat{E}\tilde{k}(X, \tilde{X}) - (n-2)\hat{E}\tilde{k}(X, \tilde{X})}.$$

By (3.3) and (3.5),

$$\begin{aligned}
\alpha_r - \tilde{\alpha}_1 &= \frac{n\hat{\Delta}}{\hat{\Delta} + (n-1)\|\hat{\mu}\|_{\mathcal{H}}^2} - \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{\mu}\|_{\mathcal{H}}^2} \\
&= \frac{\hat{E}\tilde{k}(X, X) - \hat{E}\tilde{k}(X, \tilde{X})}{\hat{E}\tilde{k}(X, X) + (n-2)\hat{E}\tilde{k}(X, \tilde{X})} - \frac{\hat{E}\tilde{k}(X, X) - \hat{E}\tilde{k}(X, \tilde{X})}{2\hat{E}\tilde{k}(X, X) + (n-2)\hat{E}\tilde{k}(X, \tilde{X})} \\
&= (\tilde{\alpha}_1 - \alpha_*)\gamma + \alpha_*\gamma,
\end{aligned}$$

Thus, $|\alpha_r - \tilde{\alpha}_1| \leq |\tilde{\alpha}_1 - \alpha_*||\gamma| + \alpha_*|\gamma|$. We know from Theorem 3.2 that as $n \to \infty$,

$$|\tilde{\alpha}_1 - \alpha_*| = O_{\mathbb{P}}(n^{-3/2}) \text{ and } \alpha_* = O(n^{-1}),$$

From Muandet et al. (2016), we have

$$|\hat{E}\tilde{k}(X, X) - E_X\tilde{k}(X, X)| = O_{\mathbb{P}}(n^{-1/2})$$

and

$$|\hat{E}\tilde{k}(X, \tilde{X}) - E_{X,\tilde{X}}\tilde{k}(X, \tilde{X})| = O_{\mathbb{P}}(n^{-1/2}).$$

Therefore, $\gamma = O_{\mathbb{P}}(n^{-1})$ and $|\alpha_r - \alpha_*| \leq |\alpha_r - \tilde{\alpha}_1| + |\tilde{\alpha}_1 - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$ as $n \to \infty$.

Moreover, notice that

$$\begin{aligned}
\left| \|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}} \right| &\leq \|\hat{\mu}_{\lambda_r} - \mu_{\alpha_*}\|_{\mathcal{H}} \\
&\leq |\alpha_r - \alpha_*| \|\hat{\mu} - \mu\|_{\mathcal{H}} + |\alpha_r - \alpha_*| \|\mu\|_{\mathcal{H}}
\end{aligned}$$

and therefore $\left| \|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}} \right| = O_{\mathbb{P}}(n^{-3/2})$ as $n \to \infty$.

By replacing $\tilde{\alpha}_1$ with α_r and following the steps of the proof of Theorem 3.2, we

can show that, as $n \to \infty$,

$$\min_{\alpha} E\|\hat{\mu}_{\alpha} - \mu\|_{\mathcal{H}}^2 \leq E\|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}}^2 \leq \min_{\alpha} E\|\hat{\mu}_{\alpha} - \mu\|_{\mathcal{H}}^2 + O(n^{-3/2}). \quad \blacksquare$$

Chapter 4

Minimax Rates

As we have seen in the Chapter 3, the estimators $\hat{\mu}$, $\hat{\mu}_{\alpha_i}$ and $\hat{\mu}_\lambda$ are $\sqrt{n}$-consistent estimators of the mean in the $\mathcal{H}$-norm. The issue that arises now is whether the rate of $n^{-1/2}$ is optimal in the minimax sense. Specifically, we show that this rate is minimax in the $\mathcal{H}$-norm over the class of discrete measures or over the class of measures that have an infinitely differentiable density. Suppose $\mathcal{P}$ is a subset of Borel probability measures on $\mathcal{X}$ and $\hat{\theta}_n$ is any estimator of mean μ. Formally, we are interested in establishing

$$\limsup_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{P\in\mathcal{P}} \sqrt{n} E \, \|\hat{\theta}_n - \mu\|_{\mathcal{H}} \leq c_r,$$

and

$$\liminf_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{P\in\mathcal{P}} \sqrt{n} E \, \|\hat{\theta}_n - \mu\|_{\mathcal{H}} \geq c_r',$$

for some constants c_r and c_r'.

4.1 A Minimax Upper Bound

We show that $\Delta_{\tilde{\alpha}_i} \leq \Delta_{\alpha_*} + O(n^{-3/2}),\ i = 1, 2$. By Theorem 3.1 we have $\Delta_{\alpha_*} \leq \Delta$ and thus $\Delta_{\tilde{\alpha}_i} \leq \Delta + O(n^{-3/2}),\ i = 1, 2$. Since $\Delta = O(n^{-1})$, we have that $\Delta_{\tilde{\alpha}_i} = E\|\hat{\mu}_{\alpha_i} - \mu\|_{\mathcal{H}}^2 = O(n^{-1}),\ i = 1, 2$, i.e., $\hat{\mu}_{\alpha_i}$ is a $\sqrt{n}$-consistent estimator of μ

$i = 1, 2$. Therefore, there exists a positive constant c_r, such that

$$\limsup_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{P\in\mathcal{P}} \sqrt{n} E \, \|\hat{\theta}_n - \mu\|_{\mathcal{H}} \leq c_r.$$

4.2 A Minimax Lower Bound

Now, we consider a minimax lower bound for the estimation of μ. In order to establish that the rate of $n^{-1/2}$ is optimal in the minimax sense, we want to show that there exists a positive constant c_r', such that

$$\liminf_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{P\in\mathcal{P}} \sqrt{n} E \, \|\hat{\theta}_n - \mu\|_{\mathcal{H}} \geq c_r'.$$

By applying Le Cam's method, it is enough to show that $r_{n,r}(\mathcal{H}, \mathcal{P}) = 1/n$ and

$$\inf_{\hat{\theta}_n} \sup_{P\in\mathcal{P}} P^n\{r_{n,r}^{-1}(\mathcal{H}, \mathcal{P})\|\hat{\theta}_n - \mu\|_{\mathcal{H}} \geq c_r(\mathcal{H}, \mathcal{P})\} > 0,$$

for some positive constant $c_r'(\mathcal{H}, \mathcal{P})$.

Now, let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let P, Q be two probability measures on $(\mathcal{X}, \mathcal{A})$. Suppose that ν is a σ-finite measure on $(\mathcal{X}, \mathcal{A})$ satisfying $P << \nu$ and $Q << \nu$. The *Kullback-Leibler divergence* (Kullback, 1959) between P and Q is defined by

$$KL(P||Q) = \int \log \frac{dP}{dQ} \, dP.$$

Also, using the chain rule of KL-divergence (Kullback, 1959), we can show that the distance between the n-fold product distribution of P, P^n, and n-fold product distribution of Q, Q^n, is given by $KL(P^n||Q^n) = nKL(P||Q)$. The following result, quoted from Tsybakov 2008 (Theorem 2.5), will be useful to obtain minimax

lower bounds for the problem at hand.

Theorem 4.1 (Tsybakov, 2008, Theorem 2.5) Let Θ be a parameter space containing the elements θ_0 and θ_1. Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a class of probabilities measures on $\mathbb{R}^d$ and $d : \Theta \times \Theta \to [0, \infty]$ is a metric on Θ. Assume that $d(\theta_0, \theta_1) \geq 2s$ and $KL(P^n_{\theta_0}||P^n_{\theta_1}) \leq \alpha$ for some $s > 0$ and $\alpha > 0$. Then,

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P^n_\theta \{d(\hat{\theta}_n, \theta) \geq s\} > \frac{1}{4} \max\left(e^{-\alpha}, 2(1 - \sqrt{\alpha/2})\right).$$

In this thesis, we examine two choices for the class $\mathcal{P}$. The first choice for $\mathcal{P}$ is the set of all Borel discrete probability measures on $\mathbb{R}^d$, and second choice is the set of all Borel absolutely continuous probability measures on $\mathbb{R}^d$ with densities that are continuously infinitely differentiable.

4.2.1 Discrete Probability Measures

First, we examine the case in which $\mathcal{P}$ is the set of all Borel discrete probability measures on $\mathbb{R}^d$. A minimax rate of $n^{-1/2}$ is obtained under specific assumptions on r.

Theorem 4.2 Let $\mathcal{P}$ be the set of all Borel discrete probability measures on $\mathcal{X} = \mathbb{R}^d$. Suppose that there exist $X, Y \in \mathbb{R}^d$ and $\beta > 0$ such that $\|r(X) - r(Y)\|^2_{\mathcal{H}} \geq \beta$. Then,

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu\|_{\mathcal{H}} \geq \frac{1}{6}\sqrt{\frac{\beta}{n}} \right\} \geq \frac{1}{4}.$$

Proof. Suppose $X, Y \in \mathbb{R}^d$, $0 < p_0, p_1 < 1$, and δ_X, δ_Y are the Dirac measures at X and Y respectively. We choose two discrete distributions, $P_0 = p_0\delta_X + (1-p_0)\delta_Y$ and $P_1 = p_1\delta_X + (1-p_1)\delta_Y$.
Define $\theta_0 = \mu(P_0) = \int_{\mathcal{X}} r(x)\, dP_0(x)$ and $\theta_1 = \mu(P_1) = \int_{\mathcal{X}} r(x)\, dP_1(x)$. Then,

$$\begin{aligned}\|\theta_0 - \theta_1\|_{\mathcal{H}}^2 &= \Big\| \int_{\mathbb{R}^d} r(x)\, dP_0(x) - \int_{\mathbb{R}^d} r(x)\, dP_1(x) \Big\|_{\mathcal{H}}^2 \\ &= \|p_0 r(X) + (1-p_0) r(Y) - p_1 r(X) - (1-p_1) r(Y)\|_{\mathcal{H}}^2 \\ &= \|(p_0 - p_1)(r(X) - r(Y))\|_{\mathcal{H}}^2 = (p_0 - p_1)^2 \|r(X) - r(Y)\|_{\mathcal{H}}^2.\end{aligned}$$

Choosing X and Y in such a way that $\|r(X) - r(Y)\|_{\mathcal{H}}^2 \geq \beta$ where $\beta > 0$, we find that $\|\theta_0 - \theta_1\|_{\mathcal{H}}^2 \geq (p_0 - p_1)^2 \beta$.

By Tolstikhin (2017),

$$KL(P_0^n || P_1^n) \leq 4n\Big(p_0 - \frac{1}{2}\Big)^2.$$

Choosing p_0 such that $\Big(p_0 - \frac{1}{2}\Big)^2 = \frac{1}{9n}$, we obtain $KL(P_0^n || P_1^n) \leq \frac{4}{9}$ and $\|r(X) - r(Y)\|_{\mathcal{H}}^2 \geq \frac{\beta}{9n}$. By applying Theorem 4.1 with $s = \frac{1}{6}\sqrt{\frac{\beta}{n}}$ and $\alpha = \frac{4}{9}$, we get the requested. ■

4.2.2 Probability Measures with Smooth Densities

In this section, we examine the minimax rate for the case in which $\mathcal{P}$ is the set of all Borel absolutely continuous probability measures on $\mathbb{R}^d$ with continuously infinitely differentiable densities. We show again that the minimax rate of $n^{-1/2}$ is obtained under certain assumptions on r.

Assume $\Lambda \in M_+^b(\mathbb{R}^d \times \mathbb{R}^d)$ and $0 < c < \infty$. First, we derive the minimax rate when $\tilde{k}(X, Y) = \langle r(X), r(Y) \rangle_{\mathcal{H}}$ has the following form,

$$\tilde{k}(X, Y) = c \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-i(\langle x, \omega \rangle) + \langle y, \eta \rangle)}\, d\Lambda(\omega, \eta),$$

This will be a stepping stone toward proving the more general case. In the following lemma, we derive a closed form expression for the distance $\|\theta_0 - \theta_1\|_{\mathcal{H}}$.

Lemma 4.3 Let θ_0 and θ_1 be the mean estimators of the Gaussian measures $N_d(\mu, \sigma_0^2 I)$ and $N_d(\mu, \sigma_1^2 I)$ for $\mu \in \mathbb{R}^d$ and $\sigma_0, \sigma_1 > 0$ respectively. Assume $\Lambda \in M_+^b(\mathbb{R}^d \times \mathbb{R}^d)$ and $0 < c < \infty$ is a constant. Suppose $\tilde{k}$ is of the form

$$\tilde{k}(X, Y) = c \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-i(\langle x, \omega \rangle) + \langle y, \eta \rangle)} \, d\Lambda(\omega, \eta).$$

Define $f_{\sigma_0, \sigma_1}(\eta) = e^{-\frac{\sigma_1^2}{2}\|\eta\|_2^2} - e^{-\frac{\sigma_0^2}{2}\|\eta\|_2^2}$ and $f_{\sigma_0, \sigma_1}(\omega) = e^{-\frac{\sigma_1^2}{2}\|\omega\|_2^2} - e^{-\frac{\sigma_0^2}{2}\|\omega\|_2^2}$. Then,

$$\|\theta_0 - \theta_1\|_{\mathcal{H}}^2 = c \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \cos\langle \mu, \omega + \eta \rangle f_{\sigma_0, \sigma_1}(\eta) f_{\sigma_0, \sigma_1}(\omega) \, d\Lambda(\omega, \eta).$$

Proof. Define $Q = P_0 - P_1$, $\underline{X} = (X, Y)$ and $\underline{\omega} = (\omega, \eta) \in \mathbb{R}^n \times \mathbb{R}^n$. We have

$$\begin{aligned} \|\theta_0 - \theta_1\|_{\mathcal{H}}^2 &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{k}(x, y) \, dQ(x) \, dQ(y) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{k}(x, y) \, dQ \otimes Q(x, y) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \tilde{k}(\underline{x}) \, d(Q \otimes Q)(\underline{x}) \end{aligned}$$

Using the fact that $\tilde{k}$ has a specific form, we obtain

$$\|\theta_0 - \theta_1\|_{\mathcal{H}}^2 = \int_{\mathbb{R}^d \times \mathbb{R}^d} c \int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-i\langle \underline{x}, \underline{w} \rangle} \, d\Lambda(\underline{\omega}) \, d(Q \otimes Q)(\underline{x})$$

Since the integral is absolutely integrable, we use Tonelli-Fubini theorem (Dudley, 2002, Theorem 4.4.5) to obtain

$$\|\theta_0 - \theta_1\|_{\mathcal{H}}^2 = \int_{\mathbb{R}^d \times \mathbb{R}^d} c \int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-i\langle \underline{x}, \underline{w} \rangle} \, d(Q \otimes Q)(\underline{x}) \, d\Lambda(\underline{\omega}),$$

Therefore, using Definition 2.7, we have

$$\|\theta_0 - \theta_1\|_{\mathcal{H}}^2 = \int_{\mathbb{R}^d \times \mathbb{R}^d} \widehat{Q \otimes Q}(\underline{\omega}) \, d\Lambda(\underline{\omega}),$$

where $\widehat{\ }$ denotes the Fourier transform. Note that

$$\widehat{Q \otimes Q}(\underline{\omega}) = \widehat{P_0 \otimes P_0}(\underline{\omega}) - \widehat{P_0 \otimes P_1}(\underline{\omega}) - \widehat{P_1 \otimes P_0}(\underline{\omega}) + \widehat{P_1 \otimes P_1}(\underline{\omega}). \tag{4.1}$$

Recall that from Wendland (2005, Theorem 5.18), we have that for any $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$ the following holds

$$\left[\frac{1}{(2\pi\sigma^2)^{d/2}} e^{\frac{-\|X-\mu\|_2^2}{2\sigma^2}}\right]^{\wedge}(\omega) = \frac{1}{(2\pi)^{d/2}} \exp\left(-i\langle\mu,\omega\rangle - \frac{\sigma^2\|\omega\|_2^2}{2}\right),\ \omega \in \mathbb{R}^d.$$

For the second term of (4.1), it follows from Wendland (2005, Theorem 5.18) that

$$\begin{aligned}
\widehat{P_0 \otimes P_1}(\underline{\omega}) &= c\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-i\langle \underline{x},\underline{\omega}\rangle}\, d(P_0\otimes P_1)(\underline{x}) \\
&= c\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i(\langle x,\omega\rangle+\langle y,\eta\rangle)}\, dP_0(x)\, dP_1(y) \\
&= c\int \frac{e^{-i\langle x,\omega\rangle}}{(2\pi\sigma_0^2)^{d/2}} e^{-\frac{1}{2\sigma_0^2}\|x-\mu\|_2^2}\, dx \int \frac{e^{-i\langle y,\eta\rangle}}{(2\pi\sigma_1^2)^{d/2}} e^{-\frac{1}{2\sigma_1^2}\|y-\mu\|_2^2}\, dy \\
&= ce^{-i\langle\mu,\omega\rangle} e^{-\frac{\sigma_0^2}{2}\|\omega\|_2^2} e^{-i\langle\mu,\eta\rangle} e^{-\frac{\sigma_1^2}{2}\|\eta\|_2^2}.
\end{aligned}$$

Similarly, we calculate the other terms of (4.1). Therefore, using the Euler formula, $e^{ix} = \cos(x) + i\sin(x)$ for all $x \in \mathbb{R}$, and the fact that Λ is symmetric while sin is an odd function, we get

$$\|\theta_0 - \theta_1\|_{\mathcal{H}}^2 = c\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \cos\langle\mu,\omega+\eta\rangle f_{\sigma_0,\sigma_1}(\eta) f_{\sigma_0,\sigma_1}(\omega)\, d\Lambda(\omega,\eta). \quad \blacksquare$$

Theorem 4.4 Let $\mathcal{P}$ be the set of all probability distributions on $\mathcal{X} = \mathbb{R}^d$ whose densities are continuously infinitely differentiable. Assume $\Lambda \in M_+^b(\mathbb{R}^d \times \mathbb{R}^d)$ and $0 < c < \infty$ is a constant. Suppose $\tilde{k}$ is of the form

$$\tilde{k}(X,Y) = c\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i(\langle x,\omega\rangle)+\langle y,\eta\rangle)}\, d\Lambda(\omega,\eta).$$

Then, there exists $c_r > 0$ such that for any n

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu\|_{\mathcal{H}} \geq \sqrt{\frac{c_r}{n}} \right\} \geq \frac{1}{8}.$$

Proof. We need to find an upper bound for the Kullback-Leibler divergence, $KL(G_0^n||G_1^n)$, where $G_0 = N_d(0, \sigma_0^2 I)$ and $G_1 = N_d(0, \sigma_1^2 I)$. By Zhou et al. (2019), we have

$$KL(G_0^n||G_1^n) = \frac{nd}{2}\Big(\frac{\sigma_0^2 - \sigma_1^2}{2\sigma_1^2} - \ln\big(\frac{\sigma_0 - \sigma_1}{\sigma_1} + 1\big)\Big).$$

Since $\ln(1+x) \geq x - \frac{x^2}{2} > 0$, we obtain

$$KL(G_0^n||G_1^n) \leq nd\frac{(\sigma_0^2 - \sigma_1^2)^2}{\sigma_1^2}.$$

By setting $\sigma_1 = 1$ and $\sigma_0 = 1 + \sqrt{\frac{1}{dn}}$, we have

$$KL(G_0^n||G_1^n) \leq 1.$$

Next, choose $A = \{\eta \in \mathbb{R}^d : 0 < \|\eta\|_2^2 < \epsilon\}$ and $B = \{\omega \in \mathbb{R}^d : 0 < \|\omega\|_2^2 < \epsilon\}$, where $\epsilon > 0$. Define $h_\omega(x) = e^{-x\|\omega\|_2^2}$ and $h_\eta(x) = e^{-x\|\eta\|_2^2}$. Since $h''_\omega(x), h''_\eta(x) > 0$, they are convex functions. Then for $x, y \in [1/2, 2]$ and $y < x$, we obtain $h_\omega(y) - h_\omega(x) \geq h'_\omega(2)(y - x)$ and $h_\eta(y) - h_\eta(x) \geq h'_\eta(2)(y - x)$. Therefore, by Lemma 4.3,

$$\begin{aligned}
\|\theta_0 - \theta_1\|_{\mathcal{H}}^2 &= c\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}(e^{-\frac{\sigma_1^2}{2}\|\eta\|_2^2} - e^{-\frac{\sigma_0^2}{2}\|\eta\|_2^2})(e^{-\frac{\sigma_1^2}{2}\|\omega\|_2^2} - e^{-\frac{\sigma_0^2}{2}\|\omega\|_2^2})\, d\Lambda(\omega, \eta) \\
&\geq c\int_A\int_B(e^{-\frac{\sigma_1^2}{2}\|\eta\|_2^2} - e^{-\frac{\sigma_0^2}{2}\|\eta\|_2^2})(e^{-\frac{\sigma_1^2}{2}\|\omega\|_2^2} - e^{-\frac{\sigma_0^2}{2}\|\omega\|_2^2})\, d\Lambda(\omega, \eta) \\
&= c\int_A\int_B\Big(h_\eta\Big(\frac{\sigma_1^2}{2}\Big) - h_\eta\Big(\frac{\sigma_0^2}{2}\Big)\Big(h_\omega\Big(\frac{\sigma_1^2}{2}\Big) - h_\omega\Big(\frac{\sigma_0^2}{2}\Big)\, d\Lambda(\omega, \eta) \\
&\geq \frac{c}{4}\int_A\int_B(\sigma_1^2 - \sigma_0^2)^2 h'_\eta(2)h'_\omega(2)\, d\Lambda(\omega, \eta) \\
&\geq \frac{c}{nd}\int_A\int_B h'_\eta(2)h'_\omega(2)\, d\Lambda(\omega, \eta).
\end{aligned}$$

Clearly, $\int_A \int_B h'_\eta(2) h'_\omega(2)\, d\Lambda(\omega, \eta) > 0$, since no open set of $\mathbb{R}^d$ is contained in the regions where $h_\eta(2) = 0$ and $h_\omega(2) = 0$. Thus, there exists $\alpha > 0$ such that

$$\|\theta_0 - \theta_1\|_{\mathcal{H}} \geq \sqrt{\frac{c\alpha}{nd}}. \qquad \blacksquare$$

Next, we present two cases in which the above condition on $\tilde{k}$ holds. First, we consider the case in which $\tilde{k}$ can be decomposed into a product of two positive definite function ψ_1 and ψ_2. Define $d\Lambda(\omega, \eta) = d\Lambda_1(\omega) d\Lambda_2(\eta)$ as a product measure and $c = \frac{1}{(2\pi)^d}$. In this case, applying Bochner's Theorem (Wendland, 2005, Theorem 6.6), we get

$$\begin{aligned}
\tilde{k}(X, Y) &= \psi_1(X)\psi_2(Y) \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle \omega, x \rangle}\, d\Lambda_1(\omega) \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle \eta, y \rangle}\, d\Lambda_2(\eta) \\
&= c \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-i(\langle x, \omega \rangle) + \langle y, \eta \rangle)}\, d\Lambda(\omega, \eta).
\end{aligned}$$

Second, when $\tilde{k}$ is translation invariant and positive definite. Define $d\Lambda(\omega, \eta) = d\delta_\eta(-\omega) d\Lambda_3(\omega)$, δ is the Dirac measure and $c = \frac{1}{(2\pi)^{d/2}}$. Applying Bochner's theorem again, we obtain

$$\begin{aligned}
\tilde{k}(X, Y) &= \psi_3(X - Y) \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle \omega, x - y \rangle}\, d\Lambda_3(\omega) \\
&= c \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-i(\langle x, \omega \rangle) + \langle y, \eta \rangle)}\, d\Lambda(\omega, \eta),
\end{aligned}$$

where .

Next, we derive the minimax rate in a more general setting.

Theorem 4.5 Let $\mathcal{P}$ be the set of all probability distributions on $\mathcal{X} = \mathbb{R}^d$ whose densities are continuously infinitely differentiable. Suppose $\tilde{k}(X, Y) = \langle r(X), r(Y) \rangle_{\mathcal{H}}$ is positive definite , $\tilde{k}(X, Y) = \tilde{k}(-X, -Y)$ for $X, Y \in \mathbb{R}^d$ and

$\tilde{k} \in L^1(\mathbb{R}^d \times \mathbb{R}^d)$. Also, assume that

$$\Big\| \int_{\mathbb{R}} r(x)\big(d - \frac{\|x\|_{\mathcal{H}}^2}{\sigma_1^2}\big) e^{-\frac{\|x\|_{\mathcal{H}}^2}{\sigma_1^2}}\, dx \Big\|_{\mathcal{H}}^2 > 0.$$

Then, there exists $c_r > 0$ such that for any n

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu\|_{\mathcal{H}} \geq \sqrt{\frac{c_r}{n}} \right\} \geq \frac{1}{8}.$$

Proof. Denote $\Psi(\omega, \eta) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{k}(x, y) e^{-i\omega^T x} e^{-i\eta^T y}\, dx\, dy$. Note that

$$\begin{aligned}
&\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{k}(x, y)\, dP_0(x)\, dP_1(y) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{k}(x, y) \frac{1}{(2\pi\sigma_0^2)^{d/2}} e^{-\frac{1}{2\sigma_0^2}\|x\|_2^2} \frac{1}{(2\pi\sigma_1^2)^{d/2}} e^{-\frac{1}{2\sigma_1^2}\|y\|_2^2}\, dx\, dy
\end{aligned}$$

By Theorem 5.18 (Wendland, 2005) and Tonelli-Fubini Theorem (Dunley, 2002, Theorem 4.4.5), we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{k}(x, y)\, dP_0(x)\, dP_1(y) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{\tilde{k}(x, y)}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{\sigma_0^2\|\omega\|_2^2}{2}} e^{-i\omega^T x}\, d\omega \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{\sigma_1^2\|\eta\|_2^2}{2}} e^{-i\eta^T y}\, d\eta\, dx\, dy \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left[\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{k}(x, y) e^{-i\omega^T x} e^{-i\eta^T y}\, dx\, dy \right] e^{-\frac{\sigma_0^2\|\omega\|_2^2}{2}} e^{-\frac{\sigma_1^2\|\eta\|_2^2}{2}}\, d\omega\, d\eta \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \Psi(\omega, \eta) e^{-\frac{\sigma_0^2\|\omega\|_2^2}{2}} e^{-\frac{\sigma_1^2\|\eta\|_2^2}{2}}\, d\omega\, d\eta.
\end{aligned}$$

By applying the same theorems to the other terms of $d(\theta_0, \theta_1) = \|\theta_0 - \theta_1\|_{\mathcal{H}}^2$, we obtain

$$\begin{aligned}
d(\theta_0, \theta_1) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \Psi(\omega, \eta) \big(e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2} - e^{\frac{-\sigma_0^2}{2}\|\omega\|_2^2}\big) \big(e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} - e^{\frac{-\sigma_0^2}{2}\|\eta\|_2^2}\big)\, d\omega\, d\eta \\
&= \frac{1}{(2\pi)^d} \iint\limits_{\{\Psi > 0\}} \Psi(\omega, \eta) \big(e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2} - e^{\frac{-\sigma_0^2}{2}\|\omega\|_2^2}\big) \big(e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} - e^{\frac{-\sigma_0^2}{2}\|\eta\|_2^2}\big)\, d\omega\, d\eta \\
&\quad + \frac{1}{(2\pi)^d} \iint\limits_{\{\Psi < 0\}} \Psi(\omega, \eta) \big(e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2} - e^{\frac{-\sigma_0^2}{2}\|\omega\|_2^2}\big) \big(e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} - e^{\frac{-\sigma_0^2}{2}\|\eta\|_2^2}\big)\, d\omega\, d\eta
\end{aligned}$$

As regards the first term of $d(\theta_0, \theta_1)$, we apply the same steps as in Theorem 4.4 and we find that

$$\begin{aligned}
&\frac{1}{(2\pi)^d} \iint\limits_{\{\Psi>0\}} \Psi(\omega,\eta)\big(e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2} - e^{\frac{-\sigma_0^2}{2}\|\omega\|_2^2}\big)\big(e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} - e^{\frac{-\sigma_0^2}{2}\|\eta\|_2^2}\big)\, d\omega\, d\eta \\
&\geq \frac{(\sigma_1^2-\sigma_0^2)^2}{4(2\pi)^d} \iint\limits_{\{\Psi>0\}} \Psi(\omega,\eta) h'_\eta\big(\frac{\sigma_1^2}{2}\big) h'_\omega\big(\frac{\sigma_1^2}{2}\big)\, d\omega\, d\eta \\
&= \frac{(\sigma_1^2-\sigma_0^2)^2}{4(2\pi)^d} \iint\limits_{\{\Psi>0\}} \Psi(\omega,\eta)\|\eta\|_2^2\|\omega\|_2^2 e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2}\, d\omega\, d\eta.
\end{aligned}$$

As regards the second term of $d(\theta_0, \theta_1)$, define $g_\eta(x) = 1 - e^{-x\|\eta\|_2^2}$. Note that $g'_\eta(x) = \|\eta\|_2^2 e^{-x\|\eta\|_2^2}$ and $g''_\eta(x) = -\|\eta\|_2^4 e^{-\|\eta\|_2^2} < 0$, i.e., g_η is concave. Thus, we find that

$$\begin{aligned}
&\frac{1}{(2\pi)^d} \iint\limits_{\{\Psi<0\}} \Psi(\omega,\eta)\big(e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2} - e^{\frac{-\sigma_0^2}{2}\|\omega\|_2^2}\big)\big(e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} - e^{\frac{-\sigma_0^2}{2}\|\eta\|_2^2}\big)\, d\omega\, d\eta \\
&= \frac{1}{(2\pi)^d} \iint\limits_{\{\Psi<0\}} \Psi(\omega,\eta)\Big(g_\omega\big(\frac{\sigma_1^2}{2}\big) - g_\omega\big(\frac{\sigma_0^2}{2}\big)\Big)\Big(g_\eta\big(\frac{\sigma_1^2}{2}\big) - g_\eta\big(\frac{\sigma_0^2}{2}\big)\Big)\, d\omega\, d\eta \\
&\geq \frac{(\sigma_1^2-\sigma_0^2)^2}{4(2\pi)^d} \iint\limits_{\{\Psi<0\}} \Psi(\omega,\eta) g'_\eta\big(\frac{\sigma_1^2}{2}\big) g'_\omega\big(\frac{\sigma_1^2}{2}\big)\, d\omega\, d\eta \\
&= \frac{(\sigma_1^2-\sigma_0^2)^2}{4(2\pi)^d} \iint\limits_{\{\Psi<0\}} \Psi(\omega,\eta)\|\eta\|_2^2\|\omega\|_2^2 e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2}\, d\omega\, d\eta.
\end{aligned}$$

Therefore, we deduce that

$$\|\theta_0 - \theta_1\|_{\mathcal{H}}^2 \geq \frac{(\sigma_1^2-\sigma_0^2)^2}{4(2\pi)^d} \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \Psi(\omega,\eta)\|\eta\|_2^2\|\omega\|_2^2 e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2}\, d\omega\, d\eta.$$

We can explore further the condition on the right hand side. By using the definition for $\Psi(\omega, \eta)$, we have that

$$
\begin{aligned}
&\iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \Psi(\omega,\eta)\|\eta\|_2^2\|\omega\|_2^2 e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2}\, d\omega\, d\eta \\
&= \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d}\iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \tilde{k}(x,y) e^{-i\omega^T x} e^{-i\eta^T y}\, dx\, dy \|\eta\|_2^2\|\omega\|_2^2 e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2}\, d\omega\, d\eta \\
&= \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \tilde{k}(x,y) \left(\int\limits_{\mathbb{R}^d} e^{-i\omega^T x}\|\omega\|^2 e^{-\frac{\sigma_1^2}{2}\|\omega\|_2^2}\, d\omega\right)\left(\int\limits_{\mathbb{R}^d} e^{-i\omega^T y}\|\omega\|^2 e^{-\frac{\sigma_1^2}{2}\|\omega\|_2^2}\, d\omega\right) dx\, dy \\
&= \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \tilde{k}(x,y) \left(\sum_{i=1}^d \int\limits_{\mathbb{R}^d} e^{-i\omega^T x}\omega_i^2 e^{-\frac{\sigma_1^2}{2}\|\omega\|_2^2}\, d\omega\right)\left(\sum_{i=1}^d \int\limits_{\mathbb{R}^d} e^{-i\omega^T y}\omega_i^2 e^{-\frac{\sigma_1^2}{2}\|\omega\|_2^2}\, d\omega\right) dxdy
\end{aligned}
$$

By using Definition 2.7, we obtain

$$
\begin{aligned}
&\iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \Psi(\omega,\eta)\|\eta\|_2^2\|\omega\|_2^2 e^{\frac{-\sigma_1^2}{2}\|\eta\|_2^2} e^{\frac{-\sigma_1^2}{2}\|\omega\|_2^2}\, d\omega\, d\eta \\
&= \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \tilde{k}(x,y) \left(\sum_{i=1}^d \mathcal{F}\left[\omega_i^2 e^{-\frac{\sigma_1^2}{2}\|\omega\|_2^2}\right](x)\right)\left(\sum_{i=1}^d \mathcal{F}\left[\omega_i^2 e^{-\frac{\sigma_1^2}{2}\|\omega\|_2^2}\right](y)\right) dx\, dy \\
&= \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \tilde{k}(x,y) \left(\sum_{i=1}^d \frac{d^2}{dx_i^2}\frac{1}{(2\pi\sigma_1^2)^{d/2}} e^{-\frac{\|x\|_{\mathcal{H}}^2}{2\sigma_1^2}}\right)\left(\sum_{i=1}^d \frac{d^2}{dy_i^2}\frac{1}{(2\pi\sigma_1^2)^{d/2}} e^{-\frac{\|y\|_{\mathcal{H}}^2}{2\sigma_1^2}}\right) dx\, dy \\
&= \frac{1}{(2\pi\sigma_1^2)^d\sigma_1^4} \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \tilde{k}(x,y) \left(\sum_{i=1}^d \frac{d}{dx_i}\left[x_i e^{-\frac{\|x\|_{\mathcal{H}}^2}{2\sigma_1^2}}\right]\right)\left(\sum_{i=1}^d \frac{d}{dy_i}\left[y_i e^{-\frac{\|y\|_{\mathcal{H}}^2}{2\sigma_1^2}}\right]\right) dx\, dy \\
&= \frac{1}{(2\pi\sigma_1^2)^d\sigma_1^4} \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \tilde{k}(x,y) \left(\sum_{i=1}^d \left[1-\frac{x_i^2}{\sigma_1^2}\right] e^{-\frac{\|x\|_{\mathcal{H}}^2}{2\sigma_1^2}}\right)\left(\sum_{i=1}^d \left[1-\frac{y_i^2}{\sigma_1^2}\right] e^{-\frac{\|y\|_{\mathcal{H}}^2}{2\sigma_1^2}}\right) dx\, dy \\
&= \frac{1}{(2\pi\sigma_1^2)^d\sigma_1^4} \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \tilde{k}(x,y) \left(d-\frac{\|x\|_{\mathcal{H}}^2}{\sigma_1^2}\right) e^{-\frac{\|x\|_{\mathcal{H}}^2}{2\sigma_1^2}} \left(d-\frac{\|y\|_{\mathcal{H}}^2}{\sigma_1^2}\right) e^{-\frac{\|y\|_{\mathcal{H}}^2}{2\sigma_1^2}}\, dx\, dy \\
&= \frac{1}{(2\pi\sigma_1^2)^d\sigma_1^4} \left\| \int\limits_{\mathbb{R}^d} r(X)\big(d-\frac{\|x\|_{\mathcal{H}}^2}{\sigma_1^2}\big) e^{-\frac{\|x\|_{\mathcal{H}}^2}{\sigma_1^2}}\, dx \right\|_{\mathcal{H}}^2 .
\end{aligned}
$$

Since $\left\| \int\limits_{\mathbb{R}^d} r(x)\big(d - \frac{\|x\|_{\mathcal{H}}^2}{\sigma_1^2}\big) e^{-\frac{\|x\|_{\mathcal{H}}^2}{\sigma_1^2}}\, dx \right\|_{\mathcal{H}}^2 > 0$, there exists $c_r > 0$ such that $\|\theta_0 - \theta_1\|_{\mathcal{H}} \geq \sqrt{\frac{c_r}{n}}$. ■

Chapter 5

Conclusion and Discussion

In this thesis, we propose shrinkage estimators for the mean function in Hilbert space, based on Stein phenomenon. Our proposed estimators are of the form $\hat{\mu}_\alpha = \alpha f^* + (1-\alpha)\hat{\mu}$, where $\alpha \geq 0$ and f^* is an arbitrary fixed function in $\mathcal{H}$ that is independent from the sample, i.e., we consider a convex combination of the standard empirical mean estimator and an arbitrary function. We show that these estimators are improved versions of the empirical average as they achieve lower risk for a certain choice of α. Furthermore, we show that the rate of convergence of the shrinkage estimators is of the order $n^{-1/2}$ and it is optimal in the minimax sense.

While we investigate the theoretical properties of the shrinkage estimators in this thesis, we intend to study the empirical behavior of the shrinkage estimators over sample mean. By running some simulations or applying them to real-world data sets, we will be able to compare our proposed shrinkage estimators to the empirical one. Also, we plan to extend this work to obtain shrinkage estimators for $\int_{\mathcal{X}} \int_{\mathcal{Y}} r(x,y)\, d\mathbb{P}(x)\, d\mathbb{P}(y)$ and investigate their theoretical and empirical behavior.

Bibliography

[1] Ann Cohen Brandwein and William E. Strawderman. *Stein estimation: The spherically symmetric case.* Statistical Science 5, 356-69, 1990.

[2] Ann Cohen Brandwein and William E. Strawderman. *Stein estimation for spherically symmetric sistributions: Recent developments.* Statistical Science 27, 11-23, 2012.

[3] Lawrence D. Brown. *On the admissibility of invariant estimators of two-dimensional location parameters.* Mimeograph notes, Birkbeck College, London, 1965.

[4] Lawrence D. Brown. *On the admissibility of invariant estimators of one or more location parameters.* Annals of Statistics, 37, 1087-1135, 1966.

[5] Raymond Cattell. *Factor Analysis.* Harper, New York, 1952.

[6] John Diestel and John J. Uhl. *Vector Measures.* Americal Mathematical Society, Providence, RI, 1977.

[7] Nicolae Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces.* Wiley, 2000.

[8] Richard M. Dudley . *Real Analysis and Probability.* Cambridge University Press, 2002.

[9] Bradley Efron and Carl Morris. *Steins paradox in statistics.* Sci. Am. 236 (5), 119127, 1975.

[10] Benjamin Fruchter . *Introduction to Factor Analysis.* Van Nostrand, 1954.

[11] Marvin Gruber. *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators.* Hong Kong, Taylor & Francis, 1998.

[12] John A. Hartigan and Anthony Wong. *Algorithm AS 136: A k-Means clustering algorithm.* Journal of the Royal Statistical Society, Series C, 28, 1979.

[13] W James and Charles Stein. *Estimation with quadratic loss.* Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, 361-379, University of California Press, Berkeley, California, 1961.

[14] Ian Jolliffe. *Principal Component Analysis.* Springer, New York, 1986.

[15] Jack Kiefer. *Invariance, minimax sequential estimation and continuous time processes.* Ann. Math. Statist. 28 573-601, 1957.

[16] Erwin Kreyszig. *Introductory Functional Analysis with Applications.* Wiley, 1989.

[17] Solomon Kullback. *Information Theory and Statistics.* Wiley, New York, 1959.

[18] John B. MacQueen. *Some methods for classification and analysis of multivariate observations.* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281–297, University of California Press, Berkeley, California, 1967.

[19] Geoffrey McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.

[20] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Arthur Gretton, and Bernhard Scholkopf. *Kernel mean estimation and Stein effect.* ICMP, 10-18, 2014a.

[21] Krikamol Muandet, Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, and Bernhard Scholkopf. *Kernel mean shrinkage estimators.* Journal of Machine Learning Research, 2016.

[22] Walter Rudin . *Real and Complex Analysis.* McGram-Hill, 3rd edition, 1987.

[23] Charles Stein. *Inadmissibility of the usual estimator for the mean of the multivariate normal distribution.* Procceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 , 197-206, University of California Press, Berkeley, California, 1955.

[24] Ingo Steinwart and Andreas Christmann. *Support Vector Machines.* Springer, New York, 2008.

[25] Ilya Tolstikhin, Bharath Sriperumbudur and Krikamol Muandet . *Minimax estimation of kernel mean embeddings.* Journal of Machine Learning Research, 18, 1-47, 2017.

[26] Alexandre B. Tsybakov. *Introduction to Non Parametric Estimation.* Springer, New York, 2008.

[27] Holger Wendland. *Scattered Data Approximation.* Cambridge University Press, Cambridge, UK, 2005.

[28] Vadim Yurinsky. *Sums and Gaussian Vectors.* Volume 1617, Lecture Notes in Mathematics, Springer, Berlin, 1995.

[29] Yang Zhou, Di-Rong Chen and Wei Huang. *A class of optimal estimators for the covariance operator in Reproducing Kernel Hilbert Spaces.* Journal of Multivariate Analysis, 166-178, 2019.