LEARNING WITH CENTERED REPRODUCING KERNELS

CHENDI WANG

M.Phil

The Hong Kong Polytechnic University

2018

The Hong Kong Polytechnic University

Department of Applied Mathematics

Learning with Centered Reproducing Kernels

Chendi Wang

A thesis submitted in partial fulfilment of the requirements

for the degree of Master of Philosophy

April 2018

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

________________________(Signed)

Chendi Wang (Name of student)

Abstract

In the past twenty years, reproducing kernels and the kernel-based learning algorithms have been widely and successfully applied to many areas of scientific research and industry, and are extensively studied. Many of these algorithms take the form of an optimization problem. Typically, the objective function consists of a fidelity term for fitting the observations, and a regularization term for preventing over-fitting. Examples include the support vector machines for classification, and the regularized least squares for regression. However, in many regression problems, the constant component should be treated differently in the regression function, and the existing kernel methods are not perfect tools to model this difference. Examples include score-based ranking function regression. In this thesis, we study a class of Centered Reproducing Kernels (CRKs), which separate the constant component from the reproducing kernel Hilbert spaces. We provide the non-asymptotic convergence analysis of the empirical CRK-based regularized least squares.

Acknowledgements

First and foremost, I especially want to express my deep thanks to my two supervisors. I would like to thank my first supervisor Prof. Xiaojun Chen, who helps me established the fundation in not only academic life but also daily life. Furthermore, I want to give my thanks to my second supervisor Dr. Xin Guo, who enlightens me in learning theory.

I finished most of my work with the resources offered by the Hong Kong Polytechnic University and the department of Applied Mathematics (AMA). I wish to thank to the PolyU and AMA for their support.

Finally, I thank my parents who gave my life and brought me up and who support me not only in economy, but also in mentality. Their selfless love will inspire me lifetime long.

Contents

Certificate of Originality iii

Abstract v

Acknowledgements vii

1 Introduction 1

1.1 Learning Problems 1

1.1.1 Regression Problems 2

1.1.2 Classification Problems 4

1.1.3 Ranking Problems 5

1.1.4 Other Learning Problems 6

1.2 Kernel-based learning algorithms 6

1.2.1 Reproducing Kernel Hilbert Spaces 6

1.2.2 Learning Algorithms for Regression Problems 9

1.2.3 Learning Algorithms for Classification Problems 12

1.2.4 Online Learning Algorithms 15

1.2.5 Algorithms for Ranking and Pairwise Learning 16

1.2.6 Deep Neural Networks 17

1.3 Centered Reproducing Kernels 17

1.3.1 Centered Reproducing Kernels 18

1.3.2 The Motivation of CRKs 19

2 Regularized Least Squares with Centered Reproducing Kernels 21

2.1 Properties of CRKs 21

2.2 CRK-based Modified KRR Learning Algorithms 25

2.3 Error Analysis for the CRK-based Learning Algorithm 27

2.4 A Simulation under Gaussian RBF kernel 43

3 Distributed KRR 45

3.1 The Convergence with High Probability of DKRR 45

3.2 Error Analysis for DKRR 47

3.3 The Semi-supervised Learning For DKRR 55

3.4 The Almost Sure Convergence of DKRR 59

Bibliography 61

Chapter 1

Introduction

1.1 Learning Problems

In the era of Big data, one important problem of data analysis is what information we can learn from massive data through automatic algorithms. The learning process through these algorithms is called machine learning. For example, a modern definition of machine learning is proposed by Mitchell [44]:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

In mathematics, we identify a function in a class T of functions based on a data set E. The difference between the identified function and the unknown target function is measured by a distance P, such as the L^2 norm based on a probability measure.

Let X denote an **input** space, and let Y be an **output** space. Usually, X is a compact metric space such as a set of genes or a domain in $\mathbb{R}^n$ and Y is a subset of the real line $\mathbb{R}$. Furthermore, we assume that there is an **unknown joint probability distribution** ρ on $X \times Y$ which can be decomposed as a conditional distribution $\rho(y|x)$ on Y for almost every x and a marginal distribution ρ_X on X. The labeled **training** sample $D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ with the size $|D| = N$ is assumed to be drawn independently from the distribution ρ. These $y_i, i = 1, 2, ..., N$

are referred to as **labels** or **supervised information**. The aim of a supervised (or semi-supervised) learning problem is to find a functional relation $f : X \to Y$ between the input and the output spaces that has the generalization power, i.e., the power to predict the label $f(x)$ of a new instance x that may not belong to the training set. The generalization power is pointwisely measured by some problem-specific loss function

$$\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+,$$

and the over-all generalization error is defined by

$$\mathcal{E}_\phi(f) = \int_{X\times Y} \phi(f(x), y) d\rho(x, y).$$

This quantitative description of the generalization power yields many learning algorithms under the empirical risk minimization scheme, which we will review below.

Based on different types of training data, learning problems can be separated into the following three parts.

1. **Supervised Learning:** the training data consist of only labeled data $D = \{(x_i, y_i)\}_{i=1}^N$.

2. **Unsupervised Learning:** the training data consist of only unlabeled data $D(x)$. For example, the clustering problems, and the association analysis problems belong to this category.

3. **Semi-supervised Learning:** the training data consist of both labeled and unlabeled data. In particular, the learning scenarios that have very cheap unlabeled data but the labels are expensive.

1.1.1 Regression Problems

The classical **least squares regression** corresponds to the **least squares loss**

$$\phi(f(x), y) = (y - f(x))^2.$$

Consider $f \in L^2_{\rho_X}(X)$ and let

$$\|f\|_\rho^2 = \int_X f(x)^2 d\rho_X(x)$$

be the square of $L^2_{\rho_X}$ norm. Then the least squares error

$$\mathcal{E}(f) = \int_{X \times Y} (y - f(x))^2 d\rho$$

is minimized by the **regression function**

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

In fact, for any $f \in L^2_{\rho_X}(X)$, a straightforward calculation shows that

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

It is worthy of mentioning that the regression function is just the conditional expection of a random variable on Y with distribution $\rho(y|x)$. Compared with learning the mean, another regression problem is called **quantile regression**.

