CONTRIBUTIONS TO STATISTICAL LEARNING AND ITS APPLICATIONS IN PERSONALIZED MEDICINE

A Thesis
Presented to
The Academic Faculty

by

Carlos Felipe Valencia Arboleda

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Enginnering

Georgia Institute of Technology
August 2013

CONTRIBUTIONS TO STATISTICAL LEARNING AND ITS APPLICATIONS IN PERSONALIZED MEDICINE

Approved by:

Ming Yuan, Advisor
H. Milton Stewart School of Industrial
and Systems Enginnering
Georgia Institute of Technology

Paul Kvam
H. Milton Stewart School of Industrial
and Systems Enginnering
Georgia Institute of Technology

Xiaoming Hou
H. Milton Stewart School of Industrial
and Systems Enginnering
Georgia Institute of Technology

Nicoleta Serban
H. Milton Stewart School of Industrial
and Systems Enginnering
Georgia Institute of Technology

Xin Qi
Department of Mathematics & Statistics
Georgia State University

Date Approved: 30 April 2013

In memory of Melisa.

I also dedicate my dissertation to my parents: the man who told me how to fly and the woman who told me how to keep my feet on the ground.

ACKNOWLEDGEMENTS

I want to start by expressing my sincere and deep gratitute to my thesis advisor, Professor Ming Yuan. During the last years of my doctoral studies he suppported and oriented my research activities. He introduced me to the fields of statistical learning and nonparametric statistics; and his knowledge, enthusiasm and academic skills continuously inspired me to finish my research. I appreciate his patience and devotion during this long venture. It has been an honor and a pleasure to work with Ming.

I am also very thankful to the people in my thesis committee: Dr. Paul Kvam, Dr. Xiaoming Huo, Dr. Nicoleta Serban and Dr. Xin Qi. They generously accepted to evaluate my thesis and their comments and suggestions contributed to my understanding of the research topics and improved the presentation of this final document. Special thanks to Dr. Kvam and Dr. Serban for agreeing to write references on my behalf.

I would like also to express my deepest gratitude to all my friends at Georgia Tech that shared with me all the challenges of this journey. Enlightening discussions and enjoyable moments with them not just made easier the work, but also made me a better person. Thanks to them, these years will last in my memory as a remarkable experience. Knowing that I can not name all of them, I just want to say thanks to Dr. Justin Vastola, Dr. Norbert Remenyi, Dr. Dong Gu Choi and Dr. Huizhi Xie.

Finally, I am endlessly grateful to my beloved family: my parents and my sister. From the distance they gave me strength to achieve my goals.

TABLE OF CONTENTS

DEDICATION . . . . **iii**

ACKNOWLEDGEMENTS . . . . **iv**

LIST OF TABLES . . . . **vii**

LIST OF FIGURES . . . . **viii**

SUMMARY . . . . **ix**

I REPRODUCING KERNEL HILBERT SPACE APPROACH TO GENERAL FUNCTIONAL LINEAR REGRESSION FOR EXPONENTIAL FAMILIES **1**

1.1 Introduction. 1

1.2 Methodology 3

1.2.1 Regularization Estimation 4

1.2.2 Derivatives and Approximation by Linearization 6

1.2.3 Simultaneous Diagonalization, Norms and Inverse Operators 7

1.2.4 Bias and Variance Approximation 13

1.3 Convergence Rates 14

1.4 Numerical Studies 15

II RADIAL BASIS REGULARIZATION FOR LINEAR INVERSE PROBLEMS WITH RANDOM NOISE **19**

2.1 Introduction 19

2.2 Radial Basis Function Regularization in Linear Inverse Problems 21

2.2.1 Sequence model via singular value decomposition 21

2.2.2 Parameter spaces 22

2.2.3 Radial basis function regularization 23

2.3 Main Results 24

2.4 Risk Analysis of Radial Basis Function Regularization 28

2.5 Numerical Experiments 30

2.6 Proof of Theorem 9 31

III A DISCRIMINANT APPROACH TO TREATMENT SELECTION. INFINITE SAMPLE CONSISTENCY ANALYSIS 36

3.1 Introduction 36

3.2 Theoretical Framework and Methodology 39

3.3 Infinite Sample Consistency. Main Results. 42

3.3.1 Treatment Selection Withholding 45

3.3.2 Multicategory Treatment Selection 48

APPENDIX A — RELEVANT PROOFS IN CHAPTER I 54

REFERENCES 66

LIST OF TABLES

1 We list here, for different combinations of parameter spaces and radial basis functions, the best achievable convergence rates of $\mathcal{R}(\lambda)$ and the order of the tuning parameter λ needed to attain the rate. α reflects the smoothness of the parameter space, β determines the ill-posedness of the inverse problem and γ depends on the choice of the radial basis function. 29

LIST OF FIGURES

1 Estimation errors in logarithmic scale. 17

2 Prediction errors in logarithmic scale. 18

3 Estimation and Prediction errors in logarithmic scale when the eigenvalues in the covariance operator are closely spaced. 18

4 Comparison of the risk of radial basis function regularization. The results are averaged over 100 replications. 31

5 Comparison of different Surrogate Value fucntions. 41

SUMMARY

The statistical anaysis of data has been regularly shaped by the challenges that new problems in science, engineering and finance have to offer. During recent years, the advances in computational information handling and the development of new technologies capable of measuring and storing massive amount of data, tilted the attention of many researches toward new kinds of statistical problems mainly characterized by a vast increment in size and complexity. The exploded number of data points favored the use of non-parametric models, in particular for prediction and feature extraction. The complexity of the data implied the use of new approaches for dealing with high dimensional problems and functional data. Using classical inference methods for fitting data in this framework lead to non-stable solutions due to the large number of parameters.

Although different in nature, the majority of these new statistical techniques require the solution of an ill-posed problem. Given the inherent over-parametrization, it is necessary to add assumptions that produce regular solutions. In general, some restrictions need to be attached to the class of models that are fitted, for example, structural smoothness is assumed in the case of non-parametric function estimation, or sparsity in the case of high dimensional linear models.

The present dissertation, in general, is about finding stable solutions to statistical models with very large number of parameters and to analyze their asymptotic statistical properties. In particular, it is centered in the study of regularization methods based on penalized estimation. Those procedures find an estimator that is the result of an optimization problem balancing out the fitting to the data with the plausability of the estimation. Adding a penalization term, in conjunction with corresponding assumptions about the class of possible values for the true parameter, permits to create a better trade-off between the deterministic and the stochastic errors. As pointed out in Bickel and Li (2006), the particular combination of restrictions and penalty may serve two purposes: to construct a good predictor and to select the relevant variables in the prediction. This thesis has three parts, each one contained in a different chapter.

The first chapter studies a smoothness regularization estimator for an infinite dimensional parameter in an exponential family model with functional predictors. The main objective is to analysis the asymptotic statistcal properties of the proposed estimator. We focused on the Reproducing Kernel Hilbert space approach and show that regardless the generality of the method, minimax optimal convergence rates are achieved. This research project enhances the set of tools that can be used for Functional Data Analysis, specifically, allowing for a computational convenient estimation of the regression problem in settings where the response is discrete or the zero mean additive error assumption is not appropiate. In order to derive the asymptotic analysis of the estimator, we developed a simultaneous diagonalization tool for two positive definite operators: the kernel operator and the operator defined by the second Frechet derivative of the expected data fit functional. By using the proposed simultaneous diagonalization tool we obtained sharper bounds on the minimax rates.

The second chapter studies the statistical properties of the method of regularization using Radial Basis Functions in the context of linear inverse problems. Radial basis function regularization is widely used in machine learning because of its demonstrated effectiveness in numerous applications and computational advantages. We consider a known compact linear operator A from $\mathcal{L}_2(-\pi,\pi)$ to $\mathcal{L}_2(-\pi,\pi)$, and suppose that y is observable, where $Af = y$. Even if the range of A is dense, finding f by inverting the operator A can be an ill-conditioned problem. The regularization here serves two purposes, one is creating a stable solution for the inverse problem (A^{-1}) and the other is prevent the over-fitting on the nonparametric estimation of f_0. The particular interest in this project is to analysis the statistical properties of the estimator $\hat{f}_\lambda$. Different degrees for the the ill-posedness in the inversion of the operator A are considered: mildly and severly ill-posed. Also, we study different types fo radial basis kernels classified by the strength of the penalization norm: Gaussian, Multiquadrics and Spline type of kernels.

The third chapter deals with the problem of Individualized Treatment Rule (ITR) and analyzes the solution of it trough Discriminant Analysis. The ITR problem is one of the primary interest of personalized medicine, where the treatment (or type of medicine) assignment is done based on the particular patient's prognosis covariates in order to maximizes some reward function (response to treatment). Data generated from a random clinical trial is considered. Given that the different treatments form a categorical set, maximizing the empirical value function is an NP-hard computational problem. The usual approach in the literature is a two stage procedure, where first the mean response is estimated and then the estimator is maximized according to the best rule. This approach is prone to generate suboptimal solutions when the functional class considered is not rich enough. We consider estimating directly the decision rule by maximizing the expected value, usign a surrogate function in order to make the optimization problem computationaly feasible (convex programming). Necessary and sufficient conditions for Infinite Sample Consistency on the surrogate function are found for different scenarios: binary treatment selection, treatment selection with witholding and milti-treatment selection.

CHAPTER I

REPRODUCING KERNEL HILBERT SPACE APPROACH TO GENERAL FUNCTIONAL LINEAR REGRESSION FOR EXPONENTIAL FAMILIES

1.1 Introduction.

Many statistical analyses require the process and manipulation of data that take the form of random curves. This is the result of new technologies, arising in different research fields, that can produce massive number of observations. Those curves are usually smoothed versions of longitudinal data measured over a very fine grid of points that can be modeled as functional data. During recent years many new statistical methodologies have been developed for functional data analysis. A compendium of these advances can be found in Ferraty and Romain (2010) and Ramsay and Silverman (2005). Special attention has been directed to the modeling of a scalar response with functional predictors, being the functional linear regression the most renowned case (see, e.g., Cardot, Ferraty and Sarda (2003), Cai and Hall (2006) and Yuan and Cai (2010)). However, in numerous applications there exist some restrictions on the characterization of the response variable, for instance when this response is categorical or when the usual zero mean additive error assumption does not seem to be appropiate. A natural alternative is the use a generallized linear model adapted for a functional predictor.

We consider the functional generalized linear model where the response Y follows a probability distribution in the exponential family with density

$$f_{\theta_0}(y) = \exp(\theta_0 y - \omega(\theta_0)), \tag{1}$$

and canonical parameter $\theta_0(X) = \left[\alpha_0 + \int_\tau \beta_0 X\right] \in \mathbb{R}$, with X being a second order stochastic process on a compact domain τ. β_0 is the unknown slope function and α_0 is the unknown scalar intercept. We assume one observes a training data $(x_1, y_1), \cdots, (x_n, y_n)$ consisting of the realization of n independent copies of (X, Y). Our purspose is to estimate the slope parameter β_0 and, based on it, present a point estimator for $\theta_0(X)$.

We propose in this chapter a regularization procedure for estimating the aforementioned parameters, for which, we will assume hereafter that β_0 belongs to a reproducing kernel

Hilbert space $\mathcal{H} \subset \mathcal{L}_2(\tau)$. In general, the method of regularization combines two non-negative functionals of the parameters (α, β). The first one is a data fit functional $\ell_n(\alpha, \beta)$ that measures how well the data is explained as a realization of a random sample with associated densities $f_{\theta_i}(y_i)$ and $\theta_i = \alpha + \int_\tau \beta x_i$. We shall use the negative loglikelihood of the data as the data fit functional. The second functional is a penalty term $J(\beta)$ that prevents the overfitting of the estimator by giving less chance of being selected to solutions that are not plausable. We choose $J(\beta)$ as a norm (or semi-norm) in the reproducing kernel Hilbert space $\mathcal{H}$. Therefore, the method of regularization estimates (α_0, β_0) by

$$\left(\hat{\alpha}_{n\lambda}, \hat{\beta}_{n\lambda}\right) = \underset{c\in\mathbb{R}, \beta\in\mathcal{H}}{\arg\min}\, \ell_n(c, \beta) + \lambda J(\beta), \tag{2}$$

where $\lambda \geq 0$ is the tuning parameter that balances out the two criteria represented by ℓ_n and J respectively. Note that the selection of $\mathcal{H}$ and J changes profoundly the nature of the numerical estimation that, in general, is a minimization problem on an infinite dimensional space. More details can be found in Wahba (1990).

The generalized linear model with functional predictor has been the subject of some previous investigations with a variety of estimation methods other than the reproducing kernel Hilbert space regularization. For example, James (2002) investigate some fitting algorithms for applied cases. Müller and Stadmüller (2005) studied quasi likelihood estimation for a truncated version of the slope coefficients in the Karhunen-Loève expansion produced by the covariance operator of $X(t)$. Cardot and Sarda (2005) proposed a B-splines estimator and presented a $\mathcal{L}_2$ rate of convergence. Dou, Pollard & Zhou (2010) extended the theory developed by Hall and Horowitz (2007) for observations coming from an exponential family. They proposed a maximun likelihood estimator for a finite dimensional projection of the canonical parameter $\theta(X)$ using the basis generated by functional principal components estimation of the covariance operator. The method of regularization for estimating the slope parameter that we investigate in the present paper offers some numerical or theoretical advantages when compared with the aforementioned procedures. In particular, it possesses the capacity of being solved numerically as a finite dimensional convex problem, but with the generality of being adapted to many reproducing Hilbert spaces associated with different types of kernel. An important theoretical property of the method of regularization is that it does not depend on the functional principal component analysis (FPCA). Therefore, it avoids some restrictive assumptions on the spacing between the eigenvalues of the covariance operator and on the Fourier coefficients of β_0 with respect to the basis generated by the principal components. In the FPCA-based methods, the estimator of β_0 is a projection on the subspace generated by the span of a truncated sequence of the eignefunctions obtained by the estimation of the principal components. That implies that a condition for the success of these approaches is that β_0 can be expressed efficiently in the principal components basis. The method of regularization estimator on the contrary, performs a shrinkage that penalizes rough solutions in terms of the $\mathcal{H}$ basis generated by the kernel operator, obtaining a smoother solution in relation to the selected kernel function.

In section 1.2 we provide the theoretical framework necessary for the aymptotic analysis of the regularization estimator. Section 1.3 presents the main results concerning the convergence rates for the estimators of β_0 and the prediction $\theta_0(X)$. Some numerical results that ilustrate the benefits of the method are shown in section 1.4. The most relevant proofs of our results are presented in Appendix A.

1.2 Methodology

We consider a family of exponential probability measures $\{F_\theta : \theta \in \mathbb{R}\}$ with corresponding densities $f_\theta(y) = \exp(y\theta - \omega(\theta))$. Using the notation $\omega^{(r)}$ to represent the r-th derivative of the fucntion $\omega(\cdot)$, the mean and the variance of a random variable with associated density f_θ are $\omega^{(1)}(\theta)$ and $\omega^{(2)}(\theta)$ respectively. In order to avoid degenerated cases it is assumed that $\omega^{(2)}(\theta) > 0$ for all $\theta \in \mathbb{R}$.

Suppose one observes the realization of n random data points $(X_1, Y_1), \cdots, (X_n, Y_n)$ as independent copies of (Y, X), where $X := \{X(t) : t \in \tau\}$ is a second order stochastic process. Let P_{yx} be the joint probability distribution of Y and X, P_x the marginal distribution of the random process X and $P_{y|x}$ the conditional distribution of Y given X. We assume that $P_{y|x} = F_{\theta_0(X)}$ for $\theta_0(X) = \alpha_0 + \int_\tau \beta_0 X$. For the second order stochastic process X taking values in $\mathcal{L}_2(\tau)$, the distribution P_x is characterized by its mean $\mu(t)$ and its covariance function $C(s,t)$ for $s, t \in \tau$. It is assumed that $C(\cdot,\cdot)$ is continuous and square integrable with some aditional conditions that will be discussed in section 1.2.3. Without loss of generality we will suposse that $X(t)$ is a centered process, e.g., $\mu(t) = 0$ for $t \in \tau$.

In order to develop the statistical analysis of our proposed estimators for β_0 and $\theta_0(X)$, we make the following assumptions concerning the distribution P_{yx}.

Assumption 1 *Assumptions on P_{yx}.*

(i) $\omega(\cdot)$ is three times continuous differentiable on $\mathbb{R}$.

(ii) There exists a constant M such that for any $f \in \mathcal{L}_2(\tau)$,

$$\mathbb{E}\left(\int_\tau f(t)\left[X(t)-\mu(t)\right]dt\right)^4 \leq M\left[\mathbb{E}\left(\int_\tau f(t)\left[X(t)-\mu(t)\right]dt\right)^2\right]^2. \tag{3}$$

(iii) For all $f \in \mathcal{L}_2(\tau)$ and $r = 1, 2, 3$

$$\mathbb{E}\left[\omega^{(r)}\left(\int_\tau Xf\right)\right]^4 < \infty. \tag{4}$$

(iv) For all $\beta, f, g, h \in \mathcal{L}_2(\tau)$,

$$\mathrm{Var}\left[\omega^{(2)}\left(\int_\tau X\beta\right)\left(\int_\tau Xf\right)\left(\int_\tau Xg\right)\right] < \infty, \tag{5}$$

$$\mathrm{Var}\left[\omega^{(3)}\left(\int_\tau X\beta\right)\left(\int_\tau Xf\right)\left(\int_\tau Xg\right)\left(\int_\tau Xh\right)\right] < \infty. \tag{6}$$

Assumption 1(i) restricts the set of exponential distributions for which our results apply. For example, the canonical parametrizatoin of a Gamma or an Inverse Gaussian distribution implies that this assumption does not hold (see e.g. McCullagh and Nelder (1989)). Nevertheless, many cases of interest are cover under this requirement. Assumption 1(ii) bounds the fourth moment of any linear funcional of X and is satisifed for a Gaussian process (with $M = 3$) for example. Although the assumption 1(iii) is stronger, we consider that still cover many of the interest cases for application. For example, if X is a Gaussian process we know that $\int_\tau Xf$ is a normal random varaible and taking $\omega(\theta) = \exp(\theta)$ it is easily checked that this assumption holds.

1.2.1 Regularization Estimation

In this paper we focus on the estimation of the function parameter β_0 using a smoothness regularization methodology. Before presenting the analysis of the statistical asymptotic properties of the resulting estimator, we start by disscusing some aspects on the characterization of the computational estimation procedure. Our purpose is to present the estimator defined in 2 as the solution of a finite dimensional convex minimization problem.

The main idea of the method of regularization is to introduce a penalty term $J(\beta)$ to prevent the overfitting in the solution. In general $J(\beta)$ penalizes the roughness of the solution, and in particular, for the Reproducing Hilbert Space approach $J(\beta)$ corrresponds to a squared norm (or a seminorm) associated with the reproducing Hilbert space $\mathcal{H}$. Let $\mathcal{H}^0$ be the null space formed by $J(\cdot)$ in $\mathcal{H}$, that is, $\mathcal{H}^0 = \{\beta \in \mathcal{H} : J(\beta) = 0\}$. It follows that $\mathcal{H}^0$ is a linear subspace of $\mathcal{H}$, for which it is required that $\dim(\mathcal{H}^0) = M \leq n$. We denote by

$\{\xi_1, \cdots, \xi_M\}$ an orthonormal basis for $\mathcal{H}^0$. Let $\mathcal{H}^1$ be the orthogonal complement of $\mathcal{H}^0$ such that $\mathcal{H} = \mathcal{H}^0 \oplus \mathcal{H}^1$. Thus, for any $f \in \mathcal{H}$ it is possible to write $f = \sum_{i \leq M} c_i \xi_i + f_1$, where $(c_1 \cdots, c_m) \in \mathbb{R}^M$ and f_1 is the projection of f in $\mathcal{H}^1$. In this context, $\mathcal{H}^1$ forms a reproducing kernel Hilbert space for which there exists one associated reproducing kernel $K : \tau \times \tau \to [0, \infty)$ such that, if $f \in \mathcal{H}^1$, $J(f) = \|f\|_{\mathcal{H}}^2 = \|f\|_K^2$. It will be assumed hereafter that $K(\cdot, \cdot)$ is continuous on $\tau \times \tau$ and square integrable.

Associated with the kernel $K(\cdot, \cdot)$ there is a nonnegative definite operator on $\mathcal{L}_2$ constructed as

$$[Kf](t) = \int_\tau K(s,t) f(s) ds \tag{7}$$

for any $f \in \mathcal{L}_2$. In what follows this operator will be represented as $Kf : \mathcal{L}_2 \to \mathcal{H}^1$ for brevity in the notation. The main computational advantage of using reproducing kernel estimators comes from the reproducing property of this operator that allows to obtain the representer of any bounded linear functional in $\mathcal{H}$. In our particular case, for any $\beta \in \mathcal{H}$ and a fixed $x \in \mathcal{L}_2$ we are interested in the functional $\int_\tau x(t)\beta(t)dt$, for which there exists a representer function $\eta_x \in \mathcal{H}$ such that

$$\int_\tau x(t)\beta(t)dt = \langle \eta_x, \beta \rangle_{\mathcal{H}},$$

and $\eta_x = Kx$. Some further details about the properties of the RKHS can be found in Aronszajn (1950) and Cucker and Smale (2001), and inside the context of general smoothing spline models, in Wahba (1990).

Recall that the regularization estimator is given by

$$(\hat{\alpha}_{n\lambda}, \hat{\beta}_{n\lambda}) := \underset{c \in \mathbb{R}, \beta \in \mathcal{H}}{\arg\min}\, \ell_{n\lambda}(c, \beta), \tag{8}$$

where $\ell_{n\lambda}(c, \beta) = \ell_n(c, \beta) + \lambda J(\beta)$. The data fit functional $\ell_n(c, \beta)$ used in 8 is the negative of the empirical loglikelihood, that is

$$\ell_n(c, \beta) = \frac{1}{n} \sum_{i \leq n} \left[\omega\left(c + \int_\tau X_i \beta\right) - Y_i \left(c + \int_\tau X_i \beta\right)\right]. \tag{9}$$

The convexity of $\ell_{n\lambda}(c, \beta)$ is guaranteed by the condition $\omega^{(2)}(\theta) > 0$ for all $\theta \in \mathbb{R}$. Although the solution 8 for $\hat{\beta}_{n\lambda}$ is defined in an infinite dimensional parameter space, the Representer theorem for spline models (see Wahba (1990)) assures that the estimator can presented as the solution of a convex minimization problem in finite dimensions. More specifically, the estimator $\hat{\beta}_{n\lambda}$ will be an element of $\mathcal{H}$ and there exist $\mathbf{d} = (d_1, \cdots, d_M) \in \mathbb{R}^M$ and $\mathbf{b} = (b_1, \cdots, b_n) \in \mathbb{R}^n$ such that

$$\hat{\beta}_{n\lambda} = \sum_{i=1}^M d_i \xi_i + \sum_{j=1}^n c_j K x_i. \tag{10}$$

This result facilitates the numerical implementation of the method given that the optimization problem is solve on the variables $c, \mathbf{d}$ and $\mathbf{b}$; and also helps to develop a simpler analysis for the statistical properties of $\hat{\beta}_{n\lambda}$.

1.2.2 Derivatives and Approximation by Linearization

Our main objective is to analysis the large sample behavior of the regularization estimator defined in 8. We place particular emphasis on the estimator for the function β_0 and in order to achieve more clarity in the presentation we will assume in what follows that $\alpha_0 = 0$. The major purpose of this study is to find the asymptotic rates of convergence for a properly defined risk function evaluated on $\left(\hat{\beta}_{n\lambda}, \beta_0\right)$. We begin by setting up the notation and some definitions that play a fundamental role in the functional analysis of the penalized likelihood $\ell_{n\lambda}$ and others related functionals on $\mathcal{H}$.

We use the notation D to represent the Fréchet derivative operator in a general normed linear space. Recall that for a functional $\ell : \mathcal{H} \to \mathbb{R}$, if ℓ is differentiable at β then $D\ell(\beta)$ is a bounded functional defined on $\mathcal{H}$. Higher order derivatives will be represented as $D^r\ell$, for $r = 2, 3$. It follows that $D^2\ell(\beta) \in \mathbf{B}(\mathcal{H}, \mathcal{H})$, where $\mathbf{B}(\mathcal{H}, \mathcal{H})$ is the class of all boundel linear operators in $\mathcal{H}$. For $f, g \in \mathcal{H}$ we use the notation $D^2\ell(\beta)fg$ to represent $\langle f, D^2\ell(\beta)g\rangle_{\mathcal{H}}$.

It will be assumed that the negative loglikelohood $\ell_n(\beta)$ converges to a limiting functional $\ell_\infty(\beta) := \mathbb{E}\ell_n(\beta)$ for all $\beta \in \mathcal{H}$. The purpose of this limiting functional is to characterize the target parameter β_0 as its minimizer. Note that for any $f \in \mathcal{H}$ and a fixed β,

$$D\ell_\infty(\beta)f = \mathbb{E}_X\left[\left(\int_\mathcal{T} Xf\right)\left(\omega^{(1)}\left(\int_\mathcal{T} X\beta\right) - \omega^{(1)}\left(\int_\mathcal{T} X\beta_0\right)\right)\right]. \tag{11}$$

CauchySchwarz inequality and assumptions 1 assures that $D\ell_\infty(\beta)f < \infty$ and therefore, $D\ell_\infty(\beta_0)f = 0$. The convexity of $\omega(\cdot)$ implies that ℓ_∞ achieves its minimun at β_0.

Similary, a limiting functional for the penalized negative loglikelihood $\ell_{n\lambda}$ is defined as $\ell_{\infty\lambda}(\beta) := \mathbb{E}\ell_{n\lambda}(\beta) = \ell_\infty(\beta) + \lambda J(\beta)$. The introduction of this functional allows us to define $\beta_{\infty\lambda}$ as the argument that minimizes it on $\mathcal{H}$, and consequently, for any $f \in \mathcal{H}$, $D\ell_{\infty\lambda}(\beta_{\infty\lambda})f = 0$. Note that for smaller values of λ, $\beta_{\infty\lambda}$ is expected to approach the target parameter β_0.

With these definitions, the estimation error can be decomposed as

$$\left(\hat{\beta}_{n\lambda} - \beta_0\right) = \left(\hat{\beta}_{n\lambda} - \beta_{\infty\lambda}\right) + \left(\beta_{\infty\lambda} - \beta_0\right). \tag{12}$$

The two terms at the right can be understood as the squared bias and the stochastic variability. To be consistent with the previous literature (see e.g. Cox (1988) and Yuan and

Cai (2010)) they will be called Deterministic error and Stochastic error respectively.

In order to facilitate the asymptotic analysis, we approximate the roots $\hat{\beta}_{n\lambda}$ and $\beta_{\infty\lambda}$ by their linearized forms derived from first order Taylor series expansion. Let $G_{\infty\lambda}(\beta) := D^2\ell_{\infty\lambda}(\beta)$, then we define

$$\bar{\beta}_{\infty\lambda} := \beta_0 - G_{\infty\lambda}^{-1}(\beta_0) D\ell_{\infty\lambda}(\beta_0) \tag{13}$$

$$\bar{\beta}_{n\lambda} := \beta_{\infty\lambda} - G_{\infty\lambda}^{-1}(\beta_{\infty\lambda}) D\ell_{n\lambda}(\beta_{\infty\lambda}). \tag{14}$$

Note that in definition 14, $G_{\infty\lambda}$ has been used to make the linearization instead of the usual form $D^2\ell_{n\lambda}$. For the purpose of this section, the existence of these linearizations is assumed to be true. Further analysis about the operator $G_{\infty\lambda}$, its inverse and other high order derivatives will be discussed in section 1.2.3. Using the Taylor series expansion of $\ell_{\infty\lambda}$ around β_0, it is easy to check that

$$\left(\bar{\beta}_{\infty\lambda} - \beta_{\infty\lambda}\right) = G_{\infty\lambda}^{-1}(\beta_0) \int_{\mathcal{T}} \int_{\mathcal{T}} x_1 \left[D^3\ell_{\infty} \left(\beta_0 + x_1 x_2 \phi_1\right) \phi_1 \phi_1 \right] dx_1 dx_2, \tag{15}$$

where $\phi_1 = (\beta_{\infty\lambda} - \beta_0)$. Similary, using the expansion of $\ell_{n\lambda}$ around $\beta_{\infty\lambda}$, for $\phi_2 = \left(\hat{\beta}_{n\lambda} - \beta_{\infty\lambda}\right)$ it follows that

$$\begin{aligned} \left(\bar{\beta}_{n\lambda} - \hat{\beta}_{\infty\lambda}\right) &= G_{\infty\lambda}^{-1}(\beta_{\infty\lambda}) D^2\ell_{n\lambda}\phi_2 - \phi_2 \\ &+ G_{\infty\lambda}^{-1}(\beta_{\infty\lambda}) \int_{\mathcal{T}} \int_{\mathcal{T}} x_1 \left[D^3\ell_n \left(\beta_{\infty\lambda} + x_1 x_2 \phi_2\right) \phi_2 \phi_2 \right] dx_1 dx_2. \end{aligned} \tag{16}$$

These first order approximations will play a strategic role in further analysis as long as $\hat{\beta}_{n\lambda}$ and $\beta_{\infty\lambda}$ will be replaced by their linearized counterparts in the error definition 12.

1.2.3 Simultaneous Diagonalization, Norms and Inverse Operators

Before we address the asymptotic analysis of the estimator 8, it is necessary to identify some operators on $\mathcal{H}$ that are involved in the analysis and to define a plausible set of norms on which the convergence rates can be derived. One of the common characteristics of the estimators obtained by the method of regularization is that the solution of 8 is expressed as a combination of operators that come separately from the data fit functional ℓ_n and the penalty term J. Consider for example the case of the sum of squared errors, that is, when $\ell_n(\beta) = \sum_{i \leq n} \left(y_i - L_i(\beta)\right)^2$ for some functionals L_i. The resulting regularization estimator has the form $\hat{\beta}_{n\lambda} = \left(D^2\ell_{n\lambda}\right)^{-1} \sum_{n \leq n} y_i L_i^*$, where L_i^* stands for the adjoint of L_i (see Cox(1988) for more details), and $D^2\ell_{n\lambda} = D^2\ell_n + \lambda D^2 J$ does not depend on β because $\ell_{n\lambda}$ is a quadratic functional. In order to study the large sample properties of $\hat{\beta}_{n\lambda}$, it is very helpful to determine the eigensystem structure $D^2\ell_{n\lambda}$, and consequently it is important to find a simultaneous diagonalization for $D^2\ell_n$ and D^2J.

In this section, we will derive a simultaneous diagonailzation for two operators related with $D^2\ell_{n\lambda}$, the covariance operator associated with the covariance function of $X(t)$ and the operator K defined in 19. This problem has been addressed already in Yuan and Cai (2010) and (2012). This result will be used to approximate a diagonalization for $G_{\infty\lambda}(\beta)$ in the linearizations 13 and 14 and to constrct a more convenient equivalent norm in $\mathcal{H}$.

By Mercer's theorem, the kernel $K(\cdot,\cdot)$ is susceptible to the following spectral decomposition

$$K(s,t) = \sum_{k=1}^{\infty} \rho_k \psi_k(s)\psi_k(t), \tag{17}$$

where $\{\psi_1, \psi_2, \cdots\}$ and $\rho_1 \geq \rho_2 \geq \cdots$ represent respectively the eigenfunctions and eigenvalues of the operator K. It follows that $K\psi_k = \rho_k\psi_k$ for all $k \geq 1$. Recall that $\{\psi_1, \psi_2, \cdots\}$ form an orthonormal basis for $\mathcal{L}_2$.

Similary, given that the covariance function $C(\cdot,\cdot)$ is assumed continuous and square integrable, it is possible to write

$$C(s,t) = \sum_{k=1}^{\infty} \mu_k \phi_k(s)\phi_k(t), \tag{18}$$

where $\{\phi_1, \phi_2, \cdots\}$ and $\mu_1 \geq \mu_2 \geq \cdots$ define the eigen structure of the positive definte operator on $\mathcal{H}$

$$\left[Cf\right](t) = \int_{\mathcal{T}} C(s,t)f(s)ds, \tag{19}$$

and consequentely, $C\phi_k = \mu_k\phi_k$ and $\langle \phi_k, \phi_j\rangle_{\mathcal{L}_2} = \delta_{kj}$, where δ_{kj} is the Kronecker's delta.

Note that the eigenfunctions associated to K and C respectively are not necessarily related, as they can form a completely different set of functions. However, we need to find a common structure in which a combination of the two operators can be diagonalized. More concretely, we need to find a linear structure for the operator $C + \lambda D^2 J$, with $\lambda \geq 0$. We start by defining a new norm for every $f \in \mathcal{H}$ as

$$\|f\|_R^2 = \langle f, Cf\rangle_{\mathcal{L}_2} + J(f). \tag{20}$$

Note that if $\langle f, Cf\rangle_{\mathcal{L}_2}$ is strictly positive for all $f \neq 0$ in $\mathcal{H}_0$, then $\|\cdot\|_R$ is a well defined norm for all functions in $\mathcal{H}$. We will make this assumption thereafter. From Proposition 2 in Yuan and Cai (2010), it follows that $\|\cdot\|_R$ and $\|\cdot\|_{\mathcal{H}}$ are equivalent norms, that is, there exist constants $0 < a \leq b$ such that for all $f \in \mathcal{H}$, $a\|f\|_R \leq \|f\|_{\mathcal{H}} \leq b\|f\|_R$. Let $B(\cdot,\cdot)$ be the quadratic functional in $\mathcal{H} \times \mathcal{H}$ such that $\|f\|_R^2 = B(f,f)$, then it is possible to define the inner product $\langle f, g\rangle_R$ for all f and g in $\mathcal{H}$ as

$$\langle f, g\rangle_R = B(f,g) = \frac{1}{4}\left[\|f+g\|_R^2 - \|f-g\|_R^2\right]. \tag{21}$$

Note that $(\mathcal{H}, \|\cdot\|_R)$ defines a Reproducing Kernel Hilbert space. Let $R(\cdot,\cdot)$ be its reproducing kernel function and R the associated operator. Defining the bounded linear operator $R^{1/2}CR^{1/2}$, with eigenvalues $\nu_1 \geq \nu_2 \geq \cdots$ and eigenfunctions $\{\zeta_1, \zeta_2, \cdots\}$, and writting $\varphi_k = \nu_k^{-1/2} R^{1/2}\zeta_k$ for $k = 1.2.\cdots$, it is possible to derive the following results

$$\langle \varphi_j, \varphi_k\rangle_R = (\nu_j\nu_k)^{-1/2}\langle \zeta_j, \zeta_k\rangle_{\mathcal{L}_2} = \nu_k^{-1}\delta_{jk}, \tag{22}$$

and

$$\langle \varphi_j, C\varphi_k\rangle_{\mathcal{L}_2} = (\nu_j\nu_k)^{-1/2}\langle \zeta_j, R^{1/2}CR^{1/2}\zeta_k\rangle_{\mathcal{L}_2} = \delta_{jk}. \tag{23}$$

The set of functions $\{\varphi_1, \varphi_2, \cdots\}$ will be the $\mathcal{H}$ basis in which the simultaneous diagonalizatoin can be constructed. From Theorem 3 in Yuan and Cai (2010) it follows that for any $f \in \mathcal{H}$, $f = \sum_{k\geq 1} f_k\varphi_k$, where $f_k = \nu_k\langle f, \varphi_k\rangle_R$. Writtng $\gamma_k = \left(\nu_k^{-1} - 1\right)^{-1}$, it is easy to show that

$$\langle f, f\rangle_R = \sum_{k\geq 1}\left(1 + \gamma_k^{-1}\right) f_k^2, \tag{24}$$

where $\langle f, Cf\rangle_{\mathcal{L}_2} = \sum_{k\geq 1} f_k^2$ and $J(f) = \sum_{k\geq 1}\gamma_k^{-1} f_k^2$. From definition 21 it follows also that for $f, g \in \mathcal{H}$

$$\langle f, g\rangle_R = \sum_{k\geq 1}\left(1 + \gamma_k^{-1}\right) f_k g_k. \tag{25}$$

Given the characteristics of the operator C we can define a class of monotone Hilbert spaces that will be necessary to the study the convergence in a suitable set of intermediate norms. In such a way, for $0 \leq a$ and $f \in \mathcal{H}$ we define the squared norm

$$\|f\|_a^2 = \sum_{k\geq 1}\left(1 + \gamma_k^{-a}\right) f_k^2. \tag{26}$$

We call $\mathcal{H}_a$ the Banach space (with respect to the norm $\|\cdot\|_a$) generated after the completion of the set $\{f \in \mathcal{H} : \|f\|_a < \infty\}$. For $0 \leq a \leq 1$, there exists a direct correspondence between definition 26 and the K-method of interpolation for Hilbert spaces (see Tartar (2000), Triebel (1978) and Cox (1988) for more details), and it could be easily derived that $\mathcal{H}_a$ is a Hilbert space with inner product

$$\langle f, g\rangle_a = \sum_{k\geq 1}\left(1 + \gamma_k^{-a}\right) f_k g_k.$$

Note that $\mathcal{H}_c \subset \mathcal{H}_d$ whenever $d \leq c$ and the inclusion $\mathcal{H}_c \hookrightarrow \mathcal{H}_d$ is continuous. Without loss of generality it can be assumed that $\|f\|_c \leq \|f\|_d$. Furthermore, making $a = 1$ we get $\|f\|_1 = \|f\|_R$, and for $a = 0$, $\|f\|_0^2 = 2\langle f, Cf\rangle_{\mathcal{L}_2} = 2\mathbb{E}\left(\int_{\mathcal{T}} \tilde{X} f\right)^2$, where $\tilde{X}$ is a sample independent copy of the random process X. It is clear that for $0 \leq \alpha \leq 1$, $\mathcal{H} \subset \mathcal{H}_\alpha$ (element wise), and therefore, $\beta_0, \hat{\beta}_{n\lambda} \in \mathcal{H}_\alpha$. For technical reasons that will become evident during the asymptotic study of the convergence rates, we will conveniently develop the variational analysis on the space $\mathcal{H}_\alpha$ for some generic $0 \leq \alpha \leq 1$. We need to assure before that D^2J, which is originally defined in $\mathcal{H}$, can be properly extended to $\mathcal{H}_\alpha$. To do so, note that there exists a non-negative self-adjoint operator $W \in \mathbf{B}(\mathcal{H}, \mathcal{H})$ such that for any $\beta, f, g \in \mathcal{H}$, $D^2J(\beta)fg = \langle f, Wg\rangle_{\mathcal{H}}$. If we select two functions such that $\beta_1 \in \mathcal{H}$ and $\beta_2 \in \mathcal{H}_{2-\alpha} \subset \mathcal{H}$, by Lemma 2.1 in Cox and O'Sullivan (1990) it follows that $\langle \beta_2, W\beta_1\rangle_{\mathcal{H}} \leq \|\beta_2\|_{2-\alpha}\|\beta_1\|_\alpha$, and therefore, $W \in \mathbf{B}\left(\mathcal{H}_\alpha, \mathcal{H}_\alpha\right)$ for $0 \leq \alpha \leq 1$.

Recall that our objective is to find a basis in $\mathcal{H}_\alpha$ such that $D^2\ell_\infty$ and D^2J can be both diagonalized. The difficulty however, is that the operator $D^2\ell_\infty(\beta^*)$ depends on some $\beta^* \in \mathcal{H}_\alpha$ and this basis changes subsequently. In the special case in which $f_\lambda(y)$ is the density function of a Normal random variable, $D^2\ell_\infty(\beta^*)$ is independent of β^* and $D^2\ell_\infty(\beta^*)fg = \langle f, Cg\rangle_{\mathcal{L}_2}$. However this is not true in general for different distributions in the exponential family. The strategy we shall follow to circunvent this problem is to create eigen structures that behave asymptoticaly like $\{(\varphi_k, \gamma_k) : k \geq 1\}$ and allow to diagonalize the operator $G_{\infty\lambda}(\beta^*)$.

Let β^* be any fixed function in $\mathcal{H}_\alpha$. From the definition of $\ell_\infty(\beta^*$, for any $f, g \in \mathcal{H}_\alpha$, we have

$$\begin{aligned} D^2\ell_\infty(\beta^*)fg &= \mathbb{E}_X\left[\omega^{(2)}\left(\int_\tau X\beta^*\right)\int_\tau Xf\int_\tau Xg\right] \\ &:= \mathbb{E}_Y\left[\int_\tau Yf\int_\tau Yg\right], \end{aligned}$$

where we define $Y(t) := X(t)\left[\omega^{(2)}\left(\int_\tau X\beta^*\right)\right]^{\frac{1}{2}}$ for $t \in \tau$. By Cauchy-Schwarz inequality and assumptions 1 (ii) and (iii), it is easy to check that $Y(t)$ is a second order process and therefore it is possible to define some covariance function $U(\beta^*)(t_1, t_2)$ for $t_1, t_2 \in \tau$, and a respective integral operator on $\mathcal{L}_2(\tau)$ such that

$$\langle f, U(\beta^*)g\rangle_{\mathcal{L}_2} := D^2\ell_\infty(\beta^*)fg. \tag{27}$$

Note that the above definition is independent of the particular $\mathcal{H}_\alpha$to which β^* belongs. Thus, for $\beta^* \in \mathcal{H}_\alpha$ and $f \in \mathcal{H}$, with a little abuse in the notation it is possible to define a seminorm in $\mathcal{H}$ as

$$\|f\|^2_{R^*} = \langle f, U(\beta^*)f\rangle_{\mathcal{L}_2} + J(f). \tag{28}$$

The following theorem shows that no matter the specific β^* it is possible to construct a common basis for diagonalizing $U(\beta^*)$ and W simultaneously using the results obtained in the analysis with the covariance operator instead.

Theorem 1 *If assumptions 1 hold, for any $\beta^* \in \mathcal{H}_\alpha$ such that $\|\beta^*\|_\alpha^2 \leq M$ for some $M < \infty$, then*

(i) $\|\cdot\|_{R^*}$ and $\|\cdot\|_{\mathcal{H}}$ are equivalent norms in $\mathcal{H}$.

(ii) There exists an eigenstructure $\{(\varphi_k^*, \gamma_k^*) : k \geq 1\}$, with $\gamma_k^* = \left(\nu_k^{*-1} - 1\right)^{-1}$ and $\nu_1^* \geq \nu_2^* \geq \cdots$, such that

$$\langle \varphi_j^*, \varphi_k^* \rangle_{R^*} = \nu_k^{*-1} \delta_{jk} \tag{29}$$

$$\langle \varphi_j^*, U(\beta^*) \varphi_k^* \rangle_{\mathcal{L}_2} = \delta_{jk}. \tag{30}$$

(iii) For any $f \in \mathcal{H}$

$$f = \sum_{k \geq 1} f_k^* \varphi_k^*, \tag{31}$$

where $f_k^* = \nu_k^* \langle f, \varphi_k^* \rangle_{R^*}$.

(iv) There exist constans $0 < m \leq M < \infty$ such that for k large enough,

$$m\gamma_k \leq \gamma_k^* \leq M\gamma_k. \square \tag{32}$$

Recall that the basis representation in $(\mathcal{H}, \|\cdot\|_{R^*})$ is associated with a particular function $\beta^* \in \mathcal{H}_\alpha$. The following corollary extends the results of theorem 1 in order to define a common system for diagonalizing $G_{\infty\lambda}(\beta^*)$.

Corollary 2 *Let β^* be a function in a bounded set in $\mathcal{H}_\alpha$. Then, for any $f \in \mathcal{H}$, $\langle f, f \rangle_{R^*} = \sum_{k \geq 1} (1 + \gamma_k^{*-1}) f^{*2}_k$. Furthermore,*

$$\langle U(\beta^*) f, f \rangle_{\mathcal{L}_2} = \sum_{k \geq 1} f^{*2}_k, \tag{33}$$

and

$$J(f) = \sum_{k \geq 1} \gamma^{*-1} f^{*2}_k. \square \tag{34}$$

We define also, for $0 \leq a \leq 1$, a continuous set of norms in $\mathcal{H}$ that are going to be useful in further analysis of the convergence rates. Similar to 26, for any $f \in \mathcal{H}$,

$$\|f\|_{a^*}^2 = (1 + \gamma_k^{*-a}) f^{*2}_k. \tag{35}$$

Following the K-method of interpolation, and noting that the inclusion $\mathcal{H} \to \mathcal{L}_2(\tau)$ is compact, we define the Hilbert space $\mathcal{H}_{a^*}$ as the completion of $\{f \in \mathcal{H} : \|f\|_{a^*} < \infty\}$, and it follows that $\mathcal{H}_{a^*} = \mathcal{H}_a$ as Banach spaces, with equivalent norms.

As a result of corollary 2 it is possible to writte

$$\begin{aligned}\langle G_{\infty\lambda}(\beta^*)f, f\rangle_{R^*} &= D^2\ell_\infty(\beta^*)ff + \lambda J(f) \\ &= \sum_{k\geq 1}\left(1+\lambda\gamma^{*-1}\right) f^{*2}_k. \end{aligned} \tag{36}$$

The following proposition shows that the inverse of the operator $G_{\infty\lambda}(\beta^*)$ is well defined in $\mathcal{H}_\alpha$ for the cases that will be used in the asymptotic analysis. In particular, it will show that the linearizations defined in 13 and 14 are correctly defined on $\mathcal{H}_\alpha$.

Proposition 3 *Let H be a bounded set in $\mathcal{H}_\alpha$. If $0 \leq \alpha \leq 1$ and $\beta^* \in H$, then the operator $G^{-1}_{\infty\lambda}(\beta^*)\theta$ is bounded and well defined in $\mathcal{H}$ for $\theta = D\ell(\xi_1)$, $D^2\ell(\xi_1)\xi_2$, or $D^3\ell(\xi_1)\xi_2\xi_3$, with $\ell = \ell_{n\lambda}$ or $\ell_{\infty\lambda}$ and $\xi_1, \xi_2, \xi_3 \in \mathcal{H}_\alpha$.*

An elementary consequence of proposition 3 and theorem 1 is that for θ (as defined in the former proposition) there exists a constant $M > 0$ such that

$$\begin{aligned}\|G_{\infty\lambda}(\beta^*)^{-1}\theta\|^2_\alpha &\leq M\|G_{\infty\lambda}(\beta^*)^{-1}\theta\|^2_{\alpha^*} \\ &= M\sum_{k\geq 1}\left(1+\gamma_k^{*-\alpha}\right)\left(1+\lambda\gamma^{*-1}\right)^{-2}\theta^{*2}_k. \end{aligned} \tag{37}$$

Although the eigenstructure $\{(\varphi_k, \gamma_k) : k \geq 1\}$ is definitive for the operational analysis of the estimator, it does not have an obvious relation with the eigenstructures of the kernel and the covariance operators, K and C respectively. The rate of decay for the eigenvaues $\{\gamma_k\}$ play a fundamental role in the asymptoric analysis of $\hat{\beta}_{n\lambda}$, however, there is not a general explicit form to determine each γ_k in terms of $\{\rho_k\}$ and $\{\mu_k\}$.
Some assumptions will have to be made in order to characterize the sequence $\{\gamma_k\}$. One option is to assume that the set of functions $\{\psi_k\}$ and $\{\phi_k\}$ form the same basis in $\mathcal{L}_2$, that is, $\psi_k = \phi_k$ for $k = 1, \cdots$. In this case, $\gamma_k = \rho_k\mu_k$ (see proposition 4 in Yuan and Cai (2010)). This relation is the usual implicit condition that is made in the FPCA based appraches for solving functional regression models (see, e.g., Cai and Hall (2006)) or more recently FLR for general exponential families (see Dou, Pollard and Zhou (2010)).
When $\mathcal{H}$ is a Sobolev space of order m in $\tau = [0, 1]$, and C satisfies the Sacks-Ylvisaker conditions of order $r \geq 0$ it is possible to show that $\gamma_k \asymp \rho_k\mu_k$. Hereafter the notation $a_k \asymp b_k$ represents that for two sequences $\{a_k\}$ and $\{b_k\}$, $\frac{a_k}{b_k}$ is bounded away from 0 and ∞ as $k \to \infty$. More details can be found in Sacks and Ylvisaker (1970), Yuan and Cai (2010), and Ritter, Wasilkowski and Woźniakowski (1995). For the purpose of our analysis, we make some more general assumtions (Assumption 2) for which the two aforementioned cases are special instances.

Assumption 2 *The following assumptions are related to the eigenvalues and eigenfunctions behavior of the kernel and covariance operators.*

(i) There existis a constant $r > \frac{1}{2}$ such that the eigenvalues for the reproducing kernel K satify $\rho_k \asymp k^{-2r}$.

(ii) There existis a constant $s > \frac{1}{2}$ such that the eigenvalues for the covariance operator C satify $\mu_k \asymp k^{-2s}$.

(iii) $\gamma_k \asymp \rho_k \mu_k$, for γ_k as described in 24.

Assumption 2[(i)] determines the smootheness properties of the functions in $\mathcal{H}$ with basis $\{\psi_k : k \geq 1\}$ in terms of the parameter r, and Assumption 2[(ii)] describes the smootheness of the random process $\{X(t) : t \in \tau\}$ in terms of s. We call $\mathcal{F}(s.M)$ the class of probability distributions in Y and X such that assumptions 1 and assumptions 2[(ii)-(iii)] are satisfied.

1.2.4 Bias and Variance Approximation

Our main objective is to analyse the asymptotic behavior of the risk fucntion defined as $\mathbb{E}d\left(\hat{\beta}_{n\lambda}, \beta_0\right)$, where $d(\cdot,\cdot)$ is a well defined distance in $\mathcal{H}_\alpha$. It follows from the bias and variance sepparation in 12 that

$$\mathbb{E}d\left(\hat{\beta}_{n\lambda}, \beta_0\right) \leq \mathbb{E}d\left(\hat{\beta}_{n\lambda}, \beta_{\infty\lambda}\right) + \mathbb{E}d\left(\beta_{\infty\lambda}, \beta_0\right).$$

In order to facilitate the analysis and produce a more comprehensive solution, we take the two quantities at the right of the last expression sepparately. The linearizations defined in 13 and 14 have the purpose of approximating the bias and the variance by replacing $\hat{\beta}_{n\lambda} \approx \bar{\beta}_{n\lambda}$ in the variance and $\beta_{\infty\lambda} \approx \bar{\beta}_{\infty\lambda}$ in the bias respectively. The two following theorems validate these approximations when using $d\left(\beta_1, \beta_2\right) = \|\beta_1 - \beta_2\|_a^2$ for any value of a in $[0, \alpha]$, where $\alpha \leq 1$.

Theorem 4 *If $\|\bar{\beta}_{\infty\lambda} - \beta_0\|_\alpha^2 \to 0$ as $\lambda \to 0$, then, for $0 \leq \alpha \leq \frac{1}{2}\left[1 - \frac{1}{2(r+s)}\right]$ there exists some λ_0 such that for $\lambda \in [0, \lambda_0]$*

$$\sup_{\beta_0 \in \mathcal{H}} \|\beta_{\infty\lambda} - \beta_0\|_a^2 \leq 4\|\bar{\beta}_{\infty\lambda} - \beta_0\|_a^2. \tag{38}$$

Theorem 5 *If there exists a sequence λ_n such that $n^{-1}\lambda_n^{-\alpha-\frac{1}{2(r+s)}} \to 0$ as $n \to \infty$ with $\frac{1}{2(r+s)} < \alpha \leq 1$, and $\|\bar{\beta}_{n\lambda_n} - \beta_{\infty\lambda_n}\|_a^2 = o_p(1)$ for $0 \leq a \leq \alpha$, then*

$$\sup_{F \in \mathcal{F}(s,M)} \|\hat{\beta}_{n\lambda} - \beta_{\infty\lambda_n}\|_a^2 = O_p\left(\|\bar{\beta}_{n\lambda_n} - \beta_{\infty\lambda_n}\|_a^2\right). \tag{39}$$

Theorems 4 and 5 allow us to find the convergence rate for $\mathbb{E}\|\hat{\beta}_{n\lambda}, \beta_0\|_a^2$ using the linearized versions of the squared bias and the varaince respectively.

1.3 Convergence Rates

We show in this section the main results for the asymptotic poperties of the regularized estimator $\hat{\beta}_{n\lambda}$. Recall that for these results are given for a class of norms $\|\cdot\|_a$ in $\mathcal{H}_\alpha$ for $0 \leq a \leq \alpha \leq 1$. The following theorem presents the optimal rate of convergence for an appropriate choosen tuning parameter λ_n.

Theorem 6 *If* $(r+s) > \frac{3}{2}$ *and*

$$\lambda_n \asymp n^{\frac{-2(r+s)}{2(r+s)+1}},$$

then, for each $\epsilon > 0$ *there exists a finite constant* C_ϵ *such that*

$$\limsup_{n\to\infty} \sup_{F\in\mathcal{F}(s,M),\beta_0\in\mathcal{H}} P\left(\|\hat{\beta}_{n\lambda} - \beta_0\|_a^2 > C_\epsilon n^{-\frac{2(1-a)(r+s)}{2(r+s)+1}}\right) < \epsilon \tag{40}$$

for any $0 \leq a \leq \alpha$ *and* $\alpha = \frac{1}{2}\left[1 - \frac{1}{2(r+s)}\right]$.

The next theorem shows that the rate of convergence of the regularization estimator $\hat{\beta}_{n\lambda}$ presented in theorem 6 is optimal (in minimax sense) among all the possible estimators obtained from the data. Let $\mathcal{B}$ be the set of all measurable functions from the observations $(X_1, Y_1), \cdots, (X_n, Y_n)$.

Theorem 7 *For* a *and* α *as defined in theorem 6, there exists a constant* $d > 0$ *such that*

$$\liminf_{n\to\infty} \inf_{\beta\in\mathcal{B}} \sup_{F\in\mathcal{F}(s,M),\beta_0\in\mathcal{H}} P\left(\|\beta - \beta_0\|_a^2 > dn^{-\frac{2(1-a)(r+s)}{2(r+s)+1}}\right) > 0. \tag{41}$$

It follows then from theorems 6 and 7 that if we choose $\lambda_n \asymp n^{\frac{-2(r+s)}{2(r+s)+1}}$, the proposed estimator $\hat{\beta}_{n\lambda}$ is rate optimal. The proof of theorem 7 is a trivial adaptation of the proof of theorem 7 in Yuan and Cai (2010), where it was showed that in the special case of $f_y(\theta)$ being a normal density with constant variance and with eigenfunctions $\psi_k = \phi_k$ for all $k \geq 1$, $\inf_{\beta\in\mathcal{B}} \|\beta - \beta_0\|_a^2 \geq c \cdot n^{-\frac{2(1-a)(r+s)}{2(r+s)+1}}$.

Having an estimation for the function parameter β_0 it is possible to define a point estimator for the canonical parameter in the exponential density as $\int_\mathcal{T} \tilde{X}\hat{\beta}_{n\lambda}$, for a new observation of the predictor $\tilde{X}$ independent of the sample. We can use the fact that $\frac{1}{2}\|\beta\|_0^2 = \mathbb{E}_{\tilde{X}}\left(\int_\mathcal{T} \tilde{X}\beta\right)^2 = \langle \beta, C\beta\rangle_{\mathcal{L}_2}$ to define the mean squared error of this estimator as $\|\hat{\beta}_{n\lambda} - \beta_0\|_0$. The following corollary of theorems 6 and 7 establish this result.

Corollary 8 *Let $\tilde{X}$ be a copy of the stochastic process X independent of the sample. Then for all $F \in \mathcal{F}(s, M)$ and $\beta_0 \in \mathcal{H}$, and with λ_n as defined in theorem 6*

$$\sup_{F \in \mathcal{F}(s,M)} \mathbb{E}_{\tilde{X}} \mathbb{E}_{(X_1,Y_1),\cdots,(X_n,Y_n)} \left[\int_{\tau} \tilde{X} \left(\hat{\beta}_{n\lambda} - \beta_0 \right) \right]^2 \asymp n^{-\frac{2(r+s)}{2(r+s)+1}}$$

for n large enough.

1.4 Numerical Studies

In order to show the computational merits of the estimator 2 and to present some supporting evidence for the derived theoretical properties, we performed a series of numerical simulations with finite sample size scenarios. The main advantage of using the RKHS approach in the present context is that the resulting estimator is a finite dimensional (of order n) well defined convex minimization problem, as $\hat{\beta}_{n\lambda}$ can be expressed as a linear combination of basis functions related with the kernel K. Standard convex programming algorithms can be used to solve $\hat{\beta}_{n\lambda}$. A very conveninet numerical approach to solve the problem is the use of Iterative Rewieghted Least Squares method. This procedure corresponds to the scoring variation of the Newton-Raphson descending method to find the root of the first derivative of $\ell_{n\lambda}$ with respect to $\mathbf{c}$ and $\mathbf{d}$ in 10. In the context of General Linear Models with exponential families, the latter method has been extensively implemented and a rather simple modification of it, taking into account the penalty terms in the weight matrix and the score function in each iteration, can be utilized to find $\hat{\beta}_{n\lambda}$.

To make our simulation results more comparable we use similar settings to the ones used in Yuan and Cai (2010) and Hall and Horowitz (2007) with pertinent modifications to adapt them into the context of general linear model and a reproducing kernel norm type of penalty. We chose the particular case when the response is binary and follows a bernoulli distribution ($y \in \{0, 1\}$), that is, the Logistic Functional Regression model. We assume $\tau = [0, 1]$ and set the second order process X to be generated as

$$X = \sum_{k=1}^{80} \zeta_k Z_k \phi_k,$$

where $\phi_1(t) = 1$ and $\phi_{k+1}(t) = \sqrt{2}\cos(k\pi t)$. Z_k are independent copies of a uniform random variable on $\left[-\sqrt{3}, \sqrt{3}\right]$ and ζ_k is a deterministic sequence. Note that the spectral decomposition of the covariance operator in 18 implies that $\mu_k = \zeta_k^2$ and $C\phi_k = \mu_k \phi_k$. To show the benefits of the method over the procedures based on Functional Principal Components Analysis, and following Hall and Horowitz (2007) we choose two different scenarios for ζ_k, one of them well spaced and the other one with respective closely-spaced eigenvalues. For the well spaced eigenvalues we define $\zeta_k = (-1)^{k+1}k^{-\nu/2}$ and for the second case

$$\zeta_k = \begin{cases} 1 & k = 1; \\ 0.2(-1)^{k+1}(1 - 0.0001k) & 2 \leq k \leq 4; \\ 0.2(-1)^{k+1}\left[(5\lfloor k/5 \rfloor)^{-\nu/2} - 0.0001(k \mod 5)\right] & k \geq 5. \end{cases}$$

In each case we consider $\nu = 1.1$ and $\nu = 2$. For ilustrative purposes we define K in the spectral domain 17 with $\psi_1(t) = 1$ and $\psi_{k+1}(t) = \sqrt{2}\cos(2k\pi t)$. Also, we select $\rho_k = k^{-2r}$ for $k \leq 40$ and $\rho_k = 0$ if $k > 40$, considering the cases $r = \frac{3}{4}$ and $r = 1$. Note that in $\mathcal{H}$ a simultaneous diagonalization for C and K may be done using the basis $\{\psi_k\}$ and with γ_k in assumption 2 (iii) being $\rho_k \mu_{2k} \asymp \rho_k \mu_k$. The true parameter function β_0 is defined as

$$\beta_0 = \sum_{k=1}^{40} 4(-1)^{k+1}k^{-2}\psi_k.$$

To explore the effect of the sample size and ilustrate the theoretical convergence rates we consider $n = 50, 100, 200$ and 500. At each simulation scenario we run the experiment 1000 times and average on two measures of statistical performance for the estimator $\hat{\beta}_{n\lambda}$: the integrated mean squared error $\|\beta_0 - \hat{\beta}_{n\lambda}\|^2_{\mathcal{L}_2}$ and the prediction error (point estimator) $\|\beta_0 - \hat{\beta}_{n\lambda}\|^2_0$. In order to select the value of the parameter λ we used a data driven approach. For $n = 50$ and 100, leave-one-out crossvalidation was used, and for $n = 200$ and 500 we used ten fold crossvalidation.

For the estimation error, note that the eigenfunctions ψ_k are used to construct the spectral decomposition of the operator R, that is, $\varphi_k = \psi_k$ in 22. Therefore, using a slightly modified proof ot Proposition 4 in Yuan and Cai (2010), it would be easy to show that in this particular setting, $\|\beta_0 - \hat{\beta}_{n\lambda}\|_{\mathcal{L}_2}$ would be an equivalent norm to $\|\beta_0 - \hat{\beta}_{n\lambda}\|_{\frac{\nu}{r+\nu}}$, as defined 26. Consequently, by theorem 6 the theoretical convergence rate for the estimation error $\|\beta_0 - \hat{\beta}_{n\lambda}\|_{\mathcal{L}_2}$ is $n^{\frac{-2r}{2(r+\nu)+1}}$. Similarly, by corollary 8, the rate of decay for $\|\beta_0 - \hat{\beta}_{n\lambda}\|_0$ is $n^{\frac{-2(r+\nu)}{2(r+\nu)+1}}$.

In figure 1, the estimation errors $\|\beta_0 - \hat{\beta}_{n\lambda}\|^2_{\mathcal{L}_2}$ for the well spaced sequence of eigenvalues ζ_k are presented. The left and right sides correspond to the case when the smootheness degree of the Kernel operator K is set up with $r = \frac{3}{4}$ and $r = 1$ respectively. The axes are reshaped to the logarithmic scale and therefore, a descending linear pattern is expected in each case. As expected from the theoretical results, increasing the value of ν makes the estimation error bigger for a fixed n and flatten the line (the slope increases). Note that our results are asymptotically relevant and for $n = 50$ the differences are not as clear as for larger sample size. Figure 2 presents in a similar layout the results for the prediction

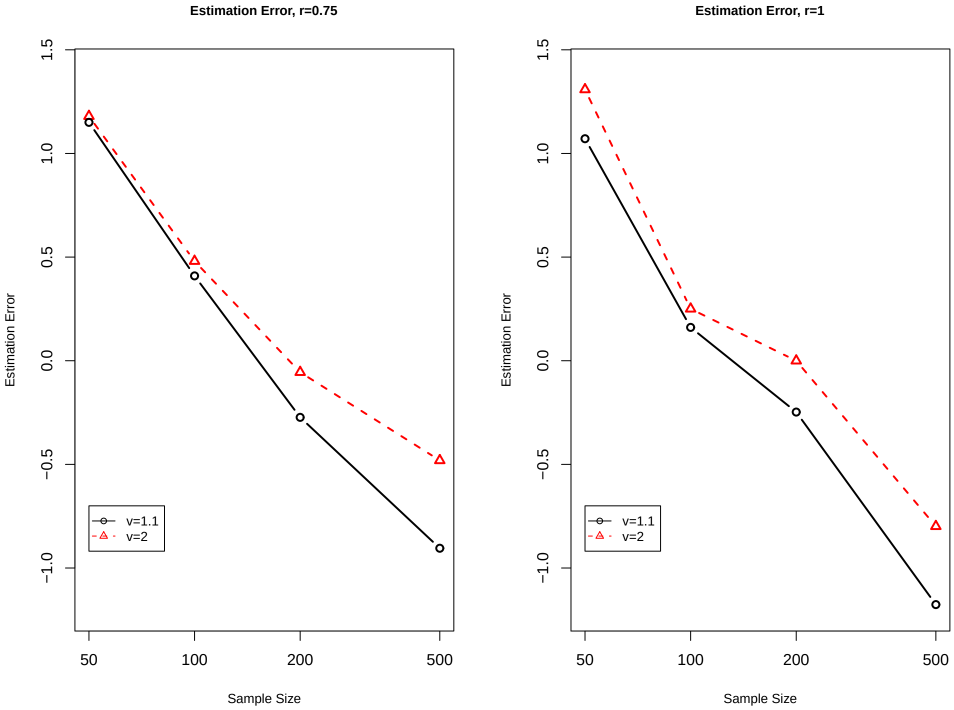

Figure 1: Estimation errors in logarithmic scale.

error $\|\beta_0 - \hat{\beta}_{n\lambda}\|_0^2$. Note that in this case increasing ν should result in a bigger prediction error and a more pronounced slope. The differences between the two panel show that for a bigger values of r, the convergence rate is less sensible to the variation in ν.

In figure 3 similar results are presented for estimation and prediction errors when the eigenvalues of the covariance operator are closely spaced. In both plots (left and right) $r = 3/4$. Note that the results in this case are similar to the well spaced scenarions and the same interpretations apply.

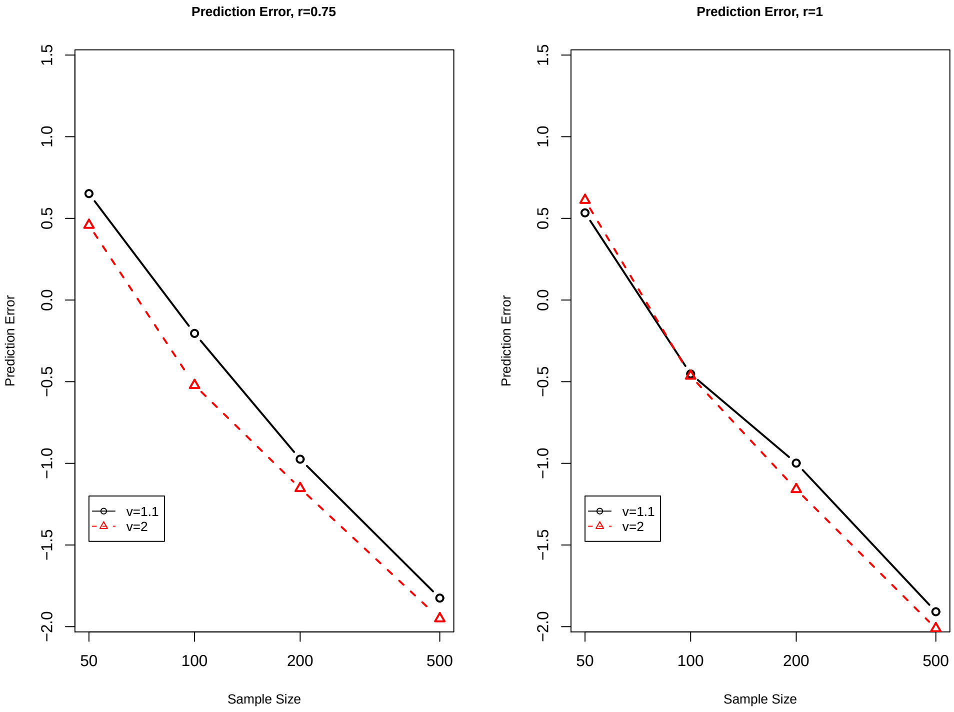

Figure 2: Prediction errors in logarithmic scale.

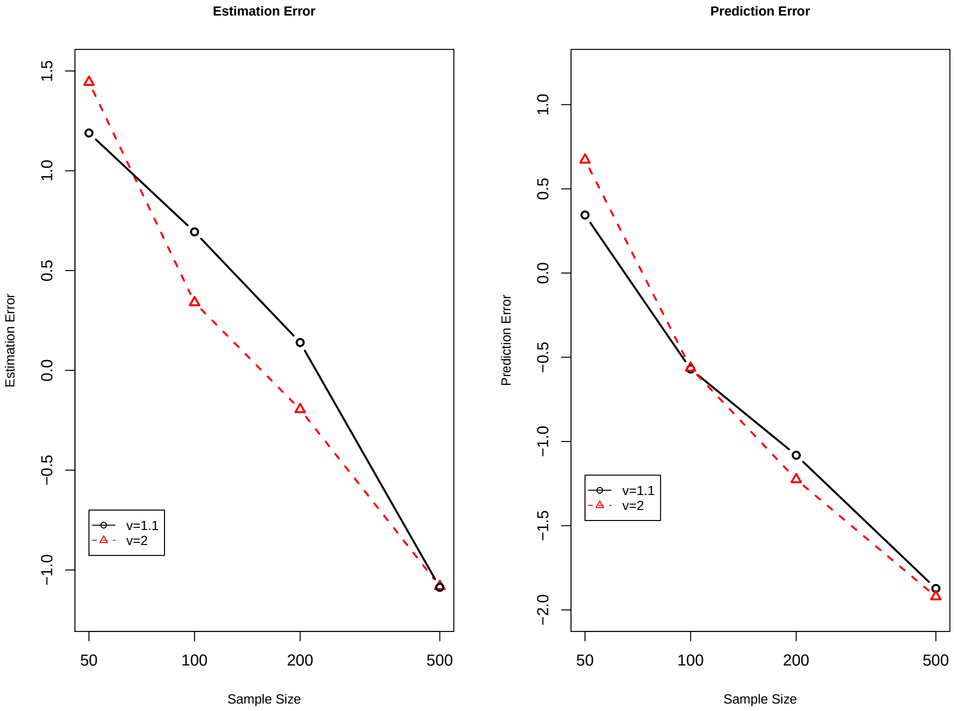

Figure 3: Estimation and Prediction errors in logarithmic scale when the eigenvalues in the covariance operator are closely spaced.

CHAPTER II

RADIAL BASIS REGULARIZATION FOR LINEAR INVERSE PROBLEMS WITH RANDOM NOISE

2.1 Introduction

In this chapter, the statistical properties of method of regularization with radial basis functions in the context of linear inverse problems are studied. Radial basis function regularization is widely used in machine learning because of its demonstrated eectiveness in numerous applications and computational advantages. From a statistical viewpoint, one of the main advantages of radial basis function regularization in general and Gaussian radial basis function regularization in particular is their ability to adapt to varying degree of smoothness in a direct problem. We show here that similar approaches for inverse problems not only share such adaptivity to the smoothness of the signal but also can accommodate dierent degrees of ill-posedness. These results render further theoretical support to the superior performance observed empirically for radial basis function regularization.

Radial basis function regularization is one of the most popular tools in statistical learning (see, e.g., Girosi, Jones, and Poggio (1993); Smola, Schölkopf, and Müller (1998); Wahba (1999); Evgeniou, Pontil, and Poggio (2000); Lin and Brown (2004); Lin and Yuan (2006)). Let $\Phi(x) = \phi(\|x\|)$ for vector $x \in \mathbb{R}^d$ be a radial basis function where $\phi : [0, +\infty) \to \mathbb{R}$ is a univariate function. Typical examples include $\phi(r) = r^{2m}\log(r)$ (thin plate spline), $\phi(r) = e^{-\varrho r^2/2}$ (Gaussian), and $\phi(r) = (c^2 + r^2)^{1/2}$ (multiquadrics) among others. When $K_\Phi(x, y) = \Phi(x - y)$ is (conditionally) positive definite in that for any $n \in \mathbb{Z}$ and any distinct $x_1, ..., x_n \in \mathbb{R}^d$,

$$\sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k K(x_j, x_k) > 0,$$

Φ can be identified with a reproducing kernel Hilbert space (Aronszajn (1950)), denoted by $\mathcal{H}_\Phi$. The squared norm in $\mathcal{H}_\Phi$ can be written as

$$J(f) = (2\pi)^{-d/2} \int_{R^d} |\tilde{f}(\omega)|^2 / \tilde{\Phi}(\omega) d\omega$$

for any function $f \in H_\Phi$, where $\tilde{f}$ stands for the Fourier transform of f, that is,

$$\tilde{f}(\omega) = (2\pi)^{-d/2} \int_{R^d} f(x) e^{-ix^T\omega} dx.$$

The method of regularization with a radial basis function estimates a functional parameter by the solution to

$$\min_{f\in H_\Phi}\{L(f,\text{data})+\lambda J(f)\},$$

where L is the empirical loss, often taken to be the negative log-likelihood. The tuning parameter $\lambda > 0$ controls the trade-off between minimizing the empirical loss and obtaining a smooth solution.

Consider in particular estimating a periodic function $f_0 : [-\pi, \pi] \to \mathbb{R}$ based on noisy observations of Af where A is a bounded linear operator, i.e.,

$$dY(t) = (Af_0)(t)dt + \epsilon dW(t), \qquad t \in [-\pi, \pi]. \tag{42}$$

Here $\epsilon > 0$ is the noise level and $W(t)$ is a standard Brownian motion on $[-\pi, \pi]$. The white noise model (42) connects to a number of common statistical problems in the light of results on its equivalence to nonparametric regression (Brown and Low, 1996), density estimation (Nussbaum, 1996), spectral density estimation (Golubev and Nussbaum, 1998), and nonparametric generalized regression (Grama and Nussbaum, 1997). The radial basis function regularization in this case gives the following estimate of f_0:

$$\hat{f}_\lambda = \underset{f\in\mathcal{H}_\Phi}{\arg\min}\left\{\|Y - Af\|^2_{\mathcal{L}_2} + \lambda J(f)\right\}.$$

Lin and Brown (2004) and Lin and Yuan (2006) recently studied the statistical properties of $\hat{f}_\lambda$ in a special case when A is the identity operator. They found that when f is a member of any finite-order Sobolev spaces, the method of regularization with many radial basis functions is rate optimal when the tuning parameter is appropriately chosen, which partially explains the success of such methods in this particular setting. Of course in many applications, A is not an identity operator but rather a general compact operator. Problems of this type can be found in almost all areas of science and engineering (see, e.g., Chalmond (2008); Kaipo and Somersalo (2004); Ramm (2009)). These problems, commonly referred to as inverse problems, are often ill-posed and therefore, fundamentally more difficult than the case when A is the identity, often referred to as direct problems (see, e.g., Cavalier (2008)). In this paper, we study the statistical properties of radial basis function regularization estimator $\hat{f}_\lambda$ in this setting.

Similar to direct problems, the difficulty in estimating f_0 in an inverse problem is determined by the complexity of the functional class it belongs to. Differing from direct problems, in an inverse problem, the difficulty of estimating f_0 also depends on the degree of ill-posedness of the linear operator A. We consider a variety of combinations of functional classes and linear operators and show that for many common choices of radial basis functions, $\hat{f}_\lambda$ is rate optimal whenever λ is appropriately tuned. Our results suggest that the superior statistical properties established earlier for the direct problems continue to hold in the inverse problems and therefore further make clear why the radial basis function regularization is so effective in a wider range of applications.

The rest of this chapter is organized as follows. In the next section, we describe in more detail the parameter spaces and the ill-posedness of the problem. We study in Section 2.3 the statistical properties of radial basis function regularization. Section 2.4 reports results from numerical experiments to illustrate the implications of our theoretical development. We close with some discussions in Section 2.5. The proof of theorem 9 is presented in section 2.6. The proofs of theorems 10 and 11 are similar to the first one and are ommited for clarity in the presentation.

2.2 Radial Basis Function Regularization in Linear Inverse Problems

The white noise model (42) can be expressed in terms of the corresponding Fourier coefficients and leads to a sequence model that is often more amenable to statistical analysis (see, e.g., Johnstone, 1998).

2.2.1 Sequence model via singular value decomposition

Let A^* be the adjoint operator of A. Because of the compactness of A, A^*A admits spectral decomposition

$$A^*Af = \sum_{k=1}^{\infty} b_k^2 \langle f, \varphi_k \rangle_{\mathcal{L}_2} \varphi_k \tag{43}$$

for any square integrable periodic function f, where the eigenfunctions $\{\varphi_1, \varphi_2, \dots\}$ constitute an orthornormal basis of $\mathcal{L}_2$, the collection of square integrable periodic functions, and the eigenvalues $\{b_1^2, b_2^2, \dots\}$ are arranged in a non-increasing order without loss of generality. Denote by ψ_k the normalized image of φ_k, that is, $A\varphi_k = b_k\psi_k$. It is easy to show that

$$A^*\psi_k = b_k\varphi_k.$$

From the singular value decomposition, we can convert the linear inverse problem (42) into a sequence model. More specifically, observe that

$$y_k := \langle Y, \psi_k \rangle_{\mathcal{L}_2} = \langle Af_0, \psi_k \rangle_{\mathcal{L}_2} + \langle \epsilon W, \psi_k \rangle_{\mathcal{L}_2} = b_k \langle f_0, \psi_k \rangle_{\mathcal{L}_2} + \epsilon \langle W, \psi_k \rangle_{\mathcal{L}_2} =: b_k\theta_k + \epsilon\xi_k$$

for $k = 1, 2, \dots$.

Unlike the direct problem where all singular values are one, in an inverse problem, $b_k \to 0$ as $k \to \infty$. The vanishing singular values poses challenges in inverting the linear operator A and makes the problem ill-posed. As a result, the estimation of f_0 becomes fundamentally more difficult for an inverse problem than for a direct problem. The rate of decay of $\{b_k : k \geq 1\}$ quantifies the ill-posedness. Typically, an inverse problem is called mildly ill-posed if $b_k \sim k^{-\beta}$ and severe ill-posed if $b_k \sim \exp(-\beta k)$ for some parameter $\beta > 0$ often referred to as the degree of ill-posedness. Hereafter, $a_k \sim b_k$ means that both a_k/b_k and b_k/a_k are bounded away from zero.

2.2.2 Parameter spaces

In addition to the ill-posedness, the difficulty of estimating f_0 in (42) is also determined by the parameter space for the functional parameter. It is often convenient to describe the parameters space using the Fourier coefficient with respect to the basis $\{\psi_k : k \geq 1\}$. Typically, f_0 belongs to the functional class corresponding to an ellipsoid Θ in the space of Fourier coefficients $\{\theta_k : k \geq 1\}$:

$$\Theta = \left\{ (\theta_k : k \geq 1) : \sum_{k \geq 1} a_k^2 \theta_k^2 \leq Q \right\}, \tag{44}$$

for a non-deceasing sequence $0 \leq a_1 \leq a_2 \leq \ldots$ such that $a_k \to \infty$ as $k \to \infty$, and a positive constant Q.

It is instructive to consider the case when $\{\psi_k : k \geq 1\}$ is the usual trigonometric basis, that is, $\psi_1(t) = (2\pi)^{-1/2}$, $\psi_{2l}(t) = \pi^{-1/2}\sin(lt)$ and $\psi_{2l+1}(t) = \pi^{-1/2}\cos(lt)$ for $l \geq 1$. In this case, the usual Sobolev spaces are perhaps the most popular examples of Θ. Let $\mathcal{S}^m(Q)$ be the mth order Sobolev space of periodic functions on $[-\pi, \pi]$, that is,

$$\mathcal{S}^m(Q) = \left\{ f \in \mathcal{L}_2 : f \text{ is } 2\pi\text{−periodic, and } \int_{-\pi}^{\pi} f^2 + (f^{(m)})^2 \leq Q \right\}.$$

Simple calculation shows that $\mathcal{S}^m(Q)$ can also be equivalently expressed as

$$\mathcal{S}^m(Q) = \left\{ f \in \mathcal{L}_2 : f = \sum_{k \geq 1} \theta_k \psi_k, \sum_{k \geq 1} a_k^2 \theta_k^2 \leq Q, a_1 = 1, a_{2l} = a_{2l+1} = k^{2m} + 1 \right\}.$$

In the same spirit, analytic functions or sometimes referred to as infinit-order Sobolev space can be described as

$$\mathcal{S}^\infty(\alpha; Q) = \left\{ f \in \mathcal{L}_2 : f = \sum_{k \geq 1} \theta_k \psi_k, \sum_{k \geq 1} a_k^2 \theta_k^2 \leq Q, a_1 = 1, a_{2l} = a_{2l+1} = e^{\alpha l} \right\}.$$

See Johnstone (1998) for details.

Appealing to this connection, in what follows, we shall write

$$\Theta^\alpha(Q) = \left\{(\theta_k : k \geq 1) : \sum_{k\geq 1} a_k^2\theta_k^2 \leq Q, a_1 = 1, a_{2l} = a_{2l+1} = k^\alpha + 1\right\}$$

as Sobolev type of spaces of order α; and

$$\Theta^\infty(\alpha; Q) = \left\{(\theta_k : k \geq 1) : \sum_{k\geq 1} a_k^2\theta_k^2 \leq Q, a_1 = 1, a_{2l} = a_{2l+1} = e^{\alpha k}\right\}$$

to represent spaces similar to $\mathcal{S}^\infty$.

2.2.3 Radial basis function regularization

We now describe the radial basis functions and the reproducing kernel Hilbert spaces they induce. Because we focus here on periodic functions, it is natural to consider periodized radial basis functions

$$\Phi_0(r) = \sum_{k\in\mathbb{Z}} \Phi(r - 2\pi k),$$

where Φ is a radial basis function. See Smola, Schölkopf and Müller (1998), Lin and Brown (2004) among others for further discussion of periodized radial basis functions and their applications in machine learning. As shown in Lin and Yuan (2006), Φ_0 (or equivalently K_{Φ_0}) is positive definite so long as Φ is positive definite and furthermore the norm of $\mathcal{H}_{\Phi_0}$ can be given by

$$\|f\|^2_{\mathcal{H}_{\Phi_0}} = \sum_{k\geq 1} \gamma_k\theta_k^2,$$

where θ_ks are the Fourier coefficients of f, and $\gamma_1 = (2\pi)^{-1/2}\{\tilde{\Phi}(0)\}^{-1}$, $\gamma_{2l} = \gamma_{2l+1} = (2\pi)^{-1/2}\{\tilde{\Phi}(l)\}^{-1}$, $l = 1, 2, \ldots$. When $\{\psi_k : k \geq 1\}$ is taken to be the classical trigonometric basis, the method of regularization with radial basis function Φ_0 can be equivalently expressed in terms of the sequence of Fourier coefficients:

$$\hat{f}_\lambda = \underset{f=\sum_{k\geq 0}\theta_k\psi_k\in\mathcal{H}_{\Phi_0}}{\arg\min} \left\{\sum_{k\geq 1}(y_k - b_k\theta_k)^2 + \lambda\sum_{k\geq 1}\gamma_k\theta_k^2\right\}.$$

Consider, for example, the periodic Gaussian kernel

$$G_0(r) = \sum_{k\in\mathbb{Z}} G(r - 2\pi k),$$

where

$$G(r) = \frac{1}{\sqrt{2\pi\varrho^2}}\exp\left(-\frac{r^2}{2\varrho^2}\right)$$

for some parameter $\varrho > 0$. Simple calculation yields that $\gamma_{2l} = \gamma_{2l+1} = e^{l^2\varrho^2/2}$. Other popular examples include periodic multiquadratics and Wendland kernels (Wendland (1998)) that corresponds to $\gamma_{2l} = \gamma_{2l+1} = e^{l\varrho}$ and $\gamma_{2l} = \gamma_{2l+1} = k^{\varrho}$ respectively. There are also other common choices of radial basis functions for which γ_k behaves similarly to these three examples. See Buhlmann (2003) for further details.

2.3 Main Results

Following the discussion before, we shall focus on the following sequence model hereafter:

$$y_k = b_k\theta_k + \epsilon\xi_k, \qquad k = 1, 2, \ldots. \tag{45}$$

The inverse problem under investigation is either mildly or severely ill-posed, that is, $b_k \sim k^{-\beta}$ or $b_k \sim e^{-\beta k}$ respectively. We shall also consider Sobolev type of parameter spaces, that is, $(\theta_k : k \geq 1) \in \Theta^\alpha$ for some $\alpha > 1/2$ or $\Theta^\infty(\alpha, Q)$. Our primary interest is to evaluate the statistical performance of radial basis function regularization:

$$(\hat{\theta}_{k\lambda} : k \geq 1) = \underset{(\eta_k : k \geq 1)}{\arg\min} \left\{ \sum_{k\geq 1} (y_k - b_k\eta_k)^2 + \lambda \sum_{k\geq 1} \gamma_k \eta_k^2 \right\}. \tag{46}$$

In particular, we consider three different types of radial basis functions: (1) $\gamma_k \sim e^{\gamma k^2}$ for some $\gamma > 0$ with periodic Gaussian kernel as a typical example; (2) $\gamma_k \sim e^{\gamma k}$ with periodic multiquadrics kernel as a typical example; and (3) $\gamma_k \sim k^\gamma$ with periodic Wendland kernel or the usual spline kernels (see, e.g., Wahba (1990)) as typical examples.

We begin with Gaussian type of kernel, that is, $\gamma_k = e^{\gamma k^2}$ for some $\gamma > 0$.

Theorem 9 *Assume that $\gamma_k \sim e^{\gamma k^2}$ for some $\gamma > 0$.*

(a) (Mildly ill-posed with Sobolev spaces) If $b_k \sim k^{-\beta}$ and

$$\lambda \sim \exp\left(-\epsilon^{\frac{4}{2\alpha+2\beta+1}}\right),$$

then

$$\sup_{(\theta_k : k\geq 1) \in \Theta^\alpha(Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^{\frac{4\alpha}{2\alpha+2\beta+1}}.$$

(b) (Mildly ill-posed with analytic functions) If $b_k \sim k^{-\beta}$ and

$$\lambda \sim \exp\left(-\frac{\gamma}{4\alpha^2}\left(\log\frac{1}{\epsilon^2}\right)^2\right),$$

then

$$\sup_{(\theta_k : k\geq 1) \in \Theta^\infty(\alpha, Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^2\left(\log\frac{1}{\epsilon^2}\right)^{2\beta+1}.$$

(c) *(Severely ill-posed with Sobolev spaces) If* $b_k \sim k^{-\beta}$ *and*

$$\lambda \sim \exp\left(-\left(\log\frac{1}{\epsilon^2}\right)^2\right),$$

then

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\alpha(Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \left(\log\frac{1}{\epsilon}\right)^{-2\alpha}.$$

(d) *(Severely ill-posed with analytic functions) If* $b_k \sim e^{-\beta k}$ *and*

$$\lambda \sim \exp\left(-\frac{\gamma}{(2\alpha+2\beta)^2}\left(\log\frac{1}{\epsilon^2}\right)^2\right),$$

then

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\infty(\alpha,Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^{\frac{2\alpha}{\alpha+\beta}}.$$

We note that all the rates obtained in Theorem 9 are minimax optimal (see, e.g., Cavalier (2008)). In other words, from Theorem 9, when the tuning parameter λ is appropriately chosen, Gaussian radial basis function regularization is rate optimal for all combinations of ill-posedness as well as parameter spaces. This result, together with similar results for direct problems (Lin and Brown (2004)), partly explain its success in numerous applications.

Next we consider the case with a multiquadrics type of kernel.

Theorem 10 *Assume that* $\gamma_k \sim e^{\gamma k}$ *for some* $\gamma > 0$.

(a) *(Mildly ill-posed with Sobolev spaces) If* $b_k \sim k^{-\beta}$ *and*

$$\lambda \sim \exp\left(-\epsilon^{-\frac{2}{2\alpha+2\beta+1}}\right),$$

then

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\alpha(Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^{\frac{4\alpha}{2\alpha+2\beta+1}}.$$

(b) *(Mildly ill-posed with analytic functions) If* $b_k \sim k^{-\beta}$, *then*

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\infty(\alpha,Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^2\left(\log\frac{1}{\epsilon^2}\right)^{2\beta+1}.$$

provided that

$$\lambda \sim \begin{cases} \epsilon^{\frac{\gamma}{\alpha}} & \gamma > \alpha - 2\beta \\ \epsilon & \gamma \leq \alpha - 2\beta \end{cases}.$$

(c) *(Severely ill-posed with Sobolev spaces) If* $b_k \sim k^{-\beta}$ *and* $\lambda \sim \epsilon^2$, *then*

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\alpha(Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \left(\log\frac{1}{\epsilon}\right)^{-2\alpha}.$$

(d) *(Severely ill-posed with analytic functions) Suppose that* $b_k \sim e^{-\beta k}$. *If* $\gamma > \alpha - 2\beta$ *and*

$$\lambda \sim \epsilon^{-\frac{\beta+\gamma}{\alpha+\beta}},$$

then

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\infty(\alpha,Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^{\frac{2\alpha}{\alpha+\beta}}.$$

If $\gamma \leq \alpha - 2\beta$, *then the best achievable rate is*

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\infty(\alpha,Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^{\frac{4\beta+2\gamma}{3\beta+\gamma}},$$

and it is attained when

$$\lambda \sim \epsilon^{\frac{2\beta+\gamma}{3\beta+\gamma}}.$$

From Theorem 10, regularization with multiquadrics type of kernel is also rate optimal for finite-order Sobolev spaces. For analytic functions, however, its behavior is more complex. When the inverse problem is mildly ill-posed, it can still achieve the optimal rate but different tuning parameters are needed to attain the optimal rate depending on whether γ is larger than $\alpha - 2\beta$. However, for severely ill-posed problems, the minimax optimal rate can only be achieved when $\gamma > \alpha - 2\beta$. The transition point $\alpha - 2\beta$ is somewhat surprising. Observe that $\mathcal{H}_{\Phi_0} \subseteq \mathcal{S}^\infty(\alpha, Q)$ if $\gamma \geq \alpha$ and $\mathcal{S}^\infty(\alpha, Q) \subset \mathcal{H}_{\Phi_0}$ otherwise. Thus Theorem 10 essentially states that regularization with multiquadrics type of kernel is always rate optimal if the reproducing kernel Hilbert space induced by the radial basis function is smaller than the parameter space. But even when the parameter space is larger than the induced space, that is, $\gamma < \alpha$, it is still capable of achieving the minimax optimal rate so long as $\gamma > \alpha - 2\beta$.

Now consider the Wendland/spline type of kernel.

Theorem 11 *Assume that* $\gamma_k \sim k^\gamma$ *for some* $\gamma > 1/2$.

(a) *(Mildly ill-posed with Sobolev spaces) Suppose that* $b_k \sim k^{-\beta}$. *If* $\gamma > \alpha - 2\beta$ *and*

$$\lambda \sim \epsilon^{\frac{4\alpha}{2\alpha+2\beta+1}},$$

then

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\alpha(Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^{\frac{4\alpha}{2\alpha+2\beta+1}}.$$

If $\gamma \le \alpha - 2\beta$, the best achivable rate is

$$\sup_{(\theta_k : k \ge 1) \in \Theta^\alpha(Q)} \sum_{k \ge 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^{\frac{2(4\beta+2\gamma)}{6\beta+2\gamma+1}},$$

and it is attained when

$$\lambda \sim \epsilon^{\frac{4\beta+2\gamma}{6\beta+2\gamma+1}}.$$

(b) (Mildly ill-posed with analytic functions) Suppose the $b_k \sim k^{-\beta}$. If $\gamma > \alpha - 2\beta$ and

$$\lambda \sim \epsilon^2 \left(\log \frac{1}{\epsilon}\right)^{-2\beta-\gamma},$$

then

$$\sup_{(\theta_k : k \ge 1) \in \Theta^\alpha(Q)} \sum_{k \ge 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^2 \left(\log \frac{1}{\epsilon}\right)^{2\beta+1}.$$

If $\gamma \le \alpha - 2\beta$, the best achievable rate is

$$\sup_{(\theta_k : k \ge 1) \in \Theta^\alpha(Q)} \sum_{k \ge 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^{\frac{2(4\beta+2\gamma)}{6\beta+2\gamma+1}},$$

and it is attained when

$$\lambda \sim \epsilon^{\frac{4\beta+2\gamma}{6\beta+2\gamma+1}}.$$

(c) (Severely ill-posed with Sobolev spaces) If $b_k \sim e^{-\beta k}$

$$\lambda \sim \epsilon^2$$

then

$$\sup_{(\theta_k : k \ge 1) \in \Theta^\alpha(Q)} \sum_{k \ge 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \left(\log \frac{1}{\epsilon}\right)^{-2\alpha}$$

(d) (Severely ill-posed with analytic functions) Suppose $b_k \sim e^{-\beta k}$. If $\gamma > \alpha - 2\beta$, then the achievable rate is

$$\sup_{(\theta_k : k \ge 1) \in \Theta^\alpha(Q)} \sum_{k \ge 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^{\frac{2\beta}{\alpha+2\beta}},$$

and it is attained when

$$\lambda \sim \epsilon^{\frac{4\beta}{\alpha+2\beta}}.$$

When $\gamma \le \alpha - 2\beta$, the best achievable rate is

$$\sup_{(\theta_k : k \ge 1) \in \Theta^\alpha(Q)} \sum_{k \ge 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k\right)^2 \sim \epsilon^{\frac{4}{3}},$$

and it is attained when $\lambda \sim \epsilon^{\frac{2}{3}}$.

Theorem 11 indicates that the method of regularization with Wendland or spline type of kernel is also capable of attaining the minimax optimal rate but only so if γ is sufficiently large, or equivalently, the reproducing kernel Hilbert space $\mathcal{H}_{\Phi_0}$ is sufficiently small.

Our main results are summarized in Table 1.

2.4 Risk Analysis of Radial Basis Function Regularization

In order to establish the results presented in the previous section we start by setting out a general framework. Recall that the regularization estimator $(\hat{\theta}_{k\lambda} : k \geq 1)$ is defined as

$$(\hat{\theta}_{k\lambda} : k \geq 1) = \underset{(\eta_k : k \geq 1)}{\arg\min} \left\{ \sum_{k\geq 1} (y_k - b_k \eta_k)^2 + \lambda \sum_{k \geq 1} \gamma_k \eta_k^2 \right\}.$$

It can be written explicitly as

$$\hat{\theta}_{k\lambda} = \frac{b_k}{b_k^2 + \lambda \gamma_k^{-1}} y_k \,, k = 1, 2, \cdots . \tag{47}$$

In particular, here we consider

$$b_k \sim \begin{cases} k^{-\beta} & \text{Mildly ill} - \text{posed} \\ \exp(-\beta k) & \text{Severely ill} - \text{posed} \end{cases} .$$

Furthermore, the true Fourier coefficients $(\theta_k : k \geq 1)$ are assumed to be in an ellipsiod

$$\Theta(Q) = \left\{ (\theta_k : k \geq 1) : \sum_{k \geq 1} a_k^2 \theta_k^2 \leq Q \right\},$$

where

$$a_k \sim \begin{cases} k^{\alpha} & \Theta = \Theta^{\alpha}(Q) \\ \exp(\alpha k) & \Theta = \Theta^{\infty}(\alpha; Q) \end{cases} .$$

Observe that the risk of the radial basis function regularization estimator $(\hat{\theta}_{k\lambda} : k \geq 1)$ can be decomposed as the sum of the squared bias and the variance:

$$\sum_{k \geq 1} \mathbb{E} \left(\hat{\theta}_{k\lambda} - \theta_k \right)^2 = \sum_{k \geq 1} \left(\mathbb{E} \hat{\theta}_{k\lambda} - \theta_k \right)^2 + \sum_{k\geq 1} \mathrm{Var}\left(\hat{\theta}_{k\lambda} \right) =: B_\theta^2 \left(\hat{\theta}_\lambda \right) + \mathrm{Var}_\theta \left(\hat{\theta}_\lambda \right). \tag{48}$$

By (47), we can further write

$$B_\theta^2 \left(\hat{\theta}_\lambda \right) = \sum_{k \geq 1} \frac{\lambda^2 \gamma_k^{-2} \theta_k^2}{\left(b_k^2 + \lambda \gamma_k^{-1} \right)^2}$$

and

$$\mathrm{Var}_\theta \left(\hat{\theta}_\lambda \right) = \epsilon^2 \sum_{k \geq 1} \frac{b_k^2}{\left(b_k^2 + \lambda \gamma_k^{-1} \right)^2}.$$

Table 1: We list here, for different combinations of parameter spaces and radial basis functions, the best achievable convergence rates of $\mathcal{R}(\lambda)$ and the order of the tuning parameter λ needed to attain the rate. α reflects the smoothness of the parameter space, β determines the ill-posedness of the inverse problem and γ depends on the choice of the radial basis function.

Kernel Type		Θ^α		Θ^∞	
		Mildly Ill-posed	Severely Ill-posed	Mildly Ill-posed	Severely Ill-posed
Gaussian	λ	$\exp\left(-\epsilon^{-\frac{4}{2\alpha+2\beta+1}}\right)$	$\exp\left(-\left(\log\frac{1}{\epsilon^2}\right)^2\right)$	$\exp\left(-\frac{\gamma}{4\alpha^2}\left(\log\frac{1}{\epsilon^2}\right)^2\right)$	$\exp\left(-\frac{\gamma}{(2\beta+2\alpha)^2}\left(\log\frac{1}{\epsilon^2}\right)^2\right)$
	$\mathcal{R}(\lambda)$	$\epsilon^{\frac{4\alpha}{2\alpha+2\beta+1}}$	$\left(\log\frac{1}{\epsilon}\right)^{-2\alpha}$	$\epsilon^2\left(\log\frac{1}{\epsilon^2}\right)^{2\beta+1}$	$\epsilon^{\frac{2\alpha}{\alpha+\beta}}$
Multiquadrics	λ	$\exp\left(-\epsilon^{-\frac{2}{2\alpha+1}}\right)$	ϵ^2	$\begin{cases}\epsilon^{\frac{\gamma}{\alpha}} & \text{if } \gamma > \alpha-2\beta\\ \epsilon & \text{if } \gamma \le \alpha-2\beta\end{cases}$	$\begin{cases}\epsilon^{-\frac{\beta+\gamma}{\alpha+\beta}} & \text{if } \gamma > \alpha-2\beta\\ \epsilon^{\frac{2\beta+\gamma}{3\beta+\gamma}} & \text{if } \gamma \le \alpha-2\beta\end{cases}$
	$\mathcal{R}(\lambda)$	$\epsilon^{\frac{4\alpha}{2\alpha+2\beta+1}}$	$\left(\log\frac{1}{\epsilon}\right)^{-2\alpha}$	$\epsilon^2\left(\log\frac{1}{\epsilon^2}\right)^{2\beta+1}$	$\begin{cases}\epsilon^{\frac{2\alpha}{\alpha+\beta}} & \text{if } \gamma > \alpha-2\beta\\ \epsilon^{\frac{4\beta+2\gamma}{3\beta+\gamma}} & \text{if } \gamma \le \alpha-2\beta\end{cases}$
Wendland	λ	$\begin{cases}\epsilon^{\frac{4\beta+2\gamma}{2\alpha+1\beta+1}} & \text{if } \gamma > \alpha-2\beta\\ \epsilon^{\frac{4\beta+2\gamma}{6\beta+2\gamma+1}} & \text{if } \gamma \le \alpha-2\beta\end{cases}$	ϵ^2	$\begin{cases}\left(\log\frac{1}{\epsilon}\right)^{-2\beta-\gamma} & \text{if } \gamma > \alpha-2\beta\\ \epsilon^{\frac{4\beta+2\gamma}{6\beta+2\gamma+1}} & \text{if } \gamma \le \alpha-2\beta\end{cases}$	$\begin{cases}\epsilon^{\frac{4\beta}{\alpha+2\beta}} & \text{if } \gamma > \alpha-2\beta\\ \epsilon^{\frac{2}{3}} & \text{if } \gamma \le \alpha-2\beta\end{cases}$
or Spline	$\mathcal{R}(\lambda)$	$\begin{cases}\epsilon^{\frac{4\alpha}{2\alpha+2\beta+1}} & \text{if } \gamma > \alpha-2\beta\\ \epsilon^{\frac{4\gamma+8\beta}{6\beta+2\gamma+1}} & \text{if } \gamma \le \alpha-2\beta\end{cases}$	$\left(\log\frac{1}{\epsilon}\right)^{-2\alpha}$	$\begin{cases}\epsilon^2\left(\log\frac{1}{\epsilon}\right)^{2\beta+1} & \text{if } \gamma > \alpha-2\beta\\ \epsilon^{\frac{8\beta+4\gamma}{6\beta+2\gamma+1}} & \text{if } \gamma \le \alpha-2\beta\end{cases}$	$\begin{cases}\epsilon^{\frac{2\alpha}{\alpha+2\beta}} & \text{if } \gamma > \alpha-2\beta\\ \epsilon^{\frac{4}{3}} & \text{if } \gamma \le \alpha-2\beta\end{cases}$

The squared bias and variance can be further bounded as follows:

$$B_\theta^2\left(\hat{\theta}_\lambda\right) \le \max_k \left\{ \frac{\lambda^2 \gamma_k^{-2} a_k^{-2}}{\left(b_k^2 + \lambda\gamma_k^{-1}\right)^2} \right\} \left(\sum_{k\ge 1} a_k^2 \theta_k^2 \right) \le \max_k \left(\frac{\lambda^2 \gamma_k^{-2} a_k^{-2}}{b_k^4 + \lambda^2 \gamma_k^{-2}} \right) \left(\sum_{k\ge 1} a_k^2 \theta_k^2 \right) , \quad (49)$$

and

$$\mathrm{Var}_\theta\left(\hat{\theta}_\lambda\right) \le \epsilon^2 \sum_{k\ge 1} \frac{b_k^2 \gamma_k^2}{b_k^4 \gamma_k^2 + \lambda^2}. \quad (50)$$

Some selected proofs will be presented in Section 2.6.

2.5 Numerical Experiments

To ilustrate the performance of the the estimator (47), we carried out some numerical experiments. The main purpose is to demonstrate the actual convergence rates when the noise level ϵ, as described in (42), goes to zero.

All the simulations are made in the domain of the coefficients for the trigonometric basis $\{\psi_k : k \ge 1\}$ of $L^2(-\pi, \pi)$, implying that all the parameters are generated as sequences in ℓ^2. We consider in particular, two functions $f = \sum_{k\ge 1} \theta_k \psi_k$ where

$$\theta_k \sim \begin{cases} k^{-2} & \text{for } (\theta_k : k \ge 1) \in \Theta^\alpha(Q) \ , \ \alpha < \frac{3}{2} \\ \exp(-2k) & \text{for } (\theta_k : k \ge 1) \in \Theta^\infty(\alpha; Q) \ , \ \alpha < 2 \end{cases} ,$$

representing Sobolev type or analytic type of functions repectively. We also consider two operators A corresponding to mildly or severe ill-posed situations respectively:

$$b_k \sim \begin{cases} k^{-2} & \text{for Mildly ill} - \text{posed} \\ \exp(-2k) & \text{for Severely ill} - \text{posed} \end{cases} .$$

For simplicity, we assure that $\gamma > \alpha - 2\beta$ in each possible sccenario choosing $\gamma = 2$ in all of the three types of kernel considered.

To understand the asymptotic behavior of the regularized estimator, we consider a set of values for the noise level as $\epsilon = \frac{j}{100}$ for $j = 1, 2, \cdots, 15$. In each case we estimate the parameter $(\theta_k : k \ge 1)$ using (47) and calculate the integrated squared error by $\|\hat{\theta}_\lambda - \theta\|_{\ell^2}$. We performed 100 replications for each setting to obtain a fair approximation of the risk.

As usual in nonparametric estimators, the tuning parameter should be selected in order to minimize the risk. To do so, in each setting we calculate $(\hat{\theta_k} : k \ge 1)$ for each $\lambda \in \Lambda$, where $\Lambda = \{\lambda_i : \lambda_i = \exp(-i/5) \, , i = 1, \cdots, 100\}$, and select the estimator $\hat{\theta}_{\lambda^*}$ such that

$$\lambda^* = \arg\min_{\lambda\in\Lambda} \|\hat{\theta}_\lambda - \theta\|_{\ell^2}.$$

The results are presented in Figure 4. In each plot, we include also the minimax optimal rate adjusted by a constant. As can be seen, the simulated rates have a similar decay as the theoretical minimax counterparts, indicating the estimate is rate optimal.

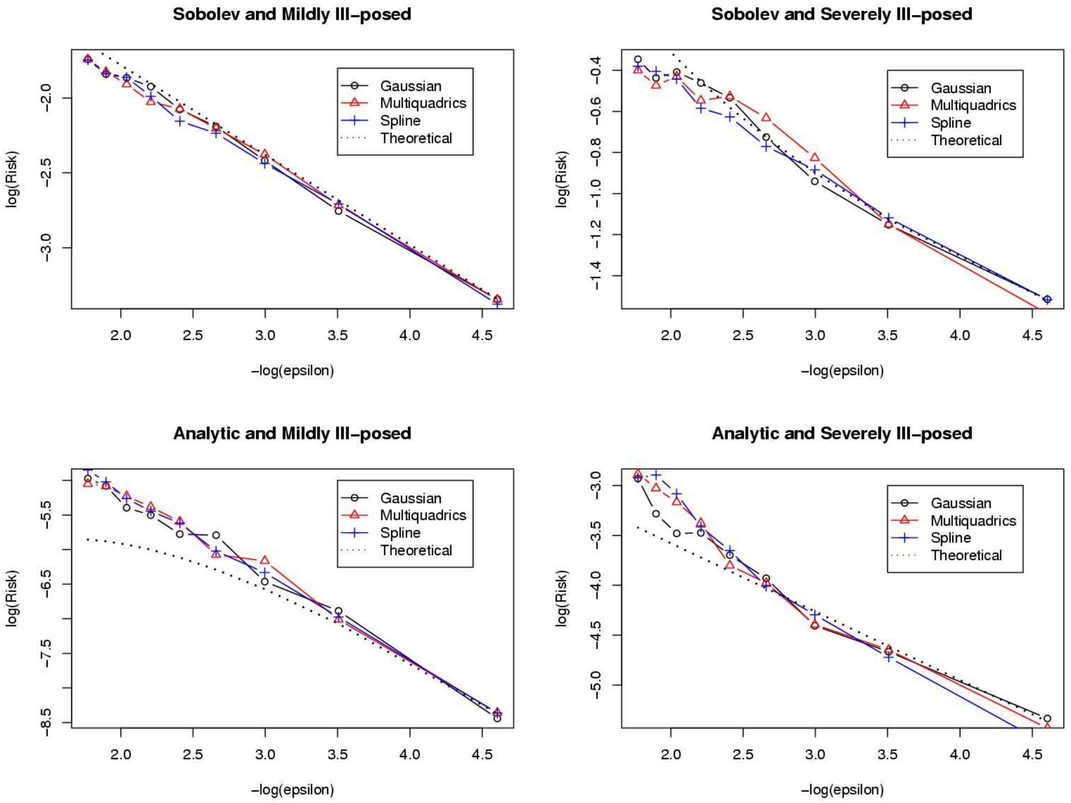

Figure 4: Comparison of the risk of radial basis function regularization. The results are averaged over 100 replications.

2.6 *Proof of Theorem 9*

We begin with the case when $\gamma_k \sim e^{\gamma k^2}$ for some $\gamma > 0$.

2.6.0.1 Mildly ill-posed with Sobolev spaces

In this case, $b_k \sim k^{-\beta}$ and $a_k \sim k^{\alpha}$. From (49)

$$\sup_{\theta\in\Theta^{\alpha}(Q)} B^2_{\theta}\left(\hat{\theta}_{\lambda}\right) \leq C\lambda^2 \left(\min_{x\geq 1}\left\{x^{2\alpha-4\beta}\exp(-2\gamma x^2)+\lambda^2 x^{2\alpha}\right\}\right)^{-1}. \tag{51}$$

Hereafter we use C as a generic positive constant which may take different values at each appearance. By the first order condition, the minimum on the right hand side is achieved at the root of

$$\left(\frac{2\alpha-4\beta}{x}-4\gamma x\right)x^{2\alpha-4\beta}\exp(-2\gamma x^2)+\frac{2\alpha\lambda^2}{x}x^{2\alpha}=0,$$

implying that

$$\sup_{\theta\in\Theta^\alpha(Q)} B_\theta^2\left(\hat{\theta}_\lambda\right)\leq C\left(-\log\lambda\right)^{-\alpha}. \tag{52}$$

Now consider $\mathrm{Var}_\theta\left(\hat{\theta}_\lambda\right)$. From (50)

$$\mathrm{Var}_\theta\left(\hat{\theta}_\lambda\right)\leq\epsilon^2\sum_{k\geq1}\frac{k^{-2\beta}\exp(-2\gamma k^2)}{k^{-4\beta}\exp(-2\gamma k^2)+\lambda^2}\approx\epsilon^2\int_1^\infty\frac{x^{-2\beta}\exp(-2\gamma x^2)}{x^{-4\beta}\exp(-2\gamma x^2)+\lambda^2}dx.$$

The integral on the rightmost hand side can be bounded by

$$\int_1^\infty\frac{1}{x^{-2\beta}+\lambda^2x^{2\beta}\exp(2\gamma x^2)}dx\leq\int_1^{x_0}x^{2\beta}dx+\int_{x_0}^\infty\lambda^{-2}x^{-2\beta}\exp(-2\gamma x^2)dx,$$

where x_0 is the positive root of

$$x^{-2\beta}=\lambda^2x^{2\beta}\exp(2\gamma x^2),$$

which is of the order $\left(-\gamma^{-1}\log\lambda\right)^{\frac{1}{2}}$. Because

$$\int_{x_0}^\infty\lambda^{-2}x^{-2\beta}\exp(-2\gamma x^2)dx=o\left(x_0^{2\beta}\right),$$

for small values of λ, we have

$$\sum_{k\geq1}\mathrm{Var}\left(\hat{\theta}_{k\lambda}\right)=O\left(\epsilon^2\left(\log\frac{1}{\lambda}\right)^{2\alpha+\frac{1}{2}}\right) \tag{53}$$

as $\lambda\to0$.

Combining (52) and (53), we have

$$\sum_{k\geq1}\mathbb{E}\left(\hat{\theta}_{k\lambda}-\theta_k\right)^2=O\left(\left(\log\frac{1}{\lambda}\right)^{-\alpha}+\epsilon^2\left(\log\frac{1}{\lambda}\right)^{\beta+1/2}\right)$$

as $\epsilon\to0$. Taking

$$\lambda=O\left(\exp\left(-\epsilon^{\frac{4}{2\alpha+2\beta+1}}\right)\right)$$

yields

$$\sup_{(\theta_k:k\geq1)\in\Theta^\alpha(Q)}\sum_{k\geq1}\mathbb{E}\left(\hat{\theta}_{k\lambda}-\theta_k\right)^2=O\left(\exp\left(-\epsilon^{\frac{4\alpha}{2\alpha+2\beta+1}}\right)\right),$$

as $\epsilon\to0$.

2.6.0.2 Mildly ill-posed with analytic function

For this case, $b_k \sim k^{-\beta}$ and $a_k \sim \exp(\alpha k)$. First observe that the variance $\mathrm{Var}_\theta\left(\hat{\theta}_\lambda\right)$ does not change with the parameter space and can still be bounded as in (53). On the other hand, from (49),

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\infty(\alpha,Q)} B_\theta^2\left(\hat{\theta}_\lambda\right) \leq C\lambda^2\left(\min_{x\geq 1}\left\{x^{-4\beta}\exp(2\alpha x-2\gamma x^2)+\lambda^2\exp(2\alpha x)\right\}\right)^{-1},$$

and following the first order condition for the minimization on the right hand side, we have

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\infty(\alpha,Q)} B_\theta^2\left(\hat{\theta}_\lambda\right) = O\left(\exp\left[-2\alpha\left(-\frac{1}{\gamma}\log\lambda\right)^{1/2}\right]\right), \tag{54}$$

as $\lambda \to 0$. Summing up, we have

$$\sum_{k\geq 1}\mathbb{E}\left(\hat{\theta}_{k\lambda}-\theta_k\right)^2 = O\left(\exp\left[-2\alpha\left(-\frac{1}{\gamma}\log\lambda\right)^{1/2}\right]+\epsilon^2\left(\log\frac{1}{\lambda}\right)^{\beta+1/2}\right), \tag{55}$$

as $\epsilon \to 0$. Consequently, if λ takes the optimal value

$$\lambda = O\left(\exp\left(-\frac{\gamma}{2\alpha^2}\left(\log\frac{1}{\epsilon^2}\right)^2\right)\right),$$

the risk is minimax rate optimal, i.e.,

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\infty(\alpha,Q)}\sum_{k\geq 1}\mathbb{E}\left(\hat{\theta}_{k\lambda}-\theta_k\right)^2 = O\left(\epsilon^2\left(\log\frac{1}{\epsilon}\right)^{2\beta+1}\right),$$

as $\epsilon \to 0$.

2.6.0.3 Severely ill-posed with Sobolev spaces

In this case, $b_k \sim \exp(-\beta k)$ and $a_k \sim k^\alpha$. Inequality (49) implies

$$\sup_{\theta\in\Theta^\alpha(Q)} B_\theta^2\left(\hat{\theta}_\lambda\right) \leq C\lambda^2\left(\min_{x\geq 1}\left\{x^{2\alpha}\exp(-4\beta x-2\gamma x^2)+\lambda^2 x^{2\alpha}\right\}\right)^{-1},$$

where, after minimizing the function inside the brackets, we get

$$\sup_{\theta\in\Theta^\alpha(Q)} B_\theta^2\left(\hat{\theta}_\lambda\right) = O\left((-\log\lambda)^{-\alpha}\right) \;\; as\ \lambda\to 0. \tag{56}$$

The variance $\mathrm{Var}_\theta\left(\hat{\theta}_\lambda\right)$ can be bounded using (50). In particular,

$$\begin{aligned}\mathrm{Var}_\theta\left(\hat{\theta}_\lambda\right) &\leq \epsilon^2\sum_{k\geq 1}\frac{\exp(-2\beta k-2\gamma k^2)}{\exp(-4\beta k-2\gamma k^2)+\lambda^2}\\ &\approx \epsilon^2\int_1^\infty\frac{\exp(-2\beta x-2\gamma x^2)dx}{\exp(-4\beta x-2\gamma x^2)+\lambda^2}\\ &\leq \epsilon^2\epsilon^2\left(\int_1^{x_0}\exp(2\beta x)dx+\int_{x_0}^\infty\lambda^{-2}\exp(-2\beta x-2\gamma x^2)dx\right),\end{aligned}$$

where x_0 is the positive root of

$$\exp(-2\beta x) = \lambda^2 \exp(2\beta x + 2\gamma x^2).$$

It can be easily derived that

$$x_0 = O\left(\left(-\gamma^{-1} \log \lambda \right)^{1/2} \right)$$

as $\lambda \to 0$. Observing that

$$\int_{x_0}^{\infty} \lambda^{-2} \exp(-2\beta x - 2\gamma x^2) dx = o\left(\exp(2\beta x_0) \right),$$

we have

$$\mathrm{Var}_\theta\left(\hat{\theta}_\lambda \right) = O\left(\epsilon^2 \exp\left(2\beta \left(-\gamma^{-1} \log \lambda \right)^{1/2} \right) \right) \tag{57}$$

as $\epsilon \to 0$. Combining (56) and (57), we have

$$\sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k \right)^2 = O\left((-\log \lambda)^{-\alpha} + \epsilon^2 \exp\left(2\beta \left(-\gamma^{-1} \log \lambda \right)^{1/2} \right) \right) \tag{58}$$

as $\epsilon \to 0$, attaining the minimax optimal rate of convergence

$$\sup_{(\theta_k : k\geq 1) \in \Theta^\alpha(Q)} \sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k \right)^2 = O\left(\left(\log \frac{1}{\epsilon} \right)^{-2\alpha} \right),$$

when

$$\lambda = O\left(\exp\left(-\left(\log \frac{1}{\epsilon^2} \right)^2 \right) \right)$$

as $\epsilon \to 0$.

2.6.0.4 Severely ill-posed with Analytic functions

For this case, $b_k \sim \exp(-\beta k)$ and $a_k \sim \exp(\alpha k)$. Following similar arguments as before, from Inequality (49)

$$\begin{aligned} \sup_{(\theta_k : k\geq 1) \in \Theta^\infty(\alpha, Q)} B_\theta^2\left(\hat{\theta}_\lambda \right) &\leq C\lambda^2 \left(\min_{x\geq 1} \left\{ \exp[(2\alpha - 4\beta)x - 2\gamma x^2] + \lambda^2 \exp(2\alpha x) \right\} \right)^{-1} \\ &= O\left(\exp\left[-2\alpha \left(-\gamma^{-1} \log \lambda \right)^{1/2} \right] \right) \end{aligned}$$

as λ goes to 0. On the other hand, the variance can still be bounded by (57). Hence,

$$\sum_{k\geq 1} \mathbb{E}\left(\hat{\theta}_{k\lambda} - \theta_k \right)^2 = O\left(\exp\left[-2\alpha \left(-\gamma^{-1} \log \lambda \right)^{1/2} \right] + \epsilon^2 \exp\left(2\beta \left(\frac{1}{\gamma} \log \frac{1}{\lambda} \right)^{1/2} \right) \right) \tag{59}$$

as $\epsilon \to 0$, which implies that if

$$\lambda = O\left(\exp\left(-\frac{\gamma}{(2\alpha+2\beta)^2}\left(\log\frac{1}{\epsilon^2}\right)^2\right)\right),$$

the radial basis function regularization achieves the optimal rate of convergence

$$\sup_{(\theta_k:k\geq 1)\in\Theta^\infty(\alpha,Q)}\sum_{k\geq 1}\mathbb{E}\left(\hat{\theta}_{k\lambda}-\theta_k\right)^2 = O\left(\epsilon^{\frac{4\alpha}{2\alpha+2\beta}}\right)$$

as $\epsilon \to 0$.

CHAPTER III

A DISCRIMINANT APPROACH TO TREATMENT SELECTION. INFINITE SAMPLE CONSISTENCY ANALYSIS

3.1 Introduction

One of the leading challenges in personalized medicine is to select treatments and clinical decisions according to the individual patient's requirements. The necessity for a individual treatment rule comes from the incresing number of studies that show heterogeneity on the patient's response, not just acrross different individuals but also in different times of the illness treatment. During the last years personalized medicine and the individualized treatment selection has become a priority in the health system (see e.g. Lesko (2007) and Piquette-Miller and Grant (2007)). In this chapter, we are interested in the problem of assigning the best treatment for a patient with particular prognosis covariates. To achieve that, we estimate the best decision rule from the data generated in a clinical trial. It is important to clarify that this approach differs from a different point of view in personalized medicine, in which the individuals are classified into risk groups, and each group has a matching treatment.

We assume that the data is generated from a clinical trial such that, for each patient in the trial, we observe a triplet: patient's individual prognostic covariates (pretreatment variables) $X \in \mathcal{X}$, treatment assignment $A \in \mathcal{A}$, and a clinical outcome called reward R that measures the benefit of the treatment with higher values being better. It is assumed that trial is randomized in the sense the for a particular $x \in \mathcal{X}$, there is a defined probability for each possible treatment with $p(a|x) := P(A = a|X = x) > 0$ for each $(a, x) \in \mathcal{A} \times \mathcal{X}$. We also assume the the reward function is bounded and without loss of generality for the purpose of our analysis R is a positive random variable with probability one. The patient selection rule follows a probability density $h_0(x)$ for $x \in \mathcal{X}$. We define P as the probability distribution of (X, A, R), with associated likelihood for an observation $h_0(x)p(a|x)h_1(r|x, a)$, where h_1 is the conditional density of R on X and A.

The main goal of treatment selection problems is to select the best treatment A^* after observing the corresponding covariates X^* so that the resulting reward is maximized. Therefore, assuming that the data generating mechanism is known, the natural course of action is dictated by the Bayes rule. To clearly express this, we define the Value function as

$$V(a; X^*) := \mathbb{E}\left(R|A = a, X = X^*\right), \qquad a \in \mathcal{A},\ X^* \in \mathcal{X},$$

that measures the expected reward when a particular treatment is assigned. Then, the optimal treatment is given by

$$A^* = \arg\max_{a \in \mathcal{A}} V(a; X^*), \tag{60}$$

and in particular, if $\mathcal{A} = \{-1, 1\}$ then

$$A^* = \begin{cases} \ \ 1 & \text{if } V(1; X^*) > V(-1; X = X^*) \\ -1 & \text{if } V(1; X = X^*) < V(-1; X = X^*) \end{cases}.$$

An *Individualized Treatment Rule* (ITR) is a decision (non random) function $d : \mathcal{X} \to \mathcal{A}$, that is, is a completely defined mechanism to assign a treatment $a \in \mathcal{A}$ to any possible patient with pretreatment variable $x \in \mathcal{X}$. If in particular we define $d_*(X^*) := A^*$, then d_* is a Bayes optimal decision rule.

In order to derive a data driven estimation procedure for d_*, it is necessary to define a performance measure for an arbitrary ITR d. For that purpose, we will use the expected value function. Let P^d be the probabillity distribution of (X, A, R) conditioned to $A = d(X)$. The assumptions on the distribution P make P^d absolutely continuous with respect to P (see Qian and Murphy (2011)) and therefore, the expected value function is

$$V(d) := \mathbb{E}V(d(X), X) = \int R\frac{dP^d}{dP}dP = \mathbb{E}\left[R\frac{\mathbb{I}(A = d(X))}{p(A|X)}\right], \tag{61}$$

where dP^d/dP corresponds to the RadonNikodym derivative and $\mathbb{I}(\cdot)$ is the indicator function. It is clear that

$$d_* = \arg\max_{d} V(d),$$

where the maximum is defined on the set of all possible desicion rules. We call $V(d_*)$ the Optimal Expected Value. Givent i.i.d observations (X_i, A_i, R_i) of n subjects in the randomized trial, it is possible to define the empirical version of V as

$$V_n(d) := \mathbb{E}_n\left[R\frac{\mathbb{I}(A = d(X))}{p(A|X)}\right] = \frac{1}{n}\sum_{i=1}^{n} R_i\frac{\mathbb{I}(A_i = d(X_i))}{p(A_i|X_i)}. \tag{62}$$

Trying to estimate d_* maximizing directly 62 is a computational NP-hard problem given the non-concavity of the objective function.

The existing methods to estimate the optimal ITR are based on three different paradigms:

- *Treatment Effect based.* The main idea is to estimate the value function $V(a, x)$ and replace in 60 to obtain d_*.

- *Treatment Difference based.* Under this paradigm, first is estimated the difference value function $H(x) = V(1, x) - V(-1, x)$ and replace $\hat{d}_* = \text{sign}(H(x))$.

- *Treatment Rule Classification based.* Under this paradigm a classification procedure is used to directly estimate the set $\mathcal{X}^+ : x \in \mathcal{X} : V(1, x) > V(-1, x)$.

Methods based on the third paradigm above usually rely on a direct maximization of 62, however, given the computational challenges, it is often assumed that d_* belongs to a very simple class of ITR functions $\mathcal{F}$ and that $\mathcal{X}$ is a low dimensional space. Some particular approaches are in Murphy et al. 2001 and Robins et al. 2008 in the seeting of Dynamic Treatment Regimes. For the first and second paradigms mentioned above, note that there is a (sometimes implicit) two stage estimation method. That is, d_* is estimated indirectly and the statistical criteria to estimate the value function (usually minimizing the prediction error) does not always match with the objective of maximizing the expected value function. Some examples of this apprach can be found in Qian and Murphy (2011) and Moodie et al. (2009). Particularly in Qian and Murphy (2011) there is an explicit description of the risk of choosing a suboptimal ITR with this two stage method when the estimated expected value function is overfitted.

We are interested in a different approach on which the empirical value function is directly maximized over a rich functional class $\mathcal{F}$ with a penalization term to regularize the overfitting. The main idea is to solve the problem usign a similar path that learning alrgorithms use to overcome the classification problem with convex minimization. Although the similarities are evident, the nature of the estimating the optimal ITR is different as long as the the interest is a classification based on the maximum of a response random variable (R). We investigate what some of the learning classification machinery have to offer to solve the ITR selection problem in a more computationally feasible way and to evaluate if that solution has good asymptotic statistical properties. Previous work on that direction is presented in Zhao, Zeng, Rush and Kosorok (2012) with emphasis on a support vector machine look alike method usign a hinge surrogate value function. We extend their methods to a much more general setting with generic type of surrogate value functions in a Reproducing Kernel Hilbert Space regularization framework.

Specifically, here we analyze the effect of using a surrogate value function in order to make the optimization problem tractable. Sufficient conditions for Infinite Sapmle Consistency

(Fisher consistency) are established, When the surrogated version of the problem is used and the maximization allows for Infinite Sample Consistency, convergence bounds on the surrogated expected value function can be translated to the 0-1 value function counterpart. That facilitates the asymptotic analysis for the convergence of the estimated ITR to d_*.

3.2 Theoretical Framework and Methodology

For clarity in the presentation we shall describe the notation an the problem set up assuming that there are two treatment options, that is, $\mathcal{A} = \{-1, 1\}$. The case of a multicategory set $\mathcal{A}$ will be described in section 3.3.2. As is typical in statistical learning classification problems, associated with a particular ITR d it is possible to define a measurable discriminant function f such that $d(x) = \text{sign}(f(x))$ for all $x \in \mathcal{X}$. Similarly, we call f_* the Bayes optimal rule, where $d_* = \text{sign}(f_*)$. Thus, the expected value function can be easily adapted as

$$V(f) := \mathbb{E}V(f(X), X) = \mathbb{E}\left[R\frac{\mathbb{I}(A = \text{sign}\,(f(X))}{p(A|X)}\right],$$

with the empirical version being $V_n(f) = \mathbb{E}_n V(f(X), X)$. It is clear that $V_n(f)$ is not concave in f and direct maximization is an NP-hard problem. The remedy that has been applied in statistical learning theory to circumvent this obstacle is to use a surrogate function ϕ that replace the step discountinuity caused by the expression $\mathbb{I}(A = \text{sign}\,(f(X))$. Replacing this part with a concave function ϕ evaluated in $Af(X)$ allows to estimate f_* using the tools of convex optimization.

For a specific surrogate function ϕ, we define the expected surrogate value function as

$$V_\phi(f) := \mathbb{E}\left[R\frac{\phi(Af(X)}{p(A|X)}\right], \tag{63}$$

and the ϕ-optimal value as $V_\phi^* = \sup_f V_\phi(f)$. Similarly, we define the empirical (surrogate) value function as

$$V_n^\phi(f) := \frac{1}{n}\sum_{i=1}^{n}\left(R_i\frac{\phi(A_i f(X_i)}{p(A_i|X_i)}\right), \tag{64}$$

for a decision function $f : \mathcal{X} \to \mathbb{R}$. We consider the estimator for the optimal ITR d_* as $\hat{d}_n^\phi = \text{sign}(\hat{f}_n^\phi)$ where

$$\hat{f}_n^\phi = \arg\max_{f \in \mathcal{F}} V_n^\phi(f), \tag{65}$$

and $\mathcal{F}$ is a closed convex functional class. We consider a rich enough class $\mathcal{F}$ to avoid selecting suboptimal treatments, but with a controlled size mechanism to avoid overfitting. Regularization is an appropriate approach to find a good estimator $\hat{f}_n^\phi$. If $\|\cdot\|_{\mathcal{H}}$ ia the norm of a Reproducing Kernel Hilbert Space $\mathcal{H}$ with associated kernel K it is possible to redefine the estimator as

$$\hat{f}_n^{\phi} = \arg\max_{f \in \mathcal{H}} \left(V_n^{\phi}(f) - \lambda \|f\|_{\mathcal{H}}^2 \right), \tag{66}$$

where λ is a tuning parameter that balances out the two terms in the equation. The Representer theorem assures that $\hat{f}_n^{\phi}$ can be solved as a convex optimization problem in finite dimensions, even that the space $\mathcal{H}$ can be infinite dimensional.

Using V_n^{ϕ} in 65 or 66 instead of V_n changes the rules of the estimation. The first important difference is that, as a function of the sample size, the sequence of estimators $\hat{f}_n^{\phi}$ can be "ϕ-consistent", that is $V_{\phi}(\hat{f}_n^{\phi}) \to V_{\phi}^*$, but that does not necessarily implies that $\hat{f}_n^{\phi}(x) \to f_*(x)$ for all $x \in \mathcal{X}$ up to null probability sets. That is, asymptoticaly, $\hat{f}_n^{\phi}$ may not lead to the optimal ITR d_*, and that depends enterely on the function ϕ that is used. When $\lim_{n\to\infty} \text{sign}(\hat{f}_n^{\phi}) = d_*$ almost sure in $\mathcal{X}$, we say that the respective surrogate value function ϕ is *Infinite Sample Consistent*. This property has received different names in the classification learning literature, as classification calibrated (e.g., Bartlett, Jordan and McAuliffe (2006)) or Fisher Consistent (e.g., Lin (2004)).

In machine learning literature, the surrogated version of the loss function has been seen as a practical solution from a computational point of view. However, in the statistical analysis of the estimator $\hat{f}_n^{\phi}$ we are interested in two measures of quality: the estimation error that concerns with the stochastic perturbation in finite samples and the approximation error that is related to the bias introduced in the class $\mathcal{H}$. As discussed in Zhang (2003) and Bartlett, Jordan and McAuliffe (2006) in the context of binary classification, for a particular f it is easier to find upper bounds in the surrogate value deficit $V_{\phi}^* - V_{\phi}(f)$ than in the 0-1 expected value deficit $V(f_*) - V(f)$. Given that we are interested in the asymptotic behavior of the last quantity, we can translate its value from the surrogate version. If ϕ is Infinite Samlple Consistent, then it is possible to find a function $\psi : [0,1] \to \mathbb{R}^+$ such that

$$\psi\left(V(f_*) - V(f)\right) \leq V_{\phi}^* - V_{\phi}(f), \tag{67}$$

where ψ is increasing on some set $[0,t)$ for $t > 0$. We want the bound produced by ψ^{-1} to be as tight as possible. Under low noise conditions (Tsybavov (2001)), this bounds can be improved. Note that the value deficit is relative to the infimum over all measurable functions f, but $\hat{f}_n^{\phi} \in \mathcal{H}$. The surrogate value deficit can be split into two components as

$$\psi\left(V(f_*) - V(f)\right) \leq \left(\sup_{h \in \mathcal{H}} V_{\phi}(h) - V_{\phi}(f) \right) + \left(V_{\phi}^* - \sup_{h \in \mathcal{H}} V_{\phi}(h) \right), \tag{68}$$

where the terms on the right of the inequality represent respectively the estimation and the approximation error measured in the surrogate value. For some universal kernels K

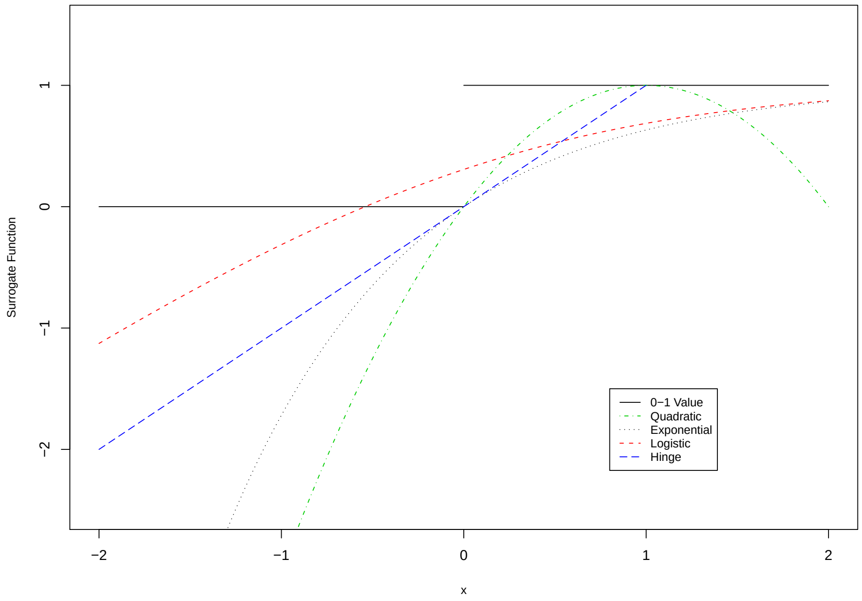

Figure 5: Comparison of different Surrogate Value fucntions.

associated with the Hilbert space $\mathcal{H}$ the approximation error vanishes and therefore it is possible to find convergent rates for $V(f_*) - V(f)$. The particular case of the Gaussian kernel function K was studied in Zhao, Zeng, Rush and Kosorok (2012) for the ITR problem.
It is therefore clear that in the statistical analysis for the estimation of the optimal ITR through the minimization of 66, Infinite Sample Consistency of ϕ plays a fundamental role. Some examples of the functions that could be used as surrogate ϕ are

- Hinge surrogate: $\phi(x) = 1 - \max(1 - x, 0)$.
- Least Squares surrogate: $\phi(x) = 1 - (1 - x)^2$.
- Exponential Surrogate: $\phi(x) = 1 - \exp(-x)$.
- Squared Hinge Surrogate: $\phi(x) = 1 - \max((1 - x)^2, 0)$.
- Logistic Surrogate: $\phi(x) = 1 - \log(1 + \exp(-x))$.

Figure 5 presents a visual comparison of them. Note that all of them are concanve on $\mathbb{R}$ and also differentiable, with exemptions in the hinge and squared hinge cases that are not differentiable at one.

3.3 Infinite Sample Consistency. Main Results.

To facilitate the analysis of the surrogate value function it is convenient to avoid dependency on X by defining the conditional ϕ-value function. We shall assume that $\mathcal{A} = \{-1, 1\}$ for the next definitions, but additional scenarions with multi-treatment selection and witholding option wiil be studied also. For all $x \in \mathcal{X}$ we say that

$$Q_x(f) := \mathbb{E}\left[R\frac{\phi(Af(X)}{p(A|X)}\middle|X = x\right] = \sum_{a \in \mathcal{A}} \phi(af(x))\mathbb{E}(R|A = a, X = x)$$
$$= \phi(f(x))\mathbb{E}(R|A = 1, X = x) + \phi(-f(x))\mathbb{E}(R|A = -1, X = x), \quad (69)$$

is the conditional ϕ-value function of $f(x)$. Note that $V_\phi(f) = \mathbb{E}_X Q_X(f)$ and that the dependency on ϕ has been left implicit. In general, it is possible to study the consistency problem pointwise, that is, the Infinite Sample Consistency can be defined and for each particular $x \in \mathcal{X}$. Thus, we drop the x dependency in definition 69. Writing $r_1 = \mathbb{E}(R|A = 1, X = x)$ and $r_{-1} = \mathbb{E}(R|A = -1, X = x)$, we express 69 as

$$Q_r(f) = \phi(f)r_1 + \phi(-r)r_{-1}. \quad (70)$$

Recall that $d_*(x) = \text{sign}(r_1 - r_{-1})$. Given that our intention is to find the conditions on which ϕ is Infinite Sample Consistent, for the following definition, we do not make any assumptions on ϕ and we can not assure that $\sup_{f\in\mathbb{R}} Q_r(f)$ is achieved or uniquely determined. To avoid the nuisance of dealing with unbounded cases for any values r_1 and r_{-1}, it is assumed hereafter that ϕ is bounded above, that is, $\sup_{y\in\mathbb{R}} \phi(y) < \infty$.

Definition 1 *ϕ is Infinite Sample Consistent if for any sequence $f^{(1)}, f^{(2)}, \dots \subset \mathbb{R}$ such that*

$$\lim_{i\to\infty} Q_r(f^{(i)}) = \sup_{g\in\mathbb{R}} Q_r(g),$$

then,

$$\lim_{i\to\infty} \text{sign}\left(f^{(i)}(r_1 - r_{-1})\right) = 1. \ \square$$

The assumptions on the probability density $h_1(r|x, a)$ imply that $0 < r_j < \infty$ with probability one for $j \in \mathcal{A}$. Therefore, if ϕ is concave, the $\sup_{f\in\mathbb{R}} Q_r(f)$ is attained and uniquely defined (ϕ bounded above implies that $\phi(x) \to -\infty$ when $x \to \infty$ or $x \to -\infty$) making possible to write

$$f_\phi(x) = \arg\max_{f\in\mathbb{R}} Q_r(f),$$

and consequently, $f_\phi(x) = \arg\max_f V_\phi(f)$, where the maximum is taken over all measurable functions. In this case, by definition 1, ϕ is Infinite Sample Consistent if

$$d_\phi(x) := \text{sign}(f_\phi(x)) = d_*.$$

For the following theorem we assume that ϕ is concave, that is the general case we are interesterd in.

Theorem 12 *Assume that ϕ is concave. Then $d_\phi = d_*$, that is, d_ϕ is infinite sample consistent, if and only if $\phi(\cdot)$ is differentiable at zero and $\phi'(0) > 0$.*

Proof. For simplicity, we define

$$\eta := \frac{r_1}{r_1 + r_{-1}} = \frac{\mathbb{E}(R|X = x, A = 1)}{\mathbb{E}(R|X = x, A = 1) + \mathbb{E}(R|X = x, A = -1)}, \tag{71}$$

and, with a slight abuse in the notation, we write

$$Q_\eta(f) := \frac{1}{r_1 + r_{-1}} Q_r(f) = \phi(f)\eta + \phi(-f)(1 - \eta). \tag{72}$$

It is clear that $f_\phi(x)$ is also a maximizing argument of $Q_\eta(f)$. Note that using this notation, it is possible to rewrite the optimal Bayes decision d_* as

$$d_*(X) = \begin{cases} 1 & \text{If } \eta(X) \geq 1/2 \\ -1 & \text{If } \eta(X) < 1/2 \end{cases}.$$

We consider the **'if'** part of the proof first, and start by the case when $\eta < 1/2$. We need to prove that for all $x \in \mathcal{X}$, $f_\phi(x) < 0$. Because ϕ is concave, for any $h > 0$ it follows that

$$\begin{aligned} \phi(0) + h\phi'(0) &\geq \phi(h) \\ \phi(0) - h\phi'(0) &\geq \phi(-h). \end{aligned} \tag{73}$$

Therefore, noting that $Q_\eta(0) = \phi(0)$, it is derived that

$$\begin{aligned} Q_\eta(h) - Q_\eta(0) &= \eta\left(\phi(h) - \phi(0)\right) + (1 - \eta)\left(\phi(-h) - \phi(0)\right) \\ &\leq \eta\phi'(0)h - (1 - \eta)\phi'(0)h = h\phi'(0)\left(2\eta - 1\right), \end{aligned}$$

that is, given $\phi'(0) > 0$, for any $h > 0$, $Q_\eta(h) - Q_\eta(0) < 0$. Consequently, $f_\phi \leq 0$ because it is a maximum. To prove the strict inequality, note that given that ϕ is differentiable at zero, by definition, for any $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that

$$\begin{aligned} \delta^{-1}\left(\phi(\delta) - \phi(0)\right) &\geq \phi'(0) - \epsilon \\ \delta^{-1}\left(\phi(0) - \phi(-\delta)\right) &\leq \phi'(0) + \epsilon. \end{aligned}$$

This implies that

$$\begin{aligned}Q_\eta(-\delta)-Q_\eta(0) &= \eta\left(\phi(-\delta)-\phi(0)\right)+(1-\eta)\left(\phi(\delta)-\phi(0)\right)\\ &\geq \eta\delta\left(-\phi'(0)-\epsilon\right)+(1-\eta)\delta\left(\phi'(0)-\epsilon\right)=\delta\left(\phi'(0)(1-2\eta)-\epsilon\right),\end{aligned}$$

thus, making ϵ small enough, $(\phi'(0)(1-2\eta)-\epsilon)>0$. It follows that $f_\phi<0$.
Now we consider the case when $\eta\geq 1/2$. From the inequalities derived in 73, for any $h>0$,

$$Q_\eta(-h)-Q_\eta(0)\leq h\phi'(0)\left(1-2\eta\right)\leq 0,$$

and in consequence $f_\phi\geq 0$. This concludes the proof for the necessary conditions in the theorem.
We proceed now with the **'only if'** part of the proof. Note that if f_ϕ maximizes $Q_\eta(f)$, it follows that $Q_\eta(f_\phi)-\phi(0)\geq 0$, where

$$Q_\eta(f_\phi)-\phi(0)=\eta\left(\phi(f_\phi)-\phi(0)\right)+(1-\eta)\left(\phi(-f_\phi)-\phi(0)\right). \tag{74}$$

Note that in the case of $\eta<1/2$ the inequality is strict. The reason of this is that if $Q_\eta(f_\phi)=\phi(0)$ then $\phi(y)=c\in\mathbb{R}$ for $y\in[f_\phi,-f_\phi]$ (by the definition of concavity), and subsequently, $-f_\phi>0$ is also a maximizing argument, which contradicts the infinite sample consistency property.
We need to prove that $\phi'(0)>0$. Let $[a,b]$ be the subderivative of ϕ at zero, that is a and b are the right and left limits when $h\to 0$ of the function $h^{-1}(\phi(h)-\phi(0))$. First is going to be proved that $a>0$. Suposse by contradiction that $a\leq 0$. If $\eta<1/2$ then $f_\phi<0$ from the infinite sample consistency definition. Therefore, $\phi(-f_\phi)\leq\phi(0)$, and replacing in 74, it is necessary that $\phi(f_\phi)>\phi(0)$ in order to keep the optimality property of f_ϕ. Given that $f_\phi<0$ and the concavity of ϕ, we have that $a\leq b\leq 0$. Consequentely, the following two inequalities hold

$$\begin{gathered}\phi(f_\phi)-\phi(0)\leq bf_\phi\\ \phi(0)-\phi(-f_\phi)\geq af_\phi,\end{gathered}$$

which implies that $Q_\eta(f_\phi)-\phi(0)\leq f_\phi\left(b\eta-(1-\eta)a\right)\leq 0$, which contradicts that f_ϕ is the maximun. Therefore, it is concluded that $a>0$.
It remains to prove that $a=b$. To do so, suppose by contradiction that $0<a<b$. This implies that it is possible to define $a/(a+b)<\eta<1/2$, and therefore $f_\phi<0$. It is clear that

$$\begin{gathered}\phi(-f_\phi)-\phi(0)<-af_\phi\\ \phi(0)-\phi(f_\phi)>-bf_\phi,\end{gathered}$$

so, $Q_\eta(f_\phi) - \phi(0) < f_\phi^*(b\eta - (1-\eta)a) < 0$, which contraticts the fact that f_ϕ maximizes Q_η. It follows that ϕ is differentiable at zero and $\phi'(0) > 0$. □

3.3.1 Treatment Selection Withholding

For many clinical treatment selections the cost of choosing the incorrect treatment can be very sustantial. In such cases, it is preferable to define some no decisive region for which the assigment to an specific procedure is inconclusive. We consider the ITR problem with the option of not making any decision because the assigment is not strong enough. We call this the rejection option.

Considering two possible treatments, that is, $\mathcal{A} = \{-1, 1\}$; we define the Individualized Treatment Rule with rejection option as a function $d : \mathcal{X} \to \bar{\mathcal{A}}$ where $\bar{\mathcal{A}} = \{-1, 0, 1\}$. The optimal (Bayes) ITR d_* is the decision that maximizes the expected value function as

$$d_* := \arg\max_d \mathbb{E}\left[\frac{R\ell(A, d(X))}{p(A|X)}\right],$$

where

$$\ell(A, d) = \begin{cases} 0 & \text{If } d \neq A \text{ and } d \neq 0 \\ c & \text{If } d = 0 \\ 1 & \text{If } d = A \end{cases},$$

for some $0 \leq c < 1/2$. Defining η as in 71,

$$\eta(x) = \frac{\mathbb{E}(R|X = x, A = 1)}{\mathbb{E}(R|X = x, A = 1) + \mathbb{E}(R|X = x, A = -1)},$$

it is easy to check (see e.g. Barlett and Wegkamp (2008) and Yuan and Wegkamp(2010) for the classification problem) that

$$d_*(X) = \begin{cases} 1 & \text{If } \eta(X) > 1 - c \\ 0 & \text{If } c \leq \eta(X) \leq 1 - c \\ -1 & \text{If } \eta(X) < c \end{cases}.$$

Let $\phi(\cdot)$ be a surrogate concave function, and

$$f_\phi := \arg\max_f \mathbb{E}\left[\frac{R\phi(Af(X))}{p(A|X)}\right],$$

where the maximum is taken over all measurable functions f. The decision rule d associated with a function f can be determined as

$$d(X) := D(f(X), \delta)) = \begin{cases} 1 & \text{If } f(X) > \delta \\ 0 & \text{If } -\delta \leq f \leq \delta \\ -1 & \text{If } f < -\delta \end{cases},$$

for a parameter $\delta(c, \phi) \geq 0$. In particular, we call $d_\phi = D(f_\phi(X), \delta))$. The following theorem gives necessary and sufficient conditions on ϕ to be Infinite Sample Consistent.

Theorem 13 *Assume ϕ is concave. Then, $d_\phi = d_*$ if and only if $\phi(\cdot)$ is differentiable at δ and $-\delta$, $\phi'(\delta) > 0$, and*

$$\frac{\phi'(\delta)}{\phi'(\delta) + \phi'(-\delta)} = c.$$

Proof. Using the definition of $Q_\eta(f)$ in 81 it is clear that f_ϕ is a maximizing argument of it. We shall consider first the **'if'** part of the proof, which is going to be divided in three different cases: $\eta < c$, $\eta > 1 - c$ and $c \leq \eta \leq 1 - c$.

Case 1: $\eta < c$. Note that ϕ concave implies that $\phi'(-\delta) \geq \phi'(\delta) > 0$, and therefore, for any $h > 0$,

$$\begin{aligned} \phi(-\delta + h) - \phi(-\delta) &< \phi'(-\delta)h \\ \phi(\delta) - \phi(\delta - h) &> \phi'(\delta)h. \end{aligned} \tag{75}$$

Consequently, using the fact that $\phi'(-\delta) = \phi'(\delta)\frac{1-c}{c}$, we have

$$\begin{aligned} Q_\eta(h - \delta) - Q_\eta(-\delta) &< \eta h\phi'(\delta)\frac{1-c}{c} - (1-\eta)h\phi'(\delta) \\ &= h\phi'(\delta)\left(\frac{\eta}{c} - 1\right) < 0. \end{aligned}$$

This means that $f_\phi \leq -\delta$. We now have to prove that $f_\phi \neq -\delta$. Given ϕ differentiable at $-\delta$ and δ, for any $\epsilon > 0$ there exist some positive constants $\xi_1(\epsilon)$ and $\xi_2(\epsilon)$ such that

$$\begin{aligned} \xi_1^{-1}\left(\phi(-\delta) - \phi(-\delta - \xi_1)\right) \leq \phi'(-\delta) + \epsilon \\ \xi_2^{-1}\left(\phi(\delta + \xi_2) - \phi(\delta)\right) \geq \phi'(\delta) - \epsilon. \end{aligned} \tag{76}$$

Writing $\xi(\epsilon) = \max\{\xi_1(\epsilon), \xi_2(\epsilon)\}$ we have

$$\begin{aligned} Q_\eta(-\delta - \xi) - Q_\eta(-\delta) &= \eta\left(\phi(-\delta - \xi) - \phi(-\delta)\right) + (1-\eta)\left(\phi(\delta + \xi) - \phi(\delta)\right) \\ &\geq \eta\xi\left(-\phi'(-\delta) - \epsilon\right) + (1-\eta)\xi\left(\phi(\delta) - \epsilon\right) \\ &= \xi\left(\phi(\delta)\left(1 - \frac{\eta}{c}\right) - \epsilon\right). \end{aligned}$$

Therefore, for ϵ small enough, $Q_\eta(-\delta - \xi) - Q_\eta(-\delta) > 0$. It follows that $f_\phi < -\delta$.

Case 2: $\eta > 1 - c$. Using the inequalities 75, for any $h > 0$ we get

$$\begin{aligned} Q_\eta(\delta - h) - Q_\eta(\delta) &< -\eta h\phi'(-\delta)\frac{c}{1-c} + (1-\eta)h\phi(-\delta) \\ &= h\phi(-\delta)\left(1 - \frac{\eta}{1-c}\right) < 0, \end{aligned}$$

and therefore, $f_\phi \geq \delta$. In order to prove the strict inequality, it is possible to use the inequalities 75 and the respective definition of $\xi(\epsilon)$ to make

$$\begin{aligned} Q_\eta(\delta+\xi) - Q_\eta(\delta) &\geq \eta\xi\left(\phi'(\delta)-\epsilon\right) + (1-\eta)\xi\left(-\phi'(-\delta)-\epsilon\right) \\ &= \xi\left(\phi'(-\delta)\left(\frac{\eta}{1-c}-1\right)-\epsilon\right). \end{aligned}$$

It follows that for a properly choosen ϵ, $Q_\eta(\delta+\xi) - Q_\eta(\delta) > 0$ and thus $f_\phi > \delta$.

Case 3: $c \leq \eta \leq 1-c$. Observe that for any $h > 0$ the following inequalities hold

$$\begin{aligned} \phi(\delta+h) - \phi(\delta) &< \phi'(\delta)h \\ \phi(-\delta) - \phi(-\delta-h) &> \phi'(-\delta)h. \end{aligned}$$

Subsequently, we have that

$$Q_\eta(\delta+h) - Q_\eta(\delta) < h\phi'(-\delta)\left(\frac{\eta}{1-c}-1\right) < 0,$$

which implies that $f_\phi \leq \delta$. In a similar way

$$Q_\eta(-\delta-h) - Q_\eta(-\delta) < h\phi'(\delta)\left(1-\frac{\eta}{c}\right) < 0,$$

which means that $f_\phi \geq -\delta$. Therefore, $f_\phi \in [-\delta, \delta]$.

We consider now the **'only if'** part of the proof. Note first that given that f_ϕ maximizes $Q_\eta(f)$ the next two inequalities follow

$$\begin{aligned} Q_\eta(f_\phi) - Q_\eta(\delta) &= \eta\left(\phi(f_\phi)-\phi(\delta)\right) + (1-\eta)\left(\phi(-f_\phi)-\phi(-\delta)\right) \geq 0 \qquad (77) \\ Q_\eta(f_\phi) - Q_\eta(-\delta) &= \eta\left(\phi(f_\phi)-\phi(-\delta)\right) + (1-\eta)\left(\phi(-f_\phi)-\phi(\delta)\right) \geq 0. \qquad (78) \end{aligned}$$

Note the first inequality is strict if $\eta > 1-c$ and so is the second one when $\eta < c$. Let $[a_-, b_-]$ and $[a_+, b_+]$ the respective subderivatives of ϕ at $-\delta$ and δ. We need to prove that $a_+ > 0$, that $a_- = b_-$ and $a_+ = b_+$, and also that $a_+/(a_+ + a_-) = c$. We shall start by proving that $a_+ > 0$. Assume by contradiction that $a_+ \leq 0$. If $\eta > 1-c$, by the optimality of f_ϕ we have that $f_\phi > \delta$. Thus, $\phi(f_\phi) \leq \phi(\delta)$, and replacing in the inequality 77 it is concluded that $\phi(-f_\phi) > \phi(-\delta)$. In consequence, $b_- \leq 0$ and $a_- \leq 0$, which implies that $\phi(f_\phi) \leq \phi(-\delta)$. Replacing in 78 it follows that $\phi(-f_\phi) - \phi(\delta) \geq 0$, and thus, $a_+ \leq b_+ \leq 0$. Using this, ee can construct the next pair of inequalities

$$\begin{aligned} \phi(f_\phi) - \phi(\delta) &\leq a_+(f_\phi - \delta) \\ \phi(-f_\phi) - \phi(-\delta) &\leq -b_-(f_\phi - \delta), \end{aligned}$$

which implies that $Q_\eta(f_\phi) - Q_\eta(\delta) \leq (f_\phi - \delta)\left(a_+\eta - (1-\eta)b_-)\right) \leq (f_\phi - \delta)(2\eta - 1)b_- \leq 0$. This contradicts the optimality of f_ϕ and therefore it is prooved that $a_+ > 0$.

We will now prove that ϕ is differentiable at δ and $-\delta$, and that $a_+/(a_- + a_+) = c$. To do so, recall that we already proved that $0 < a_+ \leq b_+ \leq a_- \leq b_-$, and consequentely, $b_-/a_+ = a_-/b_+$ if and only if $a_- = b_-$ and $a_+ = b_+$. Therefore, it is sufficient to prove that

$$\frac{a_+}{a_+ + b_-} = c = \frac{b_+}{b_+ + a_-}.$$

Assume by contradiction that $\frac{a_+}{a_+ + b_-} < c$, which implies that $\frac{b_-}{a_+ + b_-} > 1 - c$. Suppose that $1 - c < \eta < \frac{b_-}{a_+ + b_-}$. It follows that $f_\phi > \delta$ and thus

$$\begin{aligned} \phi(f_\phi) - \phi(\delta) &\leq a_+(f_\phi - \delta) \\ \phi(-\delta) - \phi(-f_\phi) &\geq b_-(f_\phi - \delta). \end{aligned}$$

Replacing in inequality 77, we have that $Q_\eta(f_\phi) - Q_\eta(\delta) \leq (f_\phi - \delta)\left(\eta(a_+ + b_-) - b_-\right) < 0$ that is a contradiciton. Therefore, $\frac{a_+}{a_+ + b_-} \geq c$. Similarly, assume now that $\frac{b_+}{b_+ + a_-} > c$ and choose $c < \eta < \frac{b_+}{b_+ + a_-}$. This implies that $-\delta \leq f_\phi \leq \delta$ and

$$\begin{aligned} \phi(f_\phi) - \phi(-\delta) &\leq a_-(f_\phi + \delta) \\ \phi(\delta) - \phi(-f_\phi) &\geq b_+(f_\phi + \delta). \end{aligned}$$

Replacing in inequality 78 it follows that $Q_\eta(f_\phi) - Q_\eta(-\delta) \leq (f_\phi + \delta)\left(\eta(a_- + b_+) - b_+\right) < 0$, which contradicts the infinite sample consistency of f_ϕ, and in consequence, $\frac{b_+}{b_+ + a_-} \leq c$. That proves that ϕ is differentiable at $-\delta$ and δ and that

$$\frac{\phi'(\delta)}{\phi'(\delta) + \phi'(-\delta)} = c. \ \square$$

3.3.2 Multicategory Treatment Selection

Suppose the the treatment selection A belongs to the set $\mathcal{A} \in \{1, 2, \cdots, K\}$, and for any $d \in \mathcal{A}$ define

$$d_* := \arg\max_d \mathbb{E}_X \mathbb{E}_{A,R}\left[R\frac{\mathbb{I}(A = d(X))}{p(A|X)}\right] = \arg\max_d \sum_{c=1}^{K} \mathbb{E}\left[\mathbb{I}(d(x) = c)R | X = x, A = c\right].$$

Then $d_*(X) = \arg\max_c \mathbb{E}(R|X, A = c)$.

For each $c \in \{1, \cdots, K\}$ let $f_c : \mathcal{X} \to \mathbb{R}$ be a real function from the sample space of X and $\mathbf{f} = (f_1, \cdots, f_K)$. Suppose that there exists some $\mathbf{f} \in \Omega \subset \mathbb{R}^K$, and $\Phi : \Omega \to \mathbb{R}^K$, such that for all $x \in \mathcal{X}$

$$\mathbf{f}_\Phi(x) := \arg\max_{\mathbf{f}} \sum_{c=1}^{K} \mathbb{E}\left[\Phi_c(\mathbf{f})R | X = x, A = c\right], \tag{79}$$

where Φ_c is the c-th component of Φ. The Φ-classification is made as

$$d_\Phi(x) = \underset{c\in\{1,\cdots,K\}}{\arg\max}\, f_c^*(x),$$

where $\mathbf{f}_\Phi = (f_1^*, \cdots, f_K^*)$. It is assumed thereafter that Φ_c is continuous and bounded above for all $c \in \mathcal{A}$. Note that the maximum element of $\mathbf{f}_\Phi$ may not be uniquely defined, however a rule for breaking ties can be easily determined and as long as it remains always the same, it does not affect the consistency analysis. We shall assume hereafter that this maximum is unique.

In the multiclass classification, checking for infinite sample consistency depends highly on the specific procedure tha is utilized, concretely, on the particular selection of the set Ω and the surrogate function Φ. The nature of the problem difficults the use of the definition 79 given that, in general, the existence of a maximum point is not assured inside the set Ω. We shall use therefore a more general definition for infinite sample consistency associated with a particular method, that is, with the definition of $\Phi_1, \cdots, \Phi_k$ and Ω, instead of the solution in 79.

For all $X^* \in \mathcal{X}$ and $a \in \mathcal{A}$, define

$$\eta_a(X^*) = \frac{\mathbb{E}(R|X = X^*, A = a)}{\sum_{k=1}^K \mathbb{E}(R|X = X^*, A = k)}, \tag{80}$$

so that $\sum_{a=1}^K \eta_a(X^*) = 1$. As a function of X^* we define

$$Q_{\eta(X^*)}(\mathbf{f}(X^*)) = \sum_{a=1}^K \eta_a(X^*)\Phi_a(\mathbf{f}(X^*)), \tag{81}$$

and $Q^*_{\eta(X^*)} := \sup_{\mathbf{f}\in\Omega} Q_{\eta(X^*)}(\mathbf{f}(X^*))$. Note that $Q^*_{\eta(X^*)} < \infty$ because each Φ_a is continuous and bounded above. We will omit the dependency on X^* in the notation herafter.

We say that the classification method with surrogate functions $\Phi_1, \cdots, \Phi_k$ on the set $\Omega \subset \mathbb{R}^K$ is *Infinite Sample Consistent* if for all $\eta_1, \eta_2, \cdots, \eta_K$ such that $\eta_a \geq 0$ for $a \in \mathcal{A}$, $\sum_{a=1}^K \eta_a = 1$ and $\eta_c < \sup_{a\in\mathcal{A}} \eta_a$, it follows that

$$Q^*_\eta > \sup_{\mathbf{f}\in\Omega: f_c = \max_{a\in\mathcal{A}} f_a} Q_\eta(\mathbf{f}). \tag{82}$$

This definition implies that no matter the values of $\eta_1, \eta_2, \cdots, \eta_K$ (and therefore no matter the particular X^*) the maximization of 81 with respect to $\mathbf{f} \in \Omega$ always leads to the Bayes rule $d_*(X^*)$. We will present now sufficient conditions for Infinite Sample Consistency for two classification methods, the first one using $\Omega = \mathbb{R}^K$, and the second with the constrained domain $\Omega = \{\mathbf{f} \in \mathbb{R}^K : \sum_{a\in\mathcal{A}} f_a = 0\}$.

3.3.2.1 Pairwise Comparison.

In this case, for $\Omega = \mathbb{R}^K$ and

$$\Phi_a(\mathbf{f}) = \sum_{k=1}^{K} \phi(f_a - f_k),$$

for a real value function ϕ. Note that it is implied that ϕ is continuous and bounded above.

Theorem 14 *If $\phi : \mathbb{R} \to \mathbb{R}$ is a non-decreasing concave differentiable function with $\phi'(0) > 0$, and there exists some x_0 such that $\phi(x) = \phi(x_0)$ for all $x \geq x_0$; then the pairwise comparison method is Infinite Sample Consistent.*

Proof.

Assume first that there exists some $\mathbf{f}_\phi = (f_1^*, \cdots, f_K^*) \in \Omega$ such that $Q_\eta(\mathbf{f}_\phi) = Q_\eta^*$, that is, $Q_\eta(\mathbf{f})$ attains its maximum value and thus $\mathbf{f}_\phi \in \arg\max_{\mathbf{f}\in\Omega} Q_\eta(\mathbf{f})$. Suppose that $\eta_{k^*} > \eta_c$ for all $c \in \mathcal{A}$ such that $c \neq k^*$. We have to prove that $f_{k^*}^* > f_c^*$.

We will start proving by contradiction that $f_{k^*}^* \geq f_c^*$. Assume by the contrary that $f_{k^*}^* < f_c^*$, and define $\mathbf{g} = (g_1, \cdots, g_K) \in \Omega$ such that $g_c = f_{k^*}^*$, $g_{k^*} = f_c^*$ and $g_k = f_k^*$ for all other $k \in \mathcal{A}$. It follows that

$$\begin{aligned} Q_\eta(\mathbf{g}) &= Q_\eta^* + (\eta_c - \eta_{k^*}) \sum_{a=1}^{K} \phi(f_{k^*}^* - f_a^*) + (\eta_{k^*} - \eta_c) \sum_{a=1}^{K} \phi(f_c^* - f_a^*) \\ &= Q_\eta * + (\eta_{k^*} - \eta_c) \left[\phi(f_c^* - f_{k^*}^*) - \phi(f_{k^*}^* - f_c^*) + \sum_{a \neq c, k^*}^{K} (\phi(f_c^* - f_a^*) - \phi(f_{k^*}^* - f_a^*)) \right], \end{aligned}$$

however, ϕ non-decreasing and $\phi'(0) > 0$ implies that $\phi(f_c^* - f_{k^*}^*) > \phi(f_{k^*}^* - f_c^*)$, and $\phi(f_c^* - f_a^*) \geq \phi(f_{k^*}^* - f_a^*)$ for all $a \neq c, k^*$. Consequently, $Q_\eta(\mathbf{g}) > Q_\eta^*$, and this is a contradiction on the optimality of $\mathbf{f}_\phi$. It follows that $f_{k^*}^* \geq f_c^*$.

Now we wil prove that the inequality has to be strict, that is, $f_{k^*}^* > f_c^*$. Assume by contradiction that $f_{k^*}^* = f_c^*$. From the first order condition on the optimality of $\mathbf{f}_\phi$ it is derived that $0 = \left.\frac{\delta Q_\eta(\mathbf{f})}{\delta f_{k^*}}\right|_{\mathbf{f}_\phi} = \left.\frac{\delta Q_\eta(\mathbf{f})}{\delta f_c}\right|_{\mathbf{f}_\phi}$. Substracting the second term from the first one, we have

$$(\eta_{k^*} - \eta_c) \sum_{a=1}^{K} \phi'(f_{k^*}^* - f_a^*) = 0.$$

However, given that ϕ is non-decreasing, $\sum_{a=1}^{K} \phi'(f_{k^*}^* - f_a^*) \geq \phi'(f_{k^*}^* - f_{k^*}^*) + \phi'(f_{k^*}^* - f_c^*) = 2\phi'(0) > 0$. Consequently, $f_{k^*}^* > f_c^*$. This proves that $d_* = d_\Phi$ if $Q_\eta(\mathbf{f})$ attains its maximum in the set Ω.

We proceed to prove Infinite Sample Consistency according to the general definition 82. Note first that it is possible to define a sequence of real vectors $\mathbf{f}^{(m)} = (f_1^{(m)}, \cdots, f_K^{(m)})$ in

Ω such that $Q_\eta(\mathbf{f}^{(m)}) \to Q^*_\eta$. Thus, a classification method based on $\Phi_1, \cdots, \Phi_K$ and Ω is infinite sample consistent if and only if

$$\left\{k \in \mathcal{A} : \lim_{m\to\infty}\left(f_k^{(m)} - \sup_{a\in\mathcal{A}} f_a^{(m)}\right) = 0\right\} \subseteq \arg\max_{a\in\mathcal{A}} \eta_a, \tag{83}$$

as long as the set in the left hand side is not empty. Assume again that $\eta_{k^*} > \eta_c$ for all $c \in \mathcal{A}$ such that $c \neq k^*$. We need to prove that $\lim_{m\to\infty}\sup_{a\in\mathcal{A}} f_a^{(m)} = \lim_{m\to\infty} f_{k^*}^{(m)}$. Suppose by contradiction that this is not true. That implies that we can find a subsequence $\mathbf{f}^{(m')}$ such that $f_c^{(m')} \geq f_{k^*}^{(m')}$ for all m'. In the case that $\eta_{k^*} = 1$ this is clearly a contradiction because $Q_\eta(\mathbf{f}^{(m')})$ will not converge to the supremum. We can assume then that $\eta_{k^*} < 1$.
We claim that we can select $\mathbf{f}^{(m')}$ to be a bounded sequence. In order to prove that, note that because ϕ is non-decreasing and concave, for $x \to -\infty$, $\phi(x) \to -\infty$. We define the partition $\mathcal{A} = \mathcal{A}_+ \cup \mathcal{A}_0$, where $\mathcal{A} = \{a \in \mathcal{A} : \eta_a > 0\}$, and $\mathcal{A}_0$ contains the indexes with $\eta_a = 0$. Suppose that for one specific index a', $f_{a'}^{(m')}$ cannot be bounded below. If $a' \in \mathcal{A}_+$, then $Q_\eta(\mathbf{f}^{(m')})$ cannot be bouded below neither, and therefore it could not converge to the supremum. Moreover, note that $Q_\eta\left(\mathbf{f}^{(m')}\right) \to Q^*_\eta < \infty$ implies that $\left(f_{a'}^{(m')} - f_a^{(m')}\right)$ are bounded below for $a' \in \mathcal{A}_+$ and $a \in \mathcal{A}$. Therefore, $f_{a'}^{(m')}$ is a bounded sequence for all $a' \in \mathcal{A}_+$ and consequently, $f_k^{(m')}$ are bounded above for all $k \in \mathcal{A}_0$. It follows that for each m' it is possible to define a new sequence $\mathbf{h}^{(m')} = \left(h_1^{(m')}, \cdots, h_K^{(m')}\right)$ such that $h_a^{(m')} = f_a^{(m')} - \min_{k\in\mathcal{A}_+} f_k^{(m')}$. Note that all the arguments in $h_{a'}^{(m')}$ are non-negative for $a' \in \mathcal{A}_+$ and $Q_\eta\left(\mathbf{h}^{(m')}\right) \to Q^*_\eta < \infty$ because the value of $Q_\eta(\cdot)$ depends only on the differences $f_{a'}^{(m')} - f_a^{(m')}$. For $k \in \mathcal{A}_0$ we can redefine $h_k^{(m')} = \max\{h_k^{(m')}, -x_0\}$ without altering the value of $Q_\eta\left(\mathbf{h}^{(m')}\right)$. Now $\mathbf{h}^{(m')}$ is a bounded sequence.
We can choose a convergent subsequence $\mathbf{h}^{(m'')}$ to an element $\mathbf{h} \in \Omega$ such that $h_c^{(m')} \geq h_{k^*}^{(m')}$ and $Q_\eta(\mathbf{h}) = Q^*_\eta$, however this is a contradiction because we already proved that if $\mathbf{h} \in \Omega$ then $h_c^{(m')} < h_{k^*}^{(m')}$. This proves that the selection method is Infinite Sample Consistent.

3.3.2.2 Constrained Comparison Method

In this case, for $\Omega = \{f \in \mathbb{R}^K : \sum_{k=1}^K f_k = 0\}$,

$$\Phi_a(\mathbf{f}) = \sum_{k\neq a}^{K} \phi(-f_k),$$

for a continuous real value function ϕ bounded above.

Theorem 15 *If $\phi(\cdot)$ is a concave function differentiable in $(-\infty, 0]$ with $\phi'(0) > 0$, then the Constrained Comparison selection method is Infinite Sample Consistent.*

Proof.

Note that it is possible to write $Q_\eta(\mathbf{f}_\phi) = \sum_{a=1}^{K}(1-\eta_a)\phi(-f_a^*)$, and whitout loss of generality it is possible to assume that $1 \geq \eta_1 > \eta_2 \geq \cdots \geq \eta_K$. Similar to the proof of theorem 14, we assume first that there exists some $\mathbf{f}_\phi = (f_1^*, \cdots, f_K^*) \in \Omega$ such that $Q_\eta(\mathbf{f}_\phi) = Q_\eta^*$, that is, $Q_\eta(\mathbf{f})$ attains its maximum value and thus $\mathbf{f}_\phi \in \arg\max_{\mathbf{f}\in\Omega} Q_\eta(\mathbf{f})$.

First it is going to be proved that $f_1^* \geq 0$. Suppose by contradiction that $f_1^* < 0$, then, given the restriction in Ω there exists at least one $c \in \mathcal{A}$ such that $c \neq 1$ and $f_c^* > 0$ with $\eta_c < \eta_1$. The concavity of ϕ, and differentiability at 0 implies that for any constant $0 < x < \min\{-f_1^*, f_c^*\}$, $\phi(-f_c^* + x) - \phi(-f_c^*) > \phi(-f_1^*) - \phi(-f_1^* - x)$. Now, let $\mathbf{g}$ be a function in Ω such that $g_a = f_a^*$ for all $a \in \mathcal{A}$ such that $a \neq 1, c$, $g_1 = f_1^* - x$ and $g_c = f_c^* + x$. It follows that

$$\begin{aligned} Q_\eta(\mathbf{g}) - Q_\eta(\mathbf{f}_\phi) &= (1-\eta_1)\left[\phi(-f_1^* - x) - \phi(-f_1^*)\right] + (1-\eta_c)\left[\phi(-f_c^* + x) - \phi(-f_c^*)\right] \\ &\geq (1-\eta_1)\left(\left[\phi(-f_c^* + x) - \phi(-f_c^*)\right] - \left[\phi(-f_1^*) - \phi(-f_1^* - x)\right]\right). \end{aligned}$$

Therefore, $Q_\eta(\mathbf{g}) - Q_\eta^* > 0$ if $\eta_1 < 1$, and because the case $\eta_1 = 1$ trivially satisfies the condition, this contradicts the optimality of $\mathbf{f}_\phi$, ant therefore, $f_1^* \geq 0$.

Now it is going to be proved that $f_1^* > f_c^*$ for any $c \in \mathcal{A}$ such that $c \neq 1$. Suppose by contradiction that there exist one c for which $f_c^* > f_1^* \geq 0$, then, by the concavity of ϕ, it follows that

$$\phi(-f_c^*) \leq \phi(-f_1^*) - \phi'(-f_1^*)(f_c^* - f_1^*).$$

If $\mathbf{h} \in \Omega$ is defined as $h_a = f_a^*$ for all $a \in \mathcal{A}$ such that $a \neq 1, c$, $h_1 = f_c^*$ and $h_c = f_1^*$, we have

$$\begin{aligned} Q_\eta(\mathbf{h}) - Q_\eta(\mathbf{f}_\phi) &= (\eta_1 - \eta_c)\left[\phi(-f_1^*) - \phi(-\phi_c^*)\right] \\ &\geq (\eta_1 - \eta_c)\phi'(-f_1^*)(f_c^* - f_1^*) \geq (\eta_1 - \eta_c)\phi'(0)(f_c^* - f_1^*) > 0. \end{aligned}$$

This contradicts the optimality of $\mathbf{f}_\phi$. We have proved then that if $Q_\eta(\mathbf{f}_\phi) = Q_\eta^*$ with $\mathbf{f}_\phi \in \Omega$, $d_\Phi = d_*$.

Now we will prove Infinite Sample Consistency according to definition 82. Suppose by contradiction that there exist a sequence $\mathbf{f}^{(m)} = \left(f_1^{(m)}, \cdots, f_K^{(m)}\right) \in \Omega$ such that $Q_\eta\left(\mathbf{f}^{(m)}\right) \to Q_\eta^*$ and $f_c^{(m)} \geq f_1^{(m)}$ for all m and $c \in \mathcal{A}$ such that $c \neq 1$. By concavity of ϕ, and given $\phi'(0) > 0$, it follows that $\phi(x) \to -\infty$ as $x \to -\infty$. Therefore, boundedness in the sequence $Q_\eta\left(\mathbf{f}^{(m)}\right)$ implies that each $f_a^{(m)}$ is bounded above for $a \in \mathcal{A}$. Given that Ω imposes the restriction $\sum_{a\in\mathcal{A}} f_a^{(m)} = 0$ for all m, we have that all the $f_a^{(m)}$ are also bounded below. Thus, it is possible to choose a convergent subsequence $\mathbf{f}^{(m')} \to \mathbf{f} = (f_1 \cdots, f_K)$ such that $Q_\eta(\mathbf{f}) = Q_\eta^*$, but this is a contradiction because we proved already that if $\mathbf{f} \in \Omega$, then $f_1 > f_c$. This proves that the Constrained Comparison Method is Infinite Sample Consistent.

APPENDIX A

RELEVANT PROOFS IN CHAPTER I

In order to make the presentations of the proofs more clear we will assume thereafter that $\mathbb{E}(X) = 0$ and $\tau = [0, 1]$.

A.1 Proof of Theorem 1

(i) For any $\beta^* \in \mathcal{H}$ it follows that

$$\begin{aligned} \langle f, U(\beta^*)f\rangle_{\mathcal{L}_2} &= D^2\ell_\infty(\beta^*)ff \\ &= \mathbb{E}_X\left[\omega^{(2)}\left(\int_\tau X\beta^*\right)\left(\int_\tau Xf\right)^2\right]. \end{aligned}$$

By Cauchy-Schwarz inequality and applying assumptions 1 (ii) and (iii), it follows that there exits a constant $0 \leq M' < \infty$ such that

$$\langle f, U(\beta^*)f\rangle_{\mathcal{L}_2} \leq M'\mathbb{E}\left(\int_\tau Xf\right)^2 = M'\langle Cf, f\rangle_{\mathcal{L}_2}. \tag{84}$$

Also, strict convexity of ω implies that for any $\theta \in \mathbb{R}$, $\omega^{(2)}(\theta) > 0$. By assumption 1 (i), $\omega^{(2)}$ is continuously differentiable, and therefore, for any $r > 0$, there exists an $\epsilon > 0$ such that $\omega^{(2)}(\theta) \geq \epsilon$ if $|\theta| \leq r$. In order to make the notation more clear, for $\beta^* \in \mathcal{H}_\alpha$ such that $\|\beta^*\|_\alpha^2 \leq M$, we will write $W^* := \int_\tau X\beta^*$. Note that boundednes of the operator $C^{1/2}$ implies that $\mathrm{Var}(W^*) = \langle\beta^*, C\beta^*\rangle_{\mathcal{L}_2} = \langle C^{1/2}\beta^*, C^{1/2}\beta^*\rangle_{\mathcal{L}_2} \leq cM$, for some positive constant c. Now, using conditional probability we have

$$\begin{aligned} \mathbb{E}\left[\omega^{(2)}(W^*)\left(\int_\tau Xf\right)^2\right] &\geq \mathbb{E}\left[\omega^{(2)}(W^*)\left(\int_\tau Xf\right)^2\middle||W^*| \leq r\right]P\left(|W^*| \leq r\right) \\ &\geq \epsilon\langle f, Cf\rangle_{\mathcal{L}_2}P\left(|W^*| \leq r\right)^2. \end{aligned}$$

By Markov inequality, it follows that $P\left(|W^*| \leq r\right) \geq (1 - \frac{cM}{r^2})$, thus, for $r^2 < cM$ we conclude that there exists some constant $m' > 0$ such that $\langle f, U(\beta^*)f\rangle_{\mathcal{L}_2} > m'\langle f, Cf\rangle_{\mathcal{L}_2}$.

From Proposition 2 in Yuan and Cai (2010) we have that there exist some constants $0 < a \leq b$ such that for any $f \in \mathcal{H}$ it follows that $a\|f\|_{\mathcal{H}}^2 \leq \langle f, Cf\rangle_{\mathcal{L}_2} + J(f) \leq b\|f\|_{\mathcal{H}}^2$, and therefore, there constants $0 < m'' \leq M''$ such that

$$m''\|f\|_{\mathcal{H}}^2 \leq \langle f, U(\beta^*)f\rangle_{\mathcal{L}_2} + J(f) \leq M''\|f\|_{\mathcal{H}}^2.$$

Consequently, $\|\cdot\|_{R^*}$ and $\|\cdot\|_{\mathcal{H}}$ are equivalent norms.

(ii) Let $R^*(\cdot,\cdot)$ be the reproducing kernel associated with the norm $\|\cdot\|_{R^*}$, and R^* the associated integral operator on $\mathcal{L}_2(\tau)$. Now we define a positive integral operator as $R^{*1/2}U(\beta^*)R^{*1/2}$, and by Mercer's theorem it has a spectral decomposition with eigenvalues $\nu_1^* \geq \nu_2^* \geq \cdots$ and eigenfunctions $\{\zeta_1^*, \zeta_2^*, \cdots\}$. Defining $\varphi_k^* = \nu_k^{*-1/2} R^{*1/2}\zeta_k^*$ for each $k \geq 1$ it follows that

$$\langle \varphi_j^*, \varphi_k^* \rangle_{R^*} = \left(\nu_j^* \nu_k^*\right)^{-1/2} \langle \zeta_j^*, R^* \zeta_k^* \rangle_{R^*} = \left(\nu_j^* \nu_k^*\right)^{-1/2} \langle \zeta_j^*, \zeta_k^* \rangle_{\mathcal{L}_2} = \nu_k^{*-1}\delta_{jk},$$

and

$$\langle \varphi_j^*, U(\beta^*)\varphi_k^* \rangle_{\mathcal{L}_2} = \left(\nu_j^* \nu_k^*\right)^{-1/2} \langle \zeta_j^*, R^{*1/2}U(\beta^*)R^{*1/2}\zeta_k^* \rangle_{\mathcal{L}_2} = \delta_{jk}.$$

(iii) Because $\{\zeta_1^*, \zeta_2^*, \cdots\}$ forms an orthonormal basis for $\mathcal{L}_2(\tau)$, it is possible to write $f = \sum_{k\geq 1} \zeta_k^* \langle f, \zeta_k^* \rangle_{\mathcal{L}_2}$, and given that $R^{*1/2}$ is a positive definite bounded operator, we have that $f = R^{*1/2}R^{*-1/2}f$, where

$$\begin{aligned} R^{*-1/2}f &= \sum_{k\geq 1} \zeta_k^* \langle R^{*-1/2}f, \zeta_k^* \rangle_{\mathcal{L}_2} = \sum_{k\geq 1} \nu_k^* R^{*-1/2}\varphi_k^* \langle R^{*-1/2}f, R^{*-1/2}\varphi_k^* \rangle_{\mathcal{L}_2} \\ &= R^{*-1/2} \sum_{k\geq 1} \nu_k^* \varphi_k^* \langle f, \varphi_k^* \rangle_{R^*}, \end{aligned}$$

and it follows that $f = \sum_{k\geq 1} f_k^* \varphi_k^*$.

(iv) This is a direct consequence of the equivalence between the norms $\|\cdot\|_R$ and $\|\cdot\|_{R^*}$ (both of them are equivalent to $\|\cdot\|_{\mathcal{H}}$). Note that for any $f \in \mathcal{H}$ there exists some constant $c \geq 1$ such that $c^{-1}\|f\|_R \leq \|f\|_{R^*} \leq c\|f\|_R$. We define two sequences of functions in $\mathcal{H}$ as $g_k := \varphi_k \nu_k$ and $h_k := \varphi_k^* \nu_k^*$. Note that $\lim_{k\to\infty} \|g_k\|_R = \lim_{k\to\infty} \nu_k = 0$, and similarly, $\lim_{k\to\infty} \|g_k\|_{R^*} = 0$. By the equivalence ot the two norms, we have that $(g_k - h_k) \to 0$, and thus, by triangle inequality

$$\|g_k - h_k\|_R \geq \Big| \|g_k\|_R - \|h_k\|_R \Big| = \Big| \nu_k - \|h_k\|_R \Big|.$$

Therefore, for k large enough there exists some constant c' such that $c'^{-1}\|h_k\|_R \leq \nu_k \leq c'\|h_k\|_R$, but because $\|h_k\|_{R^*} = \nu_k^*$, making $c'' = c \cdot c'$ it follows that

$$c''^{-1}\nu_k^* \leq \nu_k \leq c''\nu_k^*,$$

and the result follows from writting $m = c''^{-1}$ and $M = c''$. □

A.2 Proof of Proposition 3

Recall that for $\beta^* \in \mathcal{H}$, $G_{\infty\lambda}(\beta^*) = D^2\ell_\infty(\beta^*) + \lambda D^2 J$. By theorem 1 and 1.2.3 it is concluded that for $f, g \in \mathcal{H}$, $A(f,g) := \langle f, G_{\infty\lambda}(\beta^*)g\rangle_{R^*}$ is a bounded and coercive bilinear form in $\mathcal{H}$. Therefore, by Lax-Milgram lemma $G_{\infty\lambda}(\beta^*)^{-1}$ is a bounded operator on $\mathcal{H}$.

We will start by proving that for any $\xi_1 \in \mathcal{H}_{\alpha^*}$, it happens that $D\ell_\infty(\xi_1) \in \mathcal{H}$. By lemma 3.1 in Cox (1988) and lemma 2.1 in Cox and O'Sullivan (1990) it is derived that for any $\theta \in \mathcal{H}_\alpha$ and some constant c, $\theta \in \mathcal{H}_{\alpha+c^*}$ (that is, $\|\theta\|^2_{\alpha+c^*} < \infty$) if $\sup_{\theta'}\langle\theta, \theta'\rangle_{R^*} < \infty$ for all $\theta' \in \mathcal{H}_\alpha$ such that $\|\theta'\|^2_{\alpha-c^*} = 1$. Folllowing the same arguments that they present, we have that $D\ell_\infty(\xi_1)$ (as a element in the dual space of $\mathcal{H}_{\alpha^*}$) belongs to $\mathcal{H}_{2-\alpha^*}$ and therefore $D\ell_\infty(\xi_1) \in \mathcal{H}$. Similar analysis can be applied to $D\ell_n(\xi_1)$.

Now, as showed before, $J(\theta) := \langle\theta, W\theta\rangle_{\mathcal{H}}$ can be extended to an operator in $\mathbf{B}(\mathcal{H}_\alpha, \mathcal{H}_\alpha)$. Therefore, for any $\xi_1 \in \mathcal{H}_\alpha$ we have that $DJ(\xi_1) = W\xi_1 \in \mathcal{H}_\alpha$, and

$$G_{\infty\lambda}(\beta^*)^{-1}W\xi_1 = G_{\infty\lambda}(\beta^*)^{-1}\frac{1}{\lambda}\left[G_{\infty\lambda}(\beta^*) - D^2\ell_\infty(\beta^*)\right]\xi_1.$$

Thus, it is enough to prove that $D^2\ell_\infty(\beta^*)\xi_1$ is a wel defined element in $\mathcal{H}$. To check this, note that for any $\theta \in \mathcal{H}$

$$\langle\theta, D^2\ell_\infty(\beta^*)\xi_1\rangle_{R^*} = \langle\theta, U(\beta^*)\xi_1\rangle_{\mathcal{L}_2} = 2\langle\theta, \xi_1\rangle_{0^*} \leq m\|\theta\|_0\|\xi_1\|_0, \tag{85}$$

for some constant m. Therefore, $D^2\ell_\infty(\beta^*)\xi_1 \in \mathcal{H}_2 \subset \mathcal{H}$. It follows that $D\ell_\infty\lambda(\xi_1) \in \mathcal{H}$, and therefore, $G_{\infty\lambda}(\beta^*)^{-1}D^2\ell_\infty(\beta^*)\xi_1$ is a well defined element in $\mathcal{H} \subset \mathcal{H}_\alpha$.

The other different cases mentioned in the Proposition can be derived from this, or can be presented in a similar way. We will not show each of them fro brevity in the presentation.

□

A.3 Proof of Theorem 4

Note first that

$$(\beta_{\infty\lambda} - \beta_0) = \left(\beta_{\infty\lambda} - \bar{\beta}_{\infty\lambda}\right) + \left(\bar{\beta}_{\infty\lambda} - \beta_0\right),$$

and replacing $\left(\beta_{\infty\lambda} - \bar{\beta}_{\infty\lambda}\right)$ as in equation 15 we obtain

$$\bar{\beta}_{\infty\lambda} - \beta_0 = \phi_1 + G^{-1}_{\infty\lambda}(\beta_0)\int_{\mathcal{T}}\int_{\mathcal{T}} x_1\left[D^3\ell_\infty\left(\beta_0 + x_1x_2\phi_1\right)\phi_1\phi_1\right]dx_1dx_2, \tag{86}$$

for $\phi_1 = (\beta_{\infty\lambda} - \beta_0)$. In order to bound this expression, we define

$$K_3(\lambda, a) = \sup_{\beta_3\in\mathcal{H}_\alpha}\sup_{u,v}\|G^{-1}_{\infty\lambda}(\beta_0)D^3\ell_\infty(\beta_3)uv\|_a \tag{87}$$

for $0 \leq a \leq \alpha$ and $u, v \in \mathcal{H}_\alpha$ such that $\|u\|_\alpha = \|v\|_\alpha = 1$. It is easy to check from 86 that

$$\|\beta_{\infty\lambda} - \beta_0\|_a \leq \|\bar{\beta}_{\infty\lambda} - \beta_0\|_a + \frac{1}{2}K_3(\lambda, a)\|\beta_{\infty\lambda} - \beta_0\|_\alpha^2. \tag{88}$$

The following lemma is going to be crucial for the continuation of the analysis. For clarity in the presentation, its proof wil be shown after the proof of theorem 6.

Lemma 16 *If* $\alpha \leq \frac{1}{2}\left(1 - \frac{1}{2(r+s)}\right)$*, then, for any* $0 \leq a \leq \alpha$*,* $K_3(\lambda, a)\|\bar{\beta}_{\infty\lambda} - \beta_0\|_\alpha \to 0$ *as* $\lambda \to 0$. □

For $\vartheta \in \mathcal{H}_\alpha$ we define the operator

$$H_\lambda(\vartheta) = \left(\bar{\beta}_{\infty\lambda} - \beta_0\right) + G^{-1}_{\infty\lambda}(\beta_0)\int_\tau\int_\tau x_1\left[D^3\ell_\infty\left(\beta_0 + x_1x_2\vartheta\right)\vartheta\vartheta\right]dx_1dx_2. \tag{89}$$

If $\bar{\beta}_{\infty\lambda} \in \mathcal{H}_\alpha$, it is clear that $H_\lambda(\beta_{\infty\lambda} - \beta_0) = (\beta_{\infty\lambda} - \beta_0)$ and therefore $\left(\beta_{\infty\lambda} - \bar{\beta}_{\infty\lambda}\right) = H_\lambda(\beta_{\infty\lambda} - \beta_0) - H_\lambda(0)$. From the definition of $K_3(\lambda, a)$ it follows that for $a \leq \alpha$

$$\|H_\lambda(\vartheta)\|_a \leq \|\left(\bar{\beta}_{\infty\lambda} - \beta_0\right)\|_a + \frac{1}{2}K_3(\lambda, a)\|\vartheta\|_\alpha^2.$$

Let $b_\beta(r)$ a closed ball in $\mathcal{H}_\alpha$ centered in β with radious r. Defining $d_\lambda = \|\bar{\beta}_{\infty\lambda} - \beta_0\|_\alpha$, then for any $\vartheta \in b_0(2d_\lambda)$ it follows that $\|H_\lambda(\vartheta)\|_\alpha \leq d_\lambda + 2K_3(\lambda, \alpha)d_\lambda^2$. Using lemma 16 it is possible to choose a fixed λ_0 such that for $\lambda \leq \lambda_0$, $d_\lambda \leq 1$ and $K_3(\lambda, \alpha)d_\lambda \leq \frac{1}{2}$. Therefore, $H_\lambda(b_0(2d_\lambda)) \in b_0(2d_\lambda)$.

Now, using a Taylor series expansion of the functional $D\ell_{\infty\lambda}(\beta_0 + \vartheta)$ around β_0 it is easy to check that

$$H_\lambda(\vartheta) = G^{-1}_{\infty\lambda}(\beta_0)\left[D\ell_{\infty\lambda}(\beta_0 + \vartheta) - D\ell_{\infty\lambda}(\beta_0)\right] + \left(\bar{\beta}_{\infty\lambda} - \beta_0\right) - \vartheta,$$

and therefore

$$H_\lambda(\vartheta_1) - H_\lambda(\vartheta_2) = G^{-1}_{\infty\lambda}(\beta_0)\left[D\ell_{\infty\lambda}(\beta_0 + \vartheta_1) - D\ell_{\infty\lambda}(\beta_0 + \vartheta_2)\right] - (\vartheta_1 - \vartheta_2). \tag{90}$$

Expanding a Taylos series of $D\ell_{\infty\lambda}(\beta_0 + \vartheta_1)$ around $(\beta_0 + \vartheta_2)$, it is derived that

$$[D\ell_{\infty\lambda}(\beta_0 + \vartheta_1) - D\ell_{\infty\lambda}(\beta_0 + \vartheta_2) = \int_\tau D^2\ell_{\infty\lambda}(\beta_0 + \vartheta_2 + x(\vartheta_1 - \vartheta_2))(\vartheta_1 - \vartheta_2)dx.$$

Now for the term inside the integral of the last expression we can use a Taylor series expansion once again around β_0 as

$$\begin{aligned} D^2\ell_{\infty\lambda}(\beta_0 + \vartheta_1 + s(\vartheta_1 - \vartheta_2))(\vartheta_1 - \vartheta_2) &= D^2\ell_{\infty\lambda}(\beta_0)(\vartheta_1 - \vartheta_2) \\ &+ \int_\tau D^3\ell_{\infty\lambda}(\beta_0 + y(\vartheta_2 + x(\vartheta_1 - \vartheta_2)))(\vartheta_1 - \vartheta_2)(\vartheta_2 + x(\vartheta_1 - \vartheta_2))dy. \end{aligned} \tag{91}$$

If $\vartheta_1, \vartheta_2 \in b_0(2d_\lambda)$, then $(\vartheta_2 + x(\vartheta_1 - \vartheta_2)) \in b_0(2d_\lambda)$ for all $x \in [0,1]$. Replacing 91 in 90 and taking the $\mathcal{H}_\alpha$ norm, it follows that

$$\|H_\lambda(\vartheta_1) - H_\lambda(\vartheta_2)\|_\alpha \leq K_3(\lambda,\alpha)\|\vartheta_1 - \vartheta_2\|_\alpha 2d_\lambda \leq \|\vartheta_1 - \vartheta_2\|_\alpha, \tag{92}$$

if $\lambda \leq \lambda_0$. It is concluded then that $H_\lambda(\vartheta)$ is a contraction map on $b_0(2d_\lambda)$ and consequently, by the contraction theorem (see e.g. Debnath and Mikusiński (1990)) there exists an unique function ϑ^* such that $H_\lambda(\vartheta^*) = \vartheta^*$. It follows that $(\beta_{\infty\lambda} - \beta_0) \in b_0(2d_\lambda)$ and this completes the proof. □

A.4 Proof of Theorem 5

We start by defining the following expressions

$$K_2(\lambda, a) = \sup_{\beta \in \mathcal{H}_\alpha} \sup_{u} \|G^{-1}_{\infty\lambda}(\beta_{\infty\lambda})\left(D^2\ell_{n\lambda}(\beta)u - D^2\ell_{\infty\lambda}(\beta)u\right)\|_a \tag{93}$$

$$K_{3n}(\lambda, a) = \sup_{\beta \in \mathcal{H}_\alpha} \sup_{u,v} \|G^{-1}_{\infty\lambda}(\beta_{\infty\lambda})D^3\ell_{n\lambda}(\beta)uv\|_a, \tag{94}$$

for $a \leq \alpha$ and vectors $u, v \in \mathcal{H}_\alpha$ such that $\|u\|_\alpha = \|v\|_\alpha = 1$. Wrinting $d_n = \|\bar{\beta}_{n\lambda} - \beta_{\infty\lambda}\|_\alpha$ and the open ball in $b_0(2d_n) \in \mathcal{H}_\alpha$, we define the following operator for all $\vartheta \in b_0(2d_n)$

$$H_n(\vartheta) = (\bar{\beta}_{n\lambda} - \beta_{\infty\lambda}) + G^{-1}_{\infty\lambda}(\beta_{\infty\lambda})D^2\ell_{n\lambda}(\beta_{\infty\lambda})\vartheta - \vartheta \\ + G^{-1}_{\infty\lambda}(\beta_{\infty\lambda})\int_{\mathcal{T}}\int_{\mathcal{T}} x_1\left[D^3\ell_n\left(\beta_{\infty\lambda} + x_1x_2\vartheta\right)\vartheta\vartheta\right]dx_1dx_2. \tag{95}$$

Using equation 16, it is clear that $H_n(\hat{\beta}_{n\lambda} - \beta_{\infty\lambda}) = \hat{\beta}_{n\lambda} - \beta_{\infty\lambda}$. We will prove that $H_n(\cdot)$ is a contraction map in $b_0(2d_n)$, for which we need to bound some results using the following lemma.

Lemma 17 *For $\frac{1}{2(r+s)} < \alpha \leq 1$, if there exists a sequence λ_n such that $n^{-1}\lambda_n^{-\alpha-\frac{1}{2(r+s)}} \to 0$, as $n \to \infty$, then for $0 \leq a \leq \alpha$,*

$$[K_2(\lambda_n, a) + K_3(\lambda_n, a)d_n] = o_p(1). \square$$

Now, for any $\vartheta \in \mathcal{H}_\alpha$,

$$\|H_n(\vartheta)\|_\alpha \leq d_n + K_2(\lambda,\alpha)\|\vartheta\|_\alpha + \frac{1}{2}K_{3n}(\lambda,\alpha)\|\vartheta\|^2_\alpha,$$

and if $\vartheta \in b_0(2d_n)$, it is possible to select n_0 large enough to make $(K_2(\lambda_n, \alpha) + K_3(\lambda_n, \alpha)d_n) < \frac{1}{2}$ with high probability. This implies that $\|H_n(\vartheta)\|_\alpha \leq d_n(1 + 2\frac{1}{2})$, that is, $H_n(b_0(2d_n)) \in b_0(2d_n)$.

Expanding a Taylor saries of the functional $D\ell_{n\lambda}(\beta_{\infty\lambda}+\vartheta)$ around $\beta_{\infty\lambda}$ it is possible to check in 95 that

$$H_n(\vartheta) = (\bar{\beta}_{n\lambda} - \beta_{\infty\lambda}) + G^{-1}_{\infty\lambda}(\beta_{\infty\lambda})\left[D\ell_{n\lambda}(\beta_{\infty\lambda}+\vartheta) - D\ell_{n\lambda}(\beta_{\infty\lambda})\right] - \vartheta, \tag{96}$$

and therefore, for $\vartheta_1, \vartheta_2 \in \mathcal{H}_\alpha$,

$$H_n(\vartheta_1) - H_n(\vartheta_2) = G^{-1}_{\infty\lambda}(\beta_{\infty\lambda})\left[D\ell_{n\lambda}(\beta_{\infty\lambda}+\vartheta_1) - D\ell_{n\lambda}(\beta_{\infty\lambda}+\vartheta_2)\right] - (\vartheta_1 - \vartheta_2). \tag{97}$$

Making a Taylor expansion of $D\ell_{n\lambda}(\beta_{\infty\lambda}+\vartheta_1)$ around $(\beta_{\infty\lambda}+\vartheta_2)$ as in the proof of theorem 4 is can be easily derived that

$$\begin{aligned} H_n(\vartheta_1) - H_n(\vartheta_2) &= G^{-1}_{\infty\lambda}(\beta_{\infty\lambda})D^2\ell_{n\lambda}(\beta_{\infty\lambda})(\vartheta_1 - \vartheta_2) - (\vartheta_1 - \vartheta_2) \\ &+ G^{-1}_{\infty\lambda}(\beta_{\infty\lambda})\int_{\mathcal{T}} D^3\ell_{n\lambda}(\beta_{\infty\lambda} + y(\vartheta_2 + x(\vartheta_1 - \vartheta_2)))(\vartheta_1 - \vartheta_2)(\vartheta_2 + x(\vartheta_1 - \vartheta_2))dy, \end{aligned} \tag{98}$$

so, if $\vartheta_1, \vartheta_2 \in b_0(2d_n)$, with probability arbitrarily close to 1 as $n \to \infty$,

$$\|H_n(\vartheta_1) - H_n(\vartheta_2)\|_\alpha \leq \left[K_2(\lambda_n, \alpha) + K_{3n}(\lambda_n, \alpha)2d_n\right]\|\vartheta_1 - \vartheta_2\|_\alpha \leq \|\vartheta_1 - \vartheta_2\|_\alpha. \tag{99}$$

From this result follows that $H_n(\cdot)$ is a contraction map on $b_0(2d_n)$ and therefore there exists an unique $\vartheta \in b_0(2d_n)$ such that $H_n(\vartheta) = \vartheta$. It is concluded then that with probability tending to 1 as $n \to \infty$, $\|\hat{\beta}_{n\lambda_n} - \beta_{\infty\lambda_n}\|_a \leq 2d_n$ for $0 \leq a \leq \frac{1}{2}\left(1 - \frac{1}{2(r+s)}\right)$ and this completes the proof. □

A.5 Proof of Theorem 6

From decomposition 12 we know that

$$\|\hat{\beta}_{n\lambda} - \beta_0\|_a \leq \|\hat{\beta}_{n\lambda} - \beta_{\infty\lambda}\|_a + \|\beta_{\infty\lambda} - \beta_0\|_a,$$

and if conditions of theorems 4 and 5 are satisfied it follows that for $0 \leq a \leq \alpha$ and $\alpha = \left(1 - \frac{1}{2(r+s)}\right)$,

$$\sup_{F\in\mathcal{F}(s,M)}\sup_{\beta_0\in\mathcal{H}}\|\hat{\beta}_{n\lambda_n} - \beta_0\|_a = O_p\left(\|\bar{\beta}_{\infty\lambda_n} - \beta_0\|_a + \|\bar{\beta}_{n\lambda_n} - \beta_{\infty\lambda_n}\|_a\right). \tag{100}$$

In order to prove the theorem, we will focuse to convergence of the linearized versions for the bias and varaince respectively presented in the las expression.

A.5.1 Convergence rate for the linearized bias.

From definition 13 we know that $\bar{\beta}_{\infty\lambda} - \beta_0 = -G^{-1}_{\infty\lambda}(\beta_0) D\ell_{\infty\lambda}(\beta_0)$, and given that $D\ell_\infty(\beta_0) = 0$ we can write $D\ell_{\infty\lambda}(\beta_0) = D\ell_{\infty\lambda}(\beta_0) - D\ell_\infty(\beta_0) = \lambda DJ(\beta_0)$. Using the fact that $DJ(f)g = \sum_{k\geq 1} \gamma_k^{-1} f_k g_k$, and making $\beta_0 = \beta^*$ it follows that

$$\begin{aligned}
\|\bar{\beta}_{\infty\lambda} - \beta_0\|_a^2 &= \|G^{-1}_{\infty\lambda}(\beta_0) D\ell_{\infty\lambda}(\beta_0)\|_a^2 \\
&= \lambda^2 \|G^{-1}_{\infty\lambda}(\beta_0) DJ(\beta_0)\|_a^2 \\
&\leq \lambda^2 C \|G^{-1}_{\infty\lambda}(\beta_0) DJ(\beta_0)\|_{a^*}^2 \\
&= \lambda^2 C \sum_{k\geq 1} (1 + \gamma_k^{*-a})(1 + \lambda\gamma_k^{*-1})^{-2} \left(\gamma_k^{*-1} (\beta_0)_k^*\right)^2,
\end{aligned} \tag{101}$$

where $(\beta_0)_k^* = \nu_k^* \langle \beta_0, \varphi_k^* \rangle_{R^*}$. To bound the sum in 101 we can write

$$\begin{aligned}
\sum_{k\geq 1} (1 + \gamma_k^{*-a})(1 + \lambda\gamma_k^{*-1})^{-2} \left(\gamma_k^{*-1} (\beta_0)_k^*\right)^2 &\leq \sup_{k\geq 1} \frac{(1 + \gamma_k^{*-a})\gamma_k^{*-1}}{(1 + \lambda\gamma_k^{*-1})^2} \sum_{k\geq 1} \gamma_k^{*-1} (\beta_0)_k^{*2} \\
&= J(\beta_0) \sup_{k\geq 1} \frac{(1 + \gamma_k^{*-a})\gamma_k^{*-1}}{(1 + \lambda\gamma_k^{*-1})^2},
\end{aligned}$$

and

$$\begin{aligned}
\sup_{k\geq 1} \frac{(1 + \gamma_k^{*-a})\gamma_k^{*-1}}{(1 + \lambda\gamma_k^{*-1})^2} &\leq \sup_{x\geq 0} \frac{(1 + x^{-a})x^{-1}}{(1 + \lambda x^{-1})^2} \\
&\leq \sup_{x\geq 0} \frac{(1 + x^{-a})}{(1 + \lambda x^{-1})^2} + \sup_{x\geq 0} \frac{x^{-1}}{(1 + \lambda x^{-1})^2} \\
&= 4\lambda^{-1} + O\left(\lambda^{-1-a}\right).
\end{aligned}$$

Replacing in 101 it is concluded that there is a positive constant C such that

$$\|\bar{\beta}_{\infty\lambda} - \beta_0\|_a^2 \leq C J(\beta_0) \lambda^{1-a}, \tag{102}$$

and this shows the desired result. Note also that a consequence of this, because the conditions in theorem 4 are satistfied with this inequality, that

$$\|\beta_{\infty\lambda} - \beta_0\|_a^2 = O\left(J(\beta_0)\lambda^{1-a}\right). \tag{103}$$

A.5.2 Convergence rate for the linearized variance.

Note first that $\bar{\beta}_{n\lambda} - \beta_{\infty\lambda} = -G^{-1}_{\infty\lambda}(\beta_{\infty\lambda}) D\ell_{n\lambda}(\beta_{\infty\lambda})$ from definition 14. Also, given that $D\ell_{\infty\lambda}(\beta_{\infty\lambda}) = 0$, it is possible to write $D\ell_{n\lambda}(\beta_{\infty\lambda}) = D\ell_{n\lambda}(\beta_{\infty\lambda}) - D\ell_{\infty\lambda}(\beta_{\infty\lambda}) = D\ell_n(\beta_{\infty\lambda}) - D\ell_\infty(\beta_{\infty\lambda})$. We present explicit expressions for these functional derivatives in order to continue with the analysis. For any $\beta, f \in \mathcal{H}_\alpha$,

$$\begin{aligned} D\ell_n(\beta)f &= \frac{1}{n}\sum_{i\le n}\left[\omega^{(1)}\left(\int_\tau X_i\beta\right) - Y_i\right]\int_\tau X_i f \\ D\ell_\infty(\beta)f &= \mathbb{E}_X\left[\omega^{(1)}\left(\int_\tau X\beta\right) - \omega^{(1)}\left(\int_\tau X\beta_0\right)\right]\int_\tau Xf. \end{aligned}$$

Using the fact that $\ell_\infty(\beta) = \mathbb{E}\ell_n(\beta)$, we have

$$\begin{aligned} \mathbb{E}\left(D\ell_{n\lambda}(\beta_{\infty\lambda})f\right)^2 &= \mathbb{E}\left(D\ell_n(\beta_{\infty\lambda})f - D\ell_\infty(\beta_{\infty\lambda})f\right)^2 \\ &= \operatorname{Var}\left(D\ell_n(\beta_{\infty\lambda})f\right) \\ &\le \mathbb{E}\left(D\ell_n(\beta_{\infty\lambda})f\right)^2 \\ &= \frac{1}{n}\mathbb{E}\left(\left[\omega^{(1)}\left(\int_\tau X\beta_{\infty\lambda}\right) - Y\right]\int_\tau Xf\right)^2. \end{aligned}$$

Note that $\mathbb{E}(Y|X) = \omega^{(1)}\left(\int_\tau X\beta_0\right)$ and $\operatorname{Var}(Y|X) = \omega^{(2)}\left(\int_\tau X\beta_0\right)$, so we can make

$$\begin{aligned} \mathbb{E}\left(D\ell_{n\lambda}(\beta_{\infty\lambda})f\right)^2 &\le \frac{1}{n}\mathbb{E}_X\left(\left[\omega^{(1)}\left(\int_\tau X\beta_{\infty\lambda}\right) - \omega^{(1)}\left(\int_\tau X\beta_0\right)\right]\int_\tau Xf\right)^2 \\ &\quad + \frac{1}{n}\mathbb{E}\left([Y - \mathbb{E}(Y|X)]\int_\tau Xf\right)^2 \\ &\le \frac{1}{n}\left\{\mathbb{E}_X\left[\omega^{(1)}\left(\int_\tau X\beta_{\infty\lambda}\right) - \omega^{(1)}\left(\int_\tau X\beta_{\infty\lambda}\right)\right]^4 \mathbb{E}\left[\int_\tau Xf\right]^4\right\}^{1/2} \\ &\quad + \frac{1}{n}\left\{\mathbb{E}\left[\omega^{(2)}\left(\int_\tau X\beta_0\right)\right]^2 \mathbb{E}\left[\int_\tau Xf\right]^4\right\}^{1/2}, \end{aligned}$$

where we used Cauchy-Schwarz inequality and conditional variance in the last inequality. Using assumptions 1 (ii) and (iii) we conclude that there exists a constant $M > 0$ such that

$$\mathbb{E}\left(D\ell_{n\lambda}(\beta_{\infty\lambda})f\right)^2 \le \frac{M}{n}\mathbb{E}_X\left[\int_\tau Xf\right]^2 = \frac{M}{n}\langle Cf, f\rangle_{\mathcal{L}_2}. \tag{104}$$

Thus substituting this into the definition of the linearized variance we have

$$\begin{aligned} \mathbb{E}\|\bar{\beta}_{n\lambda} - \beta_{\infty\lambda}\|_a^2 &= \mathbb{E}\|G_{\infty\lambda}^{-1}(\beta_{\infty\lambda})D\ell_{n\lambda}(\beta_{\infty\lambda})\|_a^2 \\ &\le C\mathbb{E}\|G_{\infty\lambda}^{-1}(\beta_{\infty\lambda})D\ell_{n\lambda}(\beta_{\infty\lambda})\|_{a^*}^2 \\ &= C\sum_{k\ge 1}(1+\gamma_k^{*-a})(1+\lambda\gamma_k^{*-1})^{-2}\mathbb{E}\left[D\ell_{n\lambda}(\beta_{\infty\lambda})\varphi_k^*\right]^2 \\ &\le \frac{C}{n}\sum_{k\ge 1}(1+\gamma_k^{*-a})(1+\lambda\gamma_k^{*-1})^{-2}\langle C\varphi_k^*, \varphi_k^*\rangle_{\mathcal{L}_2} \\ &= \frac{C}{n}\sum_{k\ge 1}(1+\gamma_k^{*-a})(1+\lambda\gamma_k^{*-1})^{-2}, \end{aligned} \tag{105}$$

where recalling that $\gamma_k^* \asymp k^{-2(r+s)}$ for k large enough,

$$\begin{aligned}
\sum_{k\geq 1}(1+\gamma_k^{*-a})(1+\lambda\gamma_k^{*-1})^{-2} &\leq C\sum_{k\geq 1}(1+k^{2a(r+s)})(1+\lambda k^{2(r+s)})^{-2} \\
&\leq C\int_1^\infty (1+x^{2a(r+s)})(1+\lambda x^{2(r+s)})^{-2} \\
&\asymp \int_1^\infty x^{2a(r+s)}(1+\lambda x^{2(r+s)})^{-2} \\
&\asymp \int_1^\infty \left(1+\lambda x^{\frac{2(r+s)}{2a(r+s)+1}}\right)^{-2} \\
&\asymp \lambda^{-a-\frac{1}{2(r+s)}}, \qquad (106)
\end{aligned}$$

and replacing in 105 we have that $\mathbb{E}\|\bar{\beta}_{n\lambda}-\beta_{\infty\lambda}\|_a^2 \leq n^{-1}\lambda^{-a-\frac{1}{2(r+s)}}$.

A.5.3 Convergence rate for the estimator $\hat{\beta}_{n\lambda}$.

Combining the results from sections A.5.1 and A.5.2 we have that for $\alpha = \frac{1}{2}\left[1-\frac{1}{2(r+s)}\right]$ and $\lambda = o\left(n^{-1}\lambda^{-\alpha-\frac{1}{2(r+s)}}\right)$, for any $0\leq a\leq 1$ and $\epsilon>0$ there exists a finite constant C_ϵ such that

$$\limsup_{n\to\infty}\sup_{F\in\mathcal{F}(s,M),\beta_0\in\mathcal{H}} P\left(\|\hat{\beta}_{n\lambda}-\beta_0\|_a^2 > C_\epsilon\left[\lambda^{1-a}+n^{-1}\lambda^{-a-\frac{1}{2(r+s)}}\right]\right) < \epsilon.$$

Therefore, choosing $\lambda \asymp \frac{-2(r+s)}{2(r+s)+1}$ we have that

$$\limsup_{n\to\infty}\sup_{F\in\mathcal{F}(s,M),\beta_0\in\mathcal{H}} P\left(\|\hat{\beta}_{n\lambda}-\beta_0\|_a^2 > C_\epsilon n^{\frac{-2(1-a)(r+s)}{2(r+s)+1}}\right) < \epsilon,$$

and this proves the theorem. □

A.6 Proof of Lemma 16

It is easy to check that for $\beta_3\in\mathcal{H}_\alpha$,

$$D^3\ell_\infty(\beta_3)uvw = \mathbb{E}_X\left[\left(\int_\tau Xu\right)\left(\int_\tau Xv\right)\left(\int_\tau Xw\right)\cdot\omega^{(3)}\left(\int_\tau X\beta_3\right)\right] \qquad (107)$$

for u, v, $w\in\mathcal{H}_\alpha$. Therefore, by Proposition 3 and making $\beta_0=\beta^*$ in definition 37, if $0\leq a\leq\alpha$,

$$\begin{aligned}
\|G_{\infty\lambda}^{-1}(\beta_0)D^3\ell_\infty(\beta_3)uv\|_a^2 &\leq M\sum_{k\geq 1}\left(1+\gamma_k^{*-\alpha}\right)\left(1+\lambda\gamma^{*-1}\right)^{-2}\langle D^3\ell_\infty(\beta_3)uv, C\varphi_k^*\rangle_{\mathcal{L}_2}^2 \\
&= M\sum_{k\geq 1}\left(1+\gamma_k^{*-\alpha}\right)\left(1+\lambda\gamma^{*-1}\right)^{-2}\left(D^3\ell_\infty(\beta_3)uv\varphi_k^*\right)^2, (108)
\end{aligned}$$

where

$$\begin{aligned} D^3\ell_\infty(\beta_3)uv\varphi_k^* &= \mathbb{E}_X\left[\left(\int_\mathcal{T} Xu\right)\left(\int_\mathcal{T} Xv\right)\left(\int_\mathcal{T} X\varphi_k^*\right)\cdot\omega^{(3)}\left(\int_\mathcal{T} X\beta_3\right)\right] \\ &\leq \left(\mathbb{E}_X\left[\left(\int_\mathcal{T} Xu\right)\left(\int_\mathcal{T} Xv\right)\right]^2\mathbb{E}_X\left[\left(\int_\mathcal{T} X\varphi_k^*\right)\omega^{(3)}\left(\int_\mathcal{T} X\beta_3\right)\right]^2\right)^{1/2} \quad (109) \end{aligned}$$

by Cauchy-Schwarz inequality. By assumptions 1 *(i)-(iii)*, and using Cauchy-Schwarz inequality again in each term, it follows that there exists a constant $M<\infty$ such that

$$\begin{aligned} \mathbb{E}_X\left[\left(\int_\mathcal{T} Xu\right)\left(\int_\mathcal{T} Xv\right)\right]^2 &\leq M\langle Cu,u\rangle_{\mathcal{L}_2}^2\langle Cv,v\rangle_{\mathcal{L}_2}^2 \\ \mathbb{E}_X\left[\left(\int_\mathcal{T} X\varphi_k^*\right)\omega^{(3)}\left(\int_\mathcal{T} X\beta_3\right)\right]^2 &\leq M\langle C\varphi_k^*,\varphi_k^*\rangle_{\mathcal{L}_2}^2. \end{aligned}$$

Replacing in 109 we obtain that $D^3\ell_\infty(\beta_3)uv\varphi_k^* \leq M\|u\|_0\|v\|_0\|\varphi_k^*\|_0$. Now, using the expression derived in 108 and the result obtained in 106, and for $u,v\in\mathcal{H}$ such that $\|u\|_\alpha = \|v\|_\alpha = 1$, it follows that

$$\begin{aligned} K_3(\lambda,a) &= \sup_{\beta_3\in\mathcal{H}_\alpha}\sup_{u,v}\|G_{\infty\lambda}^{-1}(\beta_0)D^3\ell_\infty(\beta_3)uv\|_a \\ &\leq M\|\varphi_k^*\|_0\lambda^{-\frac{1}{2}\left(a+\frac{1}{2(r+s)}\right)}. \end{aligned} \quad (110)$$

Now, from section A.5.1 it is concluded that for $0\leq a\leq\alpha$, $\|\beta_{\infty\lambda}-\beta_0\|_a^2 = O\left(J(\beta_0)\lambda^{1-a}\right)$, and therefore,

$$K_3(\lambda,a)\|\beta_{\infty\lambda}-\beta_0\|_\alpha \leq M\lambda^{\frac{1}{2}\left(1-\alpha-a-\frac{1}{2(r+s)}\right)}.$$

It follows that if $\alpha\leq\frac{1}{2}\left(1-\frac{1}{2(r+s)}\right)$, then $K_3(\lambda,a)\|\beta_{\infty\lambda}-\beta_0\|_\alpha\to 0$ as $\lambda\to 0$, and the theorem is proved. □

A.7 Proof of Lemma 17

The second order Fréchet derivatives for the functional $\ell_{n\lambda}$ and $\ell_{\infty\lambda}$, for $\beta, f, g\in\mathcal{H}_\alpha$, can be written respectively as

$$\begin{aligned} D^2\ell_n(\beta)fg &= \frac{1}{n}\sum_{i\leq n}\left[\omega^{(2)}\left(\int_\mathcal{T} X_i\beta\right)\left(\int_\mathcal{T} X_if\right)\left(\int_\mathcal{T} X_ig\right)\right] \\ D^2\ell_\infty(\beta)fg &= \mathbb{E}_X\left[\omega^{(2)}\left(\int_\mathcal{T} X\beta\right)\left(\int_\mathcal{T} Xf\right)\left(\int_\mathcal{T} Xg\right)\right]. \end{aligned}$$

We will start by decomposing the expression $K_2(\lambda,a)$ in two separate parts. Recall that $G_{\infty\lambda}(\beta) = D^2\ell_{\infty\lambda}(\beta)$, and $D^2\ell_\infty(\beta)fg = \mathbb{E}_XD^2\ell_n(\beta)fg$. From the definition of $K_2(\lambda,a)$, for $0 \leq a \leq \alpha$, we have

$$\begin{aligned} K_2\left(\lambda, a\right) &= \sup_{\beta \in \mathcal{H}_\alpha} \sup_u \| G_{\infty\lambda}^{-1}(\beta_{\infty\lambda}) \left[D^2 \ell_{n\lambda}(\beta) u - D^2 \ell_{\infty\lambda}(\beta) u \right] \|_a \\ &= \sup_{\beta \in \mathcal{H}_\alpha} \sup_u \| G_{\infty\lambda}^{-1}(\beta_{\infty\lambda}) \left[D^2 \ell_{n}(\beta) u - D^2 \ell_{\infty}(\beta) u \right] \|_a . \end{aligned}$$

By Proposition 3, and making $\beta_{\infty\lambda} = \beta^*$, it follows that

$$\begin{aligned} &\mathbb{E} \| G_{\infty\lambda}^{-1}(\beta_{\infty\lambda}) \left[D^2 \ell_{n}(\beta) u - D^2 \ell_{\infty}(\beta) u \right] \|_a^2 \\ &\qquad \leq M \sum_{k \geq 1} \left(1 + \gamma_k^{*-\alpha}\right) \left(1 + \lambda \gamma^{*-1}\right)^{-2} \mathbb{E} \langle D^2 \ell_n(\beta) u - D^2 \ell_\infty(\beta) u, C\varphi_k^* \rangle_{\mathcal{L}_2}^2 \\ &\qquad = M \sum_{k \geq 1} \left(1 + \gamma_k^{*-\alpha}\right) \left(1 + \lambda \gamma^{*-1}\right)^{-2} \mathbb{E} \left(D^2 \ell_n(\beta) u \varphi_k^* - D^2 \ell_\infty(\beta) u \varphi_k^* \right)^2 \\ &\qquad = M \sum_{k \geq 1} \left(1 + \gamma_k^{*-\alpha}\right) \left(1 + \lambda \gamma^{*-1}\right)^{-2} \operatorname{Var} \left(D^2 \ell_n(\beta) u \varphi_k^* \right), \end{aligned}$$

and by assumption 1 (iv), it follows that $\operatorname{Var}\left(D^2 \ell_n(\beta) u \varphi_k^*\right) = O(n^{-1})$. Therefore, using the inequality in 106, we have that

$$K_2(\lambda, a) = O_p\left(n^{-1/2} \lambda^{-\frac{1}{2}\left(a + \frac{1}{2(r+s)}\right)} \right). \tag{111}$$

Now we focuse in the term $K_{3n}(\lambda, a)$, for which we have that

$$\begin{aligned} K_{3n}(\lambda, a) &= \sup_{\beta \in \mathcal{H}_\alpha} \sup_{u,v} \| G_{\infty\lambda}^{-1}(\beta_{\infty\lambda}) D^3 \ell_{n\lambda}(\beta) uv \|_a \\ &\leq K_3(\lambda, a) + \sup_{\beta \in \mathcal{H}_\alpha} \sup_{u,v} \| G_{\infty\lambda}^{-1}(\beta_{\infty\lambda}) \left[D^3 \ell_{n\lambda}(\beta) uv - D^3 \ell_{\infty\lambda}(\beta) uv \right] \|_a, \end{aligned}$$

for $u, v \in \mathcal{H}_\alpha$ such that $\|u\|_\alpha = \|v\|_\alpha = 1$. To bound the second term, note that using $\beta^* = \beta_{\infty\lambda}$ it is possible to write

$$\begin{aligned} &\mathbb{E} \| G_{\infty\lambda}^{-1}(\beta_{\infty\lambda}) \left[D^3 \ell_{n\lambda}(\beta) uv - D^3 \ell_{\infty\lambda}(\beta) uv \right] \|_a^2 \\ &\qquad \leq M \sum_{k \geq 1} \left(1 + \gamma_k^{*-\alpha}\right) \left(1 + \lambda \gamma^{*-1}\right)^{-2} \mathbb{E} \langle D^2 \ell_n(\beta) uv - D^3 \ell_\infty(\beta) uv, C\varphi_k^* \rangle_{\mathcal{L}_2}^2 \\ &\qquad = M \sum_{k \geq 1} \left(1 + \gamma_k^{*-\alpha}\right) \left(1 + \lambda \gamma^{*-1}\right)^{-2} \operatorname{Var} \left(D^3 \ell_n(\beta) uv \varphi_k^* \right). \end{aligned}$$

By assumption 1 (iv) we know that $\operatorname{Var}\left(D^3 \ell_n(\beta) uv \varphi_k^*\right) = O(n^-1)$, and together with the result in 110 it is derived that

$$K_{3n}(\lambda, a) = O\left(\lambda^{-\frac{1}{2}\left(a + \frac{1}{2(r+s)}\right)} \right) + O_p\left(n^{-1/2} \lambda^{-\frac{1}{2}\left(a + \frac{1}{2(r+s)}\right)} \right).$$

This result, together with 111 and the bound for $\|\bar{\beta}_{n\lambda} - \beta_{\infty\lambda}\|_a$ derived in section A.5.2, it follows that

$$\begin{aligned}\big[K_2(\lambda, a) + K_{3n}(\lambda, a)\|\bar{\beta}_{n\lambda} - \beta_{\infty\lambda}\|_\alpha\big] \\ = O_p\left(n^{-1/2}\lambda^{-\frac{1}{2}\left(a+\frac{1}{2(r+s)}\right)}\right) + O_p\left[n^{-1/2}\lambda^{-\left(a+\frac{1}{2(r+s)}\right)} + n^{-1}\lambda^{-\left(a+\frac{1}{2(r+s)}\right)}\right].\end{aligned}$$

Therefore, if there is a decreasing sequence $\lambda_n \to 0$ such that $n^{-1}\lambda^{-\left(a+\frac{1}{2(r+s)}\right)} \to 0$ as $n \to \infty$, then

$$\big[K_2(\lambda, a) + K_{3n}(\lambda, a)\|\bar{\beta}_{n\lambda} - \beta_{\infty\lambda}\|_\alpha\big] = o_p(1). \tag{112}$$

REFERENCES

[1] Aronszajn, N. (1950). Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, **68** (3), 337-404.

[2] Bartlett, P. L., Jordan, M. L. and McAuliffe, P. L. (2006). Convexity, classification, and risk bounds. *Journal of American Statistical Association*, **107**, 1201-1216.

[3] Bartlett, P. L., Wegkamp, M.H. (2008). Classification with a Reject Option using a Hinge Loss. *Journal of Machine Learning Research*, **9**, 1823-1840.

[4] Bickel, P. and Li, B. (2006). Regularization in Statistics. *TEST*, **15**, 271-344.

[5] Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statististics*, **101**, 138156.

[6] Buhmann, M. D. (2003). *Radial Basis Functions: Theory and Implementations.* Cambridge: University Press.

[7] Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics*, **34**, 2159-2179.

[8] Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistical Sinica*, **13**, 571-591.

[9] Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, **92**, 24-41.

[10] Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems*, **24**, 1-19.

[11] Cavalier, L. and Tsybakov, A. (2002), Sharp adaptation for inverse problems with random noise, *Probability Theory and Related Fields* **123**, 323-354.

[12] Chalmond, B. (2008). *Modeling and Inverse Problems in Image Analysis.* New York: Springer.

[13] Cox, D.D. (1988). Approximation of method of regularization estimators. *Annals of Statistics*, **18**, 694-712.

[14] Cox, D.D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics*, **18**, 1676-1695.

[15] Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of American Mathematical Society*, **39**, 1-49.

[16] Debnath, L. and Mikusiński, P. (2005). *Introduction to Hilbert Spaces with Applications*, 3rd ed. Academic Press.

[17] Dou, W., Pollard, D. and Zhou, H.H. (2012). Functional regression for general exponential families. *Annals of Statistics*, **40**, 2421-2451.

[18] Evgeniou, T., Pontil, M., and Poggio,T. (2000). Statistical Learning Theory: A Primer. *International Journal of Computer Vision*, **38** (1), 9-13.

[19] Ferraty, F. and Romain, Y. (2010). *The Oxford handbook of functional data analysis.* Oxford Handbooks in Mathematics, OUP Oxford.

[20] Girosi, F., Jones, M. and Poggio, T. (1993). Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. *Artificial Intelligence memo 1430*, MIT, Artificial Intelligence Laboratory.

[21] Golubev, G. and Nussbaum, M. (1998). Asymptotic equivalence of spectral density and regression estimation. Technical report, Weierstrass Institute for Applied Analysis and Stochastics, Berlin.

[22] Grama, I. and Nussbaum, M. (1997). Asymptotic equivalence for nonparametric generalized linear models. Technical report, Weierstrass Institute for Applied Analysis and Stochastics, Berlin.

[23] Hall, P. and Horowitz, J.L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics*, **35**, 70-91.

[24] James, G. (2002). Generalized linear models with functional predictors. *Journal of Royal Statistical Society Ser. B*, **64**, 411-432.

[25] Kaipo, J. and Somersalo, E. (2004). *Statistical and Computational Inverse Problems.* New York: Springer.

[26] Lesko, L.J. (2007). Personalized medicine: Elusive dream or imminent reality?. *Clincal Pharmacology and Therapeutics*, **81**, 807816.

[27] Lin, Y. and Brown, L. (2004). Statistical properties of the method of regularization with periodic Gaussian reproducing kernel, *Annals of Statistics*, **32**, 1723-1743.

[28] Lin, Y. (2004). A note on margin-based loss functions in classification. *Statistics & Probability Letters*, **68**, 73-82.

[29] Lin, Y. and Yuan, M. (2006). Convergence rates of compactly supported radial basis function regularization. *Statistica Sinica*, **16**, 425-439.

[30] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Monographs on Statistics and Applied Probability **37**. Chapman and Hall.

[31] Moodie, E.M. and Richardson, T.S. (2010). Estimating Optimal Dynamic Regimes: Correcting Bias under the Null. *Scandinavian Journal of Statistics*, **37**, 126-146.

[32] Müller, H.G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, **33**, 774-805.

[33] Murphy, S.A., Van Der Laan, M.J., Robins, J.M. and and CPPRG (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, **96**, 1410-1423.

[34] Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Annals of Statististics*, **24**, 2399-2430.

[35] Piquette-Miller, P. and Grant, D.M. (2007). The art and science of personalized medicine. *Clincal Pharmacology and Therapeutics*, **81**, 311315.

[36] Qian, M. and Murphy, S.A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statististics*, **39**, 1180-1210.

[37] Ramm, A. (2009). *Inverse Problems: Mathematical and Analytical Techniques with Applications to Engineering.* New York: Springer.

[38] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York.

[39] Ritter, K., Wasilkowski, G. and Woźniakowski, H. (1995). Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions. *Annals of Applied Probability*, **5**, 518-540.

[40] Robins, J., Orellana, L. and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, **27**, 46784721.

[41] Sacks, J. and Ylvisaker, D. (1970). Designs for regression problems with correlated errors III. *Annals of Mathematical Statistics*, **41**, 2057-2074.

[42] Smola, A., Schlkopf, B. and Mülller, K.R. (1998). The connection between regularization operators and support vector kernels. *Neural Networks*, **11**, 637-649.

[43] Tartar, L. (2000). *An introduction to Sobolev Spaces and Interpolation Spaces.* Lecture Notes of the Unione Matematica Italiana **3**. Springer, 2007.

[44] Triebel, H. (1978). *Interpolation theory, Function Spaces, Differential Operators.* North-Holland, Amsterdam.

[45] Tsybakov, A.B. (2004). Optimal Aggregation of Classifiers in Statistical Learning. *Annals of Statististics*, **32**, 135-166.

[46] Valencia, C. and Yuan, M. (2013). Radial basis function regularization for linear inverse problems with random noise. *Journal of Multivariate Analysis*, **116**, 92-108.

[47] Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM.

[48] Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Schoelkopf, C. Burges & A. Smola, eds, *Advances in Kernel Methods Support Vector Learning.* Cambridge: MIT Press, 69-88

[49] Weinberger, H.F. (1974). *Variational Methods for Eigenvalue Approximation.* SIAM, Philadelphia.

[50] Wendland, H. (1998). Error estimates for interpolation by compactly supported radial basis functions of minimal degree. *Journal of Approximation Theory*, **93**, 258-272.

[51] Yuan, M. and Cai, T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Annals of Statistics*, **38**, 3412-3444.

[52] Yuan, M. and Cai, T. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of American Statistical Association*, **107**, 1201-1216.

[53] Yuan, M. and Wegkamp, M.H. (2010). Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, **11**, 111-130.

[54] Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning*, **5**, 1225-1251.

[55] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**, 56-134.

[56] Zhao, Y.Q., Zeng, D., Rush A.J., Kosorok, M.R. (2012). Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, **107**, 1106-1118.