Quantile regression aims at learning the τ-quantile, in particular, the **median regression** is the 0.5-quantile regression. Precisely, for a real number $\tau \in (0, 1)$, the conditional τ-quantile of $\rho(y|x)$ is the number $t \in Y$ such that

$$\rho(y < t|x) \geqslant \tau \text{ and } \rho(y \geqslant t|x) \geqslant 1 - \tau.$$

The learning algorithms of τ-quantile (e.g. [36, 59, 49, 56, 68]) are usually based on the **pinball loss**

$$\phi_\tau(y, f(x)) = \begin{cases} (y - f(x))\tau, & y \geqslant f(x) \\ (f(x) - y)(1 - \tau), & \text{otherwise.} \end{cases}$$

In particular, for $\tau = 1/2$, the loss function is the absolute loss

$$\phi_{\text{absolute}}(y, f(x)) = \frac{1}{2} |y - f(x)| .$$

Another property of quantile regression is the **robustness** since the outlier doesn't affect the median [35].

Combining the least squares regression with the median regression, one obtains the so-called **Huber's loss**

$$\phi_{\text{Huber}}(y, f(x)) = \begin{cases} 1/2(y - f(x))^2, & |y - f(x)| \leqslant k \\ k|y - f(x)| - 1/2k^2, & \text{otherwise.} \end{cases}$$

The motivation of Huber's loss comes from a truncation to the outlier by k as a result the target function is robust to the outlier. Moreover, when k tends to infinity, the Huber's loss is close to the least squares loss which leads to learning the conditional mean [35].

1.1.2 Classification Problems

When $Y = \{\pm 1\}$, the learning problem is a **binary-classification problem** whose target function (or **classifier**) is the sign of the function f that minimizes the **misclassification risk**

$$\mathcal{R}(f) := \rho(y \neq f(x)) = \int_X \rho(y \neq f(x)|x) d\rho_X.$$

To get the explicit minimizer of the risk, one (e.g. Proposition 9.3 [24]) decomposes $\mathcal{R}(f)$ as

$$\mathcal{R}(f) = \frac{1}{2}\rho_X\left(K_\rho\right) + \int_{X/K_\rho} \rho(y \neq f(x)|x) d\rho_X$$

with $K_\rho = \{x \in X : \rho\left(y = 1|x\right) = \rho\left(y = -1|x\right)\}$. Thus the optimizer of the above functional is the **Bayes Rule**

$$f_c(x) = \begin{cases} 1, & \rho\left(y = 1|x\right) > \rho\left(y = -1|x\right), \\ -1, & \rho\left(y = -1|x\right) \geqslant \rho\left(y = 1|x\right). \end{cases}$$

When Y contains k points (k classes) for $k \geqslant 3$, the classification problem is called multi-class classification problem. The analysis of multi-class classification are studied in, for example, [67, 13].

1.1.3 Ranking Problems

Consider $(x, y), (x', y') \sim (X \times Y, \rho)$ which are independent of each other. x is regarded a better instance than x' when $y > y'$. **Ranking problems** aim at finding a suitable rule to predict the rank. In this section, we introduce some basic settings in bipartite ranking, i.e., $Y = \{\pm 1\}$.

For normalization, let (e.g. [16])

$$z := \frac{y - y'}{2}.$$

As a result, x is said to be better than x' if $z > 0$. The target function for a bipartite ranking problem is defined by the minimizer of the **ranking risk**

$$\mathcal{L}(r) = \rho\left(z \cdot r(x, x') < 0\right),$$

which is the probability of ranking mistake. Also in [16], the decomposition of $\mathcal{L}(r)$ can be presented as

$$\mathcal{L}(r) = \int_{X \times X} \left(\mathbf{1}_{[r(x,x')=1]}\rho_-(x, x') + \mathbf{1}_{[r(x,x')=-1]}\rho_+(x, x')\right) d\rho_X(x) d\rho_X(x'), \tag{1.1}$$

where

$$\rho_+(x, x') = \rho\left(z > 0 | x, x'\right) \text{ and } \rho_-(x, x') = \rho\left(z < 0 | x, x'\right).$$

As a consequence of the decomposition above, the minimizer of the ranking risk is

$$r^*(x, x') = 2 \times \mathbf{1}_{[r_+(x,x') \geqslant r_-(x,x')]} - 1$$

and the corresponding ranking risk is

$$\mathcal{L}(r^*) = \int_{X \times X} \min\{\rho_-(x, x'), \rho_+(x, x')\} d\rho_X(x) d\rho_X(x').$$

In the case where there is an **optimal scoring function** $s^* : X \to \mathbb{R}$ such that

$$r^*(x, x') = 1 \text{ if and only if } s^*(x) \geqslant s^*(x'),$$

the ranking problem is equivalent to a **scoring** problem which aims at learning the optimal scoring function s^*.

1.1.4 Other Learning Problems

There are also some other fields of learning theory such as

1. dimension reduction: reduce the dimenssion of the data points $x_1, ..., x_N$ (e.g. [65]), either linearly of nonlinearly, as an approach to reduce computation load, collinearity, or noise of the data.

2. association analysis (e.g. [60]), which tries to find interesting association rules hidden in large data sets.

1.2 Kernel-based learning algorithms

Kernel methods have been extensively studied in learning theory literature (e.g. [23, 62]). Kernel methods was used in support vector machines by Vapnik et al. in [9, 19]. The implementation of RKHS in regression problems through integral operators was studied in, for example, [55]. In this section, we will review the framework of kernel-based learning algorithms.

1.2.1 Reproducing Kernel Hilbert Spaces

In this thesis, the basic tool for analyzing algorithms is reproducing kernel Hilbert spaces (RKHS) [4, 62].

Let

$$K : X \times X \to \mathbb{R}$$

be a bivariate function on X. We say K is **symmetric** if

$$K(s,t) = K(t,s), \qquad \text{for all } t, s \in X.$$

Moreover, if K satisfies that

$$\sum_{i,j=1}^{N} c_i c_j K(x_i, x_j) \geqslant 0$$

for any finite set $\{x_i\}_{i=1}^N \subseteq X$ and any coefficients $c_i \in \mathbb{R}, i = 1, 2, ..., N$, then K is called a positive semi-definite kernel. A **Mercer kernel** on X is a positive semi-definite kernel on X which is continuous. Since we always assume X to be a compact set, for a Mercer kernel K, we have

$$\kappa := \sup_{x \in X} \sqrt{K(x,x)} < +\infty.$$

For any $x \in X$, we define

$$K_x(t) = K(x,t), \qquad t \in X.$$

The reproducing kernel Hilbert space corresponding to the kernel function K is defined by

$$(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K) = \overline{\text{span}\,\{K_x, x \in X\}},$$

where the completion is done with respect to the inner product $\langle \cdot, \cdot \rangle_K$ defined on span $\{K_x, x \in X\}$ that satisfies

$$\langle K_s, K_t \rangle_K = K(s,t).$$

The word “reproducing” comes from the **reproducing property**

$$\langle f, K_x \rangle_K = f(x), \qquad \text{for all } f \text{ in } \mathcal{H}_K \text{ and } x \text{ in } X,$$

which implies that

$$\|f\|_\infty = \operatorname{ess}\sup_{x \in X} |f(x)| \leqslant \sup_{x \in X} \|f\|_K \|K_x\|_K = \|f\|_K \sup_{x \in X} \sqrt{K(x,x)} = \kappa \|f\|_K. \qquad (1.2)$$

Define the **integral operator**

$$\begin{aligned} L_K : &L^2_{\rho_X}(X) \to L^2_{\rho_X}(X) \\ f &\mapsto \int_X f(x) K_x d\rho_X(x). \end{aligned}$$

The integral operator L_K is a compact, symmetric, positive semi-definite, and Hilbert-Schmidt opertor [55]. Thus we can write its **eigen-system** as $\{\lambda_i, \phi_i\}_{i=1}^{+\infty}$ with non-negative **eigenvalues** $\lambda_1 \geqslant \lambda_2 \geqslant ...$ and **eigenfunctions** $\{\phi_i\}_{i=1}^{+\infty}$ normalized in

$L^2_{\rho_X}(X)$. Based on the integral operator, we have the following famous **Mercer's Theorem** [37].

Theorem (Mercer). *Let X be a compact metric measure space with finite measure ρ_X and let K be a Mercer kernel. Moreover, let $\{\lambda_i, \phi_i\}_{i=1}^{+\infty}$ denote the eigen-system of the integral operator L_K with $\|\phi_i\|_\rho = 1$, for all $i = 1, 2, ...$ Then*

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(s)\phi_i(t), \qquad \text{for all } \; s, t \in X. \tag{1.3}$$

Here the series converges absolutely and uniformly.

The RKHS was generalized to the reproducing kernel Banach space (RKBS) B_K in [74] by regarding K_x as a continuous linear functional on its dual space B_K^* such that

$$K_x(f) = f(x), \qquad \text{for all } f \text{ in } B_K \text{ and } x \text{ in } X.$$

In this chapter, we introduce the regularized learning scheme in learning theory based on the RKHS. With the RKHS, the regularized learning algorithms have the form

$$f^D_{\phi,\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{i=1}^{N} \phi(y_i, f(x_i)) \right\} + \lambda \Omega(f) \tag{1.4}$$

for a given loss function ϕ and a regularization (or penality) term $\Omega(f)$ with functional

$$\Omega : \mathcal{H}_K \to \mathbb{R}.$$

When one takes $\Omega(f) := \|f\|_K^2$, the well-known representer theorem [51, 62] guarantees that one can always find a solution to (1.4) with the form

$$f^D_{\phi,\lambda} = \sum_{i=1}^{N} c_i K_{x_i}. \tag{1.5}$$

1.2.2 Learning Algorithms for Regression Problems

In the previous section, we introduced two kinds of regression problems: the least squares regression problems and the quantile regression problems.

For the least squares regression, the regularized least squares (RLS) learning algorithm

$$f_\lambda^D := \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{N}\sum_{i=1}^N (f(x_i)-y_i)^2+\lambda\Omega(f)\right\}$$

has been extensively studied. The RLS with penalty term $\Omega(f)=\|f\|_K^2$ is also named as kernel ridge regression (KRR). The convergence of the output function generated by KRR are studied in [46, 23, 55, 54, 58, 10, 57], with consistency and robustness [14]. Moreover, the empirical feature based RLS (regularized kernel priciple component analysis (RKPCA)) [7, 8, 81, 80] is a powerful tool for solving not only KRR, but also RLS with l_1 penalty [30] and folded concave penalty [28].

Define the **empirical integral operator**

$$\begin{aligned} L_K^{D(x)} : \mathcal{H}_K &\to \mathcal{H}_K \\ f &\mapsto \frac{1}{N}\sum_{i=1}^N f(x_i)K_{x_i}. \end{aligned} \tag{1.6}$$

Since $L_K^{D(x)}$ is also a compact positive semi-definite Hilbert-Schmidt operator, we denote $\{\lambda_i^{D(x)},\phi_i^{D(x)}\}_i$ its eigensystem normalized in $\mathcal{H}_K$. These eigenfunctions $\left\{\phi_i^{D(x)}\right\}_i$ are called **empirical features**. The concentration results of $\lambda_i^{D(x)}$ and $L_K^{D(x)}$ to λ_i and L_K, respectively, are well studied in [32, 33].

The output function of the empirical feature based learning algorithm has the form

$$f_{\omega,\lambda}^D = \sum_{i=1}^N c_i^{D,\omega}\phi_i^{D(x)},$$

where $c^{D,\omega} = (c_1^{D,\omega}, ..., c_N^{D,\omega})$, and

$$c^{D,\omega} = \arg\min_{c\in\mathbb{R}^N} \left\{ \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} c_j \phi_j^{D(x)}(x_i) - y_i\right)^2 + \lambda\sum_{j=1}^{N}\omega(|c_j|) \right\}$$

with $\omega : \mathbb{R} \to \mathbb{R}$. In [30], the RLS with l_1 regularization term $\omega(|c_i|) = |c_i|$ is analyzed while the convergence of $f_{\omega,\lambda}^D$ with folded-concave ω was studied in [28].

For the KRR

$$f_\lambda^D = \arg\min \left\{ \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2 + \lambda\|f\|_K^2 \right\}, \tag{1.7}$$

by taking the gradient with respect to f in $\mathcal{H}_K$ and letting the gradient vanish, we have the explicit solution given by

$$f_\lambda^D = (L_K^{D(x)} + \lambda I)^{-1}\frac{1}{N}\sum_{i=1}^{N} y_i K_{x_i}. \tag{1.8}$$

By the representer theorem (1.5), f_λ^D has the form

$$f_\lambda^D = \sum_{i=1}^{N} c_i K_{x_i}. \tag{1.9}$$

Substituting (1.9) into (1.7), we obtain

$$\begin{aligned} c = (c_1, ..., c_N) &= \arg\min_{c\in\mathbb{R}^{\mathbb{N}}} \left\{ \frac{1}{N}\|K_{[\mathbf{x}]}c - \mathbf{y}\|_2^2 + \lambda c^T K_{[\mathbf{x}]}c \right\} \\ &= (N\lambda I + K_{[\mathbf{x}]})^{-1}\mathbf{y}, \end{aligned}$$

with

$$\mathbf{y} = (y_1, ..., y_N)$$

and

$$K_{[\mathbf{x}]} = (K(x_i, x_j))_{N\times N}.$$

The convergence of f_λ^D to the regression function f_ρ are widely studied (e.g. [55]). The convergence analysis of KRR is usually under the assumptions on

1. the regularity of f_ρ (e.g. [55]);

2. the covering numbers of the unit ball in $\mathcal{H}_K$ (e.g. [79]);

3. the decay of the eigenvalues λ_i of the integral operator L_K ([57]);

4. the effective dimension of L_K (e.g. [75]).

In our error analysis, we will use the assmputions on the regularity of f_ρ, that is, $f_\rho = L_K^r h_\rho$ for some $h_\rho \in L^2_{\rho_X}(X)$ and $r > 0$. Moreover, we also assume that the effective dimension

$$\mathcal{N}_{L_K}(\lambda) := \mathrm{Tr}((L_K + \lambda I)^{-1} L_K) = O\left(\lambda^{-s}\right), \qquad \text{for some } s > 0. \tag{1.10}$$

In [10, 57], the optimal convergence rate for KRR in the sense of minimax under the $L^2_{\rho_X}(X)$ norm was obtained

$$\|f_\lambda^D - f_\rho\|_\rho^2 = O\left(N^{\frac{2r}{2r+s}}\right).$$

To reduce the memory requirement and computing time for analyzing big data, distributed learning algorithms have been widely used. A distributed learning algorithm usually consists of the following three steps:

1. partitioning the data set D into subsets $D_1, ..., D_m$;

2. implementing a learning algorithm to the data subset on each computing node to produce an individual predicted function;

3. synthesizing a global output by, for example, averaging the individual outputs.

For distributed KRR, on each local machine, we use the RLS learning scheme

$$f_\lambda^{D_j} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{|D_j|} \sum_{(x,y) \in D_j} (f(x) - y)^2 + \lambda \|f\|_K^2 \right\}$$

to get a local output function. Then we approximate the regression function f_ρ by averaging the local output functions as

$$\tilde{f}_\lambda^D = \sum_{i=1}^{m} \frac{|D_i|}{|D|} f_\lambda^{D_j}.$$

The convergence of $\tilde{f}_\lambda^D$ to the regression function f_ρ was first studied in [78]. [40] obtained the minimax optimal rate for distributed KRR. Furthermore, in [11], unlabeled data (semi-supervised) was used to loosen the restriction of the maximum possible concurrent computing nodes.

For the quantile regression, the convergence of the output function of the regularized learning scheme with the form

$$f_{\phi_\tau,\lambda}^D = \arg\min_{f\in\mathcal{H}_K} \frac{1}{N}\sum_{i=1}^{N} \phi_\tau(y_i - f(x_i)) + \lambda\|f\|_K^2 \tag{1.11}$$

is studied in [36, 59, 49, 56, 68, 15, 34].

1.2.3 Learning Algorithms for Classification Problems

For classification problems, the loss function can be rewritten as a univariate function by $\phi(y, f(x)) = \phi_C(yf(x))$. To approximate the Bayes classifier, in [6, 77], the convex analysis techniques were applied and several loss functions for binary-classification problems were studied, including

- Least squares loss: $\phi_{\mathrm{ls}}(t) = (1-t)^2$.
- Modified least squares loss: $\phi_{\mathrm{mls}}(t) = \max(1-t, 0)^2$.
- Hinge loss: $\phi_{\mathrm{h}}(t) = \max(1-t, 0)$.
- Exponential loss: $\phi_{\exp}(t) = \exp(-t)$.
- Logistic loss: $\phi_{\log}(t) = \log(1+\exp(-t))$.

In particular, the hinge loss ϕ_{h} was employed in the famous supported vector machines (SVM), introduced by Vapnik and his collaborators [9, 19]. The aim is to separate two classes $\mathcal{C}_{\mathrm{I}} := \{i : y_i = 1\}$ and $\mathcal{C}_{\mathrm{II}} := \{i : y_i = -1\}$ of a data set $\{(x_i, y_i)\}_{i=1}^N$ for $X = \mathbb{R}^n$ by a hyperplane $H_w^b := \{x : w \cdot x - b = 0, \|w\|_2 = 1\}$, i.e.,

$$\begin{cases} w \cdot x_i > b, & i \in \mathcal{C}_{\mathrm{I}}, \\ w \cdot x_i < b, & i \in \mathcal{C}_{\mathrm{II}}. \end{cases}$$

Generally, we say that this two classes are **separable** if there is a measurable f such that

$$\begin{cases} f(x_i) > 0, & i \in \mathcal{C}_{\mathrm{I}}, \\ f(x_i) < 0, & i \in \mathcal{C}_{\mathrm{II}}. \end{cases}$$

Moreover, the hyperplane H_w^b is called **separating hyperplane**. The solution of the linear case was obtained in [61]. Specifically, define the **margin** $\Delta(w)$ as the distance of two classes to the hyperplane H_w^b, that is,

$$\Delta(w) = \frac{1}{2}\left\{\min_{i \in \mathcal{C}_{\mathrm{I}}} w \cdot x_i - \max_{i \in \mathcal{C}_{\mathrm{II}}} w \cdot x_i\right\}.$$

Let $\tilde{w}$ be the solution of the following minimization problem

$$\begin{aligned} &\min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \|w\|_2^2 \\ &\text{s.t. } y_i(w \cdot x_i - b) \geqslant 1, i = 1, 2, ..., N. \end{aligned} \tag{1.12}$$

Then the separating hyperplane is H_{w*}^{b*} with

$$\begin{aligned} w^* &= \frac{\tilde{w}}{\|\tilde{w}\|_2}, \\ b^* &= b(w^*) = \frac{1}{2}\left\{\min_{i \in \mathcal{C}_{\mathrm{I}}} w^* \cdot x_i + \max_{i \in \mathcal{C}_{\mathrm{II}}} w^* \cdot x_i\right\}. \end{aligned}$$

In this separable case, the margin $\Delta(w)$ is called **hard margin**.

For the non-separable case, the minimization problem could be modified by slack variables $\xi \in \mathbb{R}^N$ by

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^N} \|w\|_2^2 + \frac{1}{\lambda N}\sum_{i=1}^N \xi_i$$

$$\text{s.t. } y_i(w \cdot x_i - b) \geqslant 1 - \xi_i, \qquad i = 1, 2, ..., N,$$

$$\xi_i \geqslant 0, \qquad \text{for all } i = 1, 2, ..., N,$$

whose solution is the same as

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}} \frac{1}{N}\sum_{i=1}^N \phi_{\mathrm{h}}(y_i(w \cdot x_i - b)) + \lambda\|w\|_2^2.$$

Similarly, the regularized SVM based on RKHS is defined by

$$\min_{f\in\mathcal{H}_K, b\in\mathbb{R}} \frac{1}{N}\sum_{i=1}^N \phi_{\mathrm{h}}(y_i(f(x_i) - b)) + \lambda\|f\|_K^2. \tag{1.13}$$

For non-linear case, we say that ρ is **strictly separable** by $\mathcal{H}_K$ with **margin** $\Delta > 0$ if there is some f_{sp} in $\mathcal{H}_K$ such that

$$\|f_{\mathrm{sp}}\|_K = 1$$

and

$$yf_{\mathrm{sp}}(x) \geqslant \Delta \text{ almost surely.}$$

The weakly separable classification problems was considered in [12]. Precisely, ρ is said to be **weakly separated** by $\mathcal{H}_K$ if there is an $f_{\mathrm{sp}} \in \mathcal{H}_K, \|f_{\mathrm{sp}}\|_K = 1$ such that $yf_{\mathrm{sp}}(x) > 0$ a.s.ρ. Moreover, we say that it has **separation triple** (θ, Δ, C_ρ) for $0 < \theta \leqslant +\infty$ and $0 < \Delta, C_\rho < +\infty$ if

$$\rho_X \{x \in X : |f_{\mathrm{sp}}(x)| < \Delta t\} \leqslant C_\rho t^\theta, \qquad \text{for all } t > 0. \tag{1.14}$$

The largest θ satisfying (1.14) is called the **separation exponent** of ρ.

The following comparisom theorem was studied in[12, 6, 77].

Theorem. *For any measurable $f : X \to \mathbb{R}$, it holds*

$$\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f_c) \leqslant \sqrt{(\mathcal{E}_{\phi_{\mathrm{h}}}(f) - \mathcal{E}_{\phi_{\mathrm{h}}}(f_c))}.$$

In general, the ψ-transform

$$\psi : [0, 1] \to \mathbb{R}_+$$

satisfying

$$\psi\left(\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f_c)\right) \leqslant \mathcal{E}_\phi(f) - \mathcal{E}_\phi(f_c), \qquad \text{for all measurable } f : X \to \mathbb{R},$$

where

$$\mathcal{E}_\phi(f) = \int_{X \times Y} \phi_{\mathrm{h}}(yf(x))d\rho.$$

was studied in [6].

The solution analysis of regularized SVM (1.13) could be found in [22, 52, 61, 76] while the convergence results are obtained in [25, 42, 61, 62, 63, 77].

1.2.4 Online Learning Algorithms

In the previous section, we considered only learning algorithms handling the whole data set at one time, which is called **batch learning**. Meanwhile, there is a large class of algorithms that use data points one by one. These algorithms are referred to as online algorithms, and are some times used as fast substitutes for batch learning algorithms. Another important application of online algorithms is when users need to update the predicted function on the fly, while they keep obtaining new data points. In [38, 53, 70], a stochastic gradient descent (SGD) learning algorithm of least squares loss

$$f_{t+1} = f_t - \eta_t\left((f_t(x_t) - y_t)K_{x_t} + \lambda f_t\right), t = 1, 2, ..., N,$$

where $\eta_t > 0$ are step sizes, was studied to approximate the regression function f_ρ. Furthermore, in [69, 71], an SGD algorithm for a general loss function ϕ, which is convex and differentiable at 0 with

$$\phi'(0) < 0,$$

was utilized as

$$f_{t+1} = f_t - \eta_t \left(\phi'_-(y_t f_t(x_t)) y_t K_{x_t} + \lambda f_t\right), \qquad t = 1, 2, ..., N,$$

where ϕ'_- is the left derivative of ϕ. For non-strongly convex case, convergence analysis has been done in [5]. The optimal rate is attained by [26] using an averaged unregularized least squares algorithm under a large step-size assumption.

1.2.5 Algorithms for Ranking and Pairwise Learning

Bipartite ranking problems have been considered in learning theory for a long history (e.g. [1, 2, 17, 16, 18, 20, 21]). For learning the scoring functions, a method focused on maximizing the **AUC criterion**

$$\mathrm{AUC}(s) = \rho[s(x) \geqslant s(x')|y = 1, y' = -1]$$

was analyzed in [2, 18], while in [16], the convergence of the empirical version

$$L_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{1}_{\{z_{ij} f(x_i, x_j) < 0\}} \tag{1.15}$$

of $L(r)$ defined by (1.1), where

$$z_{ij} = \frac{y_i - y_j}{2},$$

was estimated. Note that $L_n(r)$ is a U-statistic of order 2, the estimation of training error between $L_n(r)$ and $L(r)$ has been done based on U-statistic theory combined with VC-dimension techniques by [16]. In [3, 47], the ranking learning problems was formulated under the framework of **pairwise learning** with a loss function

$$\phi : \mathbb{R}^X \times (X \times Y) \times (X \times Y) \to [0, +\infty)$$

$$(f, (x, y), (x', y')) \mapsto \phi(f, (x, y), (x', y'))$$

which is symmetric about (x, y) and (x', y'), where $\mathbb{R}^X := \{f : X \to \mathbb{R}\}$ is the set of all real functions on X and some error analysis based on kernel methods were achieved. In this settings, define the ranking error as

$$\mathcal{L}_\phi(f) = \int_{(X,Y)\times(X,Y)} \phi(f, (x, y), (x', y'))d\rho(x, y)d\rho(x', y').$$

For pairwise learning, online learning algorithms were used and analyzed in, for example, [72, 73, 64].

1.2.6 Deep Neural Networks

Artificial neural networks (ANN) dated back to multilayer perceptrons [48, 50]. ANN's, especially those with many layers (thus called deep neural networks), provide an alternative way (perhaps more successful nowadays) to generate nonlinear hypothesis spaces for learning, that parallel kernel methods. In recent years, the development of computing hardware has boosted a fast development of deep neural networks and deep learning. Many new algorithms, using different network structures, are studied and have achieved big successes in speech recognition and natural language processing (e.g. recurrent neural networks, RNN [27, 66]), and in image processing and computer vision (e.g. convolutional neural networks, CNN [39, 50]). However, we will not expand ANN in this thesis.

1.3 Centered Reproducing Kernels

In this Chapter, we first introduce the definition of a class of centered reproducing kernels and the motivations behind. In Chapter 2, some properties of the CRKs will be sumarized.

1.3.1 Centered Reproducing Kernels

Given a Mercer kernel K from $X \times X$ to $\mathbb{R}$, we define a new kernel $\bar{K}$ with respect to the marginal distribution ρ_X by

$$\bar{K}(x,u) = K(x,u) - \int_X K(\xi,u)d\rho_X(\xi) - \int_X K(x,\xi)d\rho_X(\xi) + \int_{X\times X} K(\xi,\xi')d\rho_X(\xi)\rho_X(\xi').$$

It's obvious that $\bar{K}$ is symmetric. It is shown in [29] that $\bar{K}$ is also a Mercer kernel (see Lemma 2.1 below). Since ρ_X is a probability measure,

$$\sqrt{\sup_{x\in X} \bar{K}(x,x)} \leqslant \sqrt{4\sup_{x\in X} K(x,x)} = 2\kappa.$$

For $\bar{K}$, an important property of the corresponding RKHS $\mathcal{H}_{\bar{K}}$ is that it contains no non-zero constant function. In fact, a straightforward calculation shows that

$$\int_X \bar{K}_x d\rho_X(x) = 0. \tag{1.16}$$

Since $\mathcal{H}_{\bar{K}}$ is spaned by $\bar{K}_x$ and completed with respect to the norm $\|\cdot\|_{\bar{K}}$ which is stronger than the $L^2_{\rho_X}$ norm, we have

$$\int_X f(x)d\rho_X(x) = 0, \qquad \text{for all } f \in \mathcal{H}_{\bar{K}}, \tag{1.17}$$

i.e., $\mathcal{H}_{\bar{K}}$ is perpendicular to the constant function $\mathbf{1}$ in $L^2_{\rho_X}(X)$.

Since the definition of $\bar{K}$ is based on the unknown marginal distribution ρ_X, in practice, we need to discretize $\bar{K}$ by

$$\hat{K}(x,u) := K(x,u) - \frac{1}{N}\sum_{i=1}^N K(x,x_i) - \frac{1}{N}\sum_{i=1}^N K(x_i,u) + \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N K(x_i,x_j).$$

Similar as (1.16) and (1.17), we have

$$\frac{1}{N}\sum_{i=1}^N \hat{K}_{x_i} = 0 \tag{1.18}$$

and

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) = 0, \qquad \text{for all } f \text{ in } \mathcal{H}_{\hat{K}}.$$

The relationship between $\hat{K}$ and $\bar{K}$ is given in section 2.1.

1.3.2 The Motivation of CRKs

In many regression problems, such as the score-based ranking problems, the constant component shoule be treated differently from the other part of the regression function. However, there is no existing approach serving this intuition. For example, we will give a simulation later which shows that the convergence rate of the output functions generated by KRR based on Gaussian RBF kernel could be much slower if we add the original regression function by a constant, since the RKHS spaned by a Gaussian kernel contains no non-zero constant [43]. Under this circumstance, we first separate the constant from the RKHS by centering the kernel to $\bar{K}$, and consider the constant term independently. Moreover, note that $\bar{K}$ denpends on the unknown distribution ρ, we approximate $\bar{K}$ by the discrete centered kernel $\hat{K}$.

Chapter 2

Regularized Least Squares with Centered Reproducing Kernels

In the previous chapter, we introduced the centered reproducing kernel (CRK) with respect to the marginal distribution ρ_X,

$$\bar{K}(x,u) = K(x,u) - \int_X K(\xi,u)d\rho_X(\xi) - \int_X K(x,\xi)d\rho_X(\xi) + \int_{X\times X} K(\xi,\xi')d\rho_X(\xi)d\rho_X(\xi'),$$

and the CRK with respect to the empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ concentrated on the observations,

$$\hat{K}(x,u) = K(x,u) - \frac{1}{N}\sum_{i=1}^N K(x,x_i) - \frac{1}{N}\sum_{i=1}^N K(x_i,u) + \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N K(x_i,x_j).$$

In this chapter, we first summarize the properties of CRKs. Then we study a modified kernel ridge regression based on CRKs and give the error analysis.

2.1 Properties of CRKs

Summarized in [29], the centered reproducing kernel $\bar{K}$ and the corresponding integral operator $L_{\bar{K}}$ and RKHS $\mathcal{H}_{\bar{K}}$ possess the following properties.

Lemma 2.1. *For a given Mercer kernel K and the corresponding centered kernel $\bar{K}$, integral operator $L_{\bar{K}}$ and the RKHS $\mathcal{H}_{\bar{K}}$, one has*

1. *$\bar{K}$ is a Mercer kernel on X. Moreover, $\mathcal{H}_{\bar{K}}$ is perpendicular to the constant function $\mathbf{1}(x) \equiv 1$ in $L^2_{\rho_X}(X)$.*

2. *Let $P_{\mathbb{1}} : L^2_{\rho_X}(X) \to L^2_{\rho_X}(X)$ denote the orthogonal projection operator onto the space spanned by $\mathbf{1}(x) \equiv 1$, i.e.,*

$$P_{\mathbb{1}} f = \int_X f(x) d\rho_X(x).$$

Moreover, let I be the identity operator on $L^2_{\rho_X}(X)$. We have

$$L_{\bar{K}} f = (I - P_{\mathbb{1}}) L_K (I - P_{\mathbb{1}}) f, \qquad \text{for any } f \text{ in } L^2_{\rho_X}(X).$$

3. *For the eigenvalues $\{\bar{\lambda}_i\}_{i=1}^{\infty}$ of $L_{\bar{K}}$ arranged in decreasing order, one has the interlacing relationship*

$$\lambda_1 \geqslant \bar{\lambda}_1 \geqslant \lambda_2 \geqslant \bar{\lambda}_2 \geqslant \ldots \geqslant \lambda_n \geqslant \bar{\lambda}_n \geqslant \ldots \tag{2.1}$$

between $\{\bar{\lambda}_i\}_{i=1}^{\infty}$ and $\{\lambda_i\}_{i=1}^{\infty}$.

4. *For any f in $L^2_{\rho_X}(X)$ and $1/2 \leqslant r \leqslant 1$, there is a real constant c and a function g in $L^2_{\rho_X}(X)$ with $\|g\|_\rho \leqslant \|f\|_\rho$ such that*

$$L_K^r g = L_{\bar{K}}^r f + c.$$

Moreover, if L_K has eigenfunction $\mathbf{1}$, then the requirement on r can be slacked to $r \in (0, +\infty)$.

5. *We have*

$$\overline{\mathcal{H}_K}^{\rho_X} \subseteq \overline{\operatorname{span}\{\mathbf{1}\} + \mathcal{H}_{\bar{K}}}^{\rho_X} \text{ and } \overline{\mathcal{H}_{\bar{K}}}^{\rho_X} \subseteq \overline{\operatorname{span}\{\mathbf{1}\} + \mathcal{H}_K}^{\rho_X},$$

where the completion $\bar{\cdot}^{\rho_X}$ is done with respect to the $L^2_{\rho_X}$ norm.

For the relationship between $\hat{K}$ and $\bar{K}$, we have the following Lemma.

Lemma 2.2. *For $\hat{K}$ and $\bar{K}$, we have*

1. *If we take the maps $K \mapsto \bar{K}$ and $K \mapsto \hat{K}$ as transformations of kernels and denote them by $\hat{\cdot}$ and $\bar{\cdot}$, respectively, then we have the following relations:*

$$\hat{K} = \hat{\bar{K}}, \qquad \bar{K} = \bar{\hat{K}}, \qquad \bar{\bar{K}} = \bar{K}, \qquad \hat{\hat{K}} = \hat{K}. \tag{2.2}$$

2. *Define $P_N = e_N e_N^T$ as the matrix of the orthogonal projection onto the space spanned by $e_N = \frac{1}{\sqrt{N}}(1, ..., 1)^T$ in $\mathbb{R}^N$. Let $I_N : \mathbb{R}^N \to \mathbb{R}^N$ be the identity matrix on $\mathbb{R}^N$. Then we have*

$$\hat{K}_{[\mathbf{x}]} = \bar{K}_{[\mathbf{x}]} - P_N \bar{K}_{[\mathbf{x}]} - \bar{K}_{[\mathbf{x}]} P_N + P_N \bar{K}_{[\mathbf{x}]} P_N = (I_N - P_N)\bar{K}_{[\mathbf{x}]}(I_N - P_N), \tag{2.3}$$

where $\hat{K}_{[\mathbf{x}]} = \left(\hat{K}(x_i, x_j)\right)_{N \times N}$ and $\bar{K}_{[\mathbf{x}]} = \left(\bar{K}(x_i, x_j)\right)_{N \times N}$ are the kernel matrices of $\hat{K}$ and $\bar{K}$, respectively. As a result, we have

$$\hat{K}_{[\mathbf{x}]} e_N = 0, \tag{2.4}$$

so e_N is an eigenvector of $\hat{K}_{[\mathbf{x}]}$ associated with the eigenvalue 0.

Proof. We only prove $\bar{\hat{K}} = \bar{K}$, and the rest proof of Item 1 follows from tedious but similar calculation.

Note that

$$\bar{\hat{K}}(s,t) = \hat{K}(s,t) - \int_X \hat{K}(\xi, t) d\rho_X(\xi) - \int_X \hat{K}(s, \xi') d\rho_X(\xi') + \int_{X \times X} \hat{K}(\xi, \xi') d\rho_X(\xi) d\rho_X(\xi')$$

$$= K(s,t) - \frac{1}{N}\sum_{i=1}^N K(x_i, t) - \frac{1}{N}\sum_{i=1}^N K(s, x_i) + \frac{1}{N^2}\sum_{p,q=1}^N K(x_p, x_q)$$

$$- \left(\int_X K(\xi, t) d\rho_X(\xi) - \frac{1}{N}\sum_{i=1}^N K(x_i, t) - \frac{1}{N}\sum_{i=1}^N \int_X K(\xi, x_i) d\rho_X(\xi) + \frac{1}{N^2}\sum_{p,q=1}^N K(x_p, x_q) \right)$$

$$-\left(\int_X K(s,\xi^{'})d\rho_X(\xi^{'})-\frac{1}{N}\sum_{i=1}^N\int_X K(x_i,\xi^{'})d\rho_X(\xi^{'})-\frac{1}{N}\sum_{i=1}^N K(s,x_i)+\frac{1}{N^2}\sum_{p,q=1}^N K(x_p,x_q)\right)$$

$$+\left(\int_{X\times X} K(\xi,\xi^{'})d\rho_X(\xi)d\rho_X(\xi^{'})-\frac{1}{N}\sum_{i=1}^N\int_X K(x_i,\xi^{'})d\rho_X(\xi^{'})\right.$$

$$\left.-\frac{1}{N}\sum_{i=1}^N\int_X K(\xi,x_i)d\rho_X(\xi)+\frac{1}{N^2}\sum_{p,q=1}^N K(x_p,x_q)\right)$$

$$=K(s,t)-\int_X K(\xi,t)d\rho_X(\xi)-\int_X K(s,\xi^{'})d\rho_X(\xi^{'})+\int_{X\times X} K(\xi,\xi^{'})d\rho_X(\xi)d\rho_X(\xi')$$

$$=\bar{K}(s,t).$$

Wo obtain $\bar{\hat{K}}=\bar{K}$ in (2.2).

For Item 2, since

$$P_N\bar{K}_{[\mathbf{x}]}=\frac{1}{\sqrt{N}}e_N\left(\sum_{i=1}^N\bar{K}(x_i,x_1),\sum_{i=1}^N\bar{K}(x_i,x_2),...,\sum_{i=1}^N\bar{K}(x_i,x_N)\right),$$

$$\bar{K}_{[\mathbf{x}]}P_N=\frac{1}{\sqrt{N}}\left(\sum_{i=1}^N\bar{K}(x_1,x_i),\sum_{i=1}^N\bar{K}(x_2,x_i),...,\sum_{i=1}^N\bar{K}(x_N,x_i)\right)^T e_N^T,$$

and

$$P_N\bar{K}_{[\mathbf{x}]}P_N=\left(\frac{1}{N}\sum_{i,j=1}^N\bar{K}(x_i,x_j)\right)e_Ne_N^T,$$

we obtain (2.3) by comparing each entry of both sides of the equation.

Equation (2.4) follows from

$$(I_N-P_N)\,e_N=e_N-e_N=0.$$

□

2.2 CRK-based Modified KRR Learning Algorithms

Recall the classical kernel ridge regression

$$f_\lambda^D = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{N}\sum_{i=1}^N (f(x_i)-y_i)^2 + \lambda\|f\|_K^2 \right\}. \tag{2.5}$$

As we mentioned before, to separate the constant component from the RKHS, we introduced centered reproducing kernels $\hat{K}$ and $\bar{K}$ for KRR. Based on $\hat{K}$, we modify the classical KRR as

$$(\hat{f}_\lambda^D, \hat{b}_\lambda^D) = \arg\min_{f\in\mathcal{H}_{\hat{K}}, b\in\mathbb{R}} \left\{ \frac{1}{N}\sum_{i=1}^N (f(x_i)+b-y_i)^2 + \lambda\|f\|_{\hat{K}}^2 \right\}. \tag{2.6}$$

Recall that

$$\sum_{i=1}^N \hat{K}_{x_i} = 0.$$

Note that

$$\sum_{i=1}^N f(x_i) = 0, \qquad \text{for all } f \text{ in } \mathcal{H}_{\hat{K}}.$$

By letting the gradient of (2.6) with respect to (f, b) in $\mathcal{H}_{\hat{K}} \oplus \mathbb{R}$ vanish, we get

$$\hat{b}_\lambda^D = \frac{1}{N}\sum_{i=1}^N y_i - \frac{1}{N}\sum_{i=1}^N f(x_i) = \frac{1}{N}\sum_{i=1}^N y_i, \tag{2.7}$$

$$\hat{f}_\lambda^D = \left(L_{\hat{K}}^{D(x)} + \lambda I\right)^{-1} \frac{1}{N}\sum_{i=1}^N (y_i - b)\hat{K}_{x_i} = \left(L_{\hat{K}}^{D(x)} + \lambda I\right)^{-1} \frac{1}{N}\sum_{i=1}^N y_i \hat{K}_{x_i}. \tag{2.8}$$

By using (1.18), we get

$$\hat{f}_\lambda^D = \left(L_{\hat{K}}^{D(x)} + \lambda I\right)^{-1} \frac{1}{N}\sum_{i=1}^N \left(y_i - \int_X f_\rho(x) d\rho_X(x)\right) \hat{K}_{x_i}.$$

Here the operator $L_{\hat{K}}^{D(x)}$ is defined similarly as $L_K^{D(x)}$, by replacing K by $\hat{K}$ in (1.6). By (1.8), one has

$$\hat{f}_\lambda^D = \arg\min_{f\in\mathcal{H}_{\hat{K}}} \left\{ \frac{1}{N}\sum_{i=1}^N \left(f(x_i) - y_i + \int_X f_\rho(x) d\rho_X(x) \right)^2 + \lambda \|f\|_{\hat{K}}^2 \right\}.$$

Thus by the representer theorem (1.9), we can represent $\hat{f}_\lambda^D$ as

$$\hat{f}_\lambda^D = \sum_{i=1}^N \hat{c}_i \hat{K}_{x_i},$$

where $\hat{c} = (\hat{c}_1, ..., \hat{c}_N)^T \in \mathbb{R}^N$ is the solution to

$$\min_{c\in\mathbb{R}^N} \left\{ \frac{1}{N} \left\| \hat{K}_{[\mathbf{x}]} c - \bar{\mathbf{y}} \right\|_2^2 + \lambda c^T \hat{K}_{[\mathbf{x}]} c \right\}, \tag{2.9}$$

with

$$\bar{\mathbf{y}} = \left(y_1 - \int_X f_\rho(x) d\rho_X(x), ..., y_N - \int_X f_\rho(x) d\rho_X(x) \right)^T \in \mathbb{R}^N.$$

By (2.4) in Lemma 2.2, (2.9) is equivalent to

$$\min_{c\in\mathbb{R}^N} \left\{ \frac{1}{N} \left\| \hat{K}_{[\mathbf{x}]} (I_N - P_N) c - \bar{\mathbf{y}} \right\|_2^2 + \lambda c^T (I_N - P_N) \hat{K}_{[\mathbf{x}]} (I_N - P_N) c \right\}.$$

As a result, $(I_N - P_N)\hat{c}$ is also a solution of (2.9). Since

$$\sum_{i=1}^N ((I_N - P_N)\hat{c})_i = \sqrt{N} e_N^T (I_N - P_N) \hat{c} = 0,$$

without loss of generality, we can assume that $\hat{c}$ has already been centered, i.e.,

$$\sum_{i=1}^N \hat{c}_i = 0. \tag{2.10}$$

So we define

$$\hat{c} = (I_N - P_N) \left(N\lambda I_N + \hat{K}_{[\mathbf{x}]} \right)^{-1} \bar{\mathbf{y}}.$$

The output function of CRKs-based learning algorithm is given by

$$\hat{f}_\lambda^D + \hat{b}_\lambda^D.$$

In classical KRR, the regularization assumption on f_ρ is

$$f_\rho = L_K^r h_\rho, \qquad \text{for some } h_\rho \text{ in } L^2_{\rho_X}(X) \text{ and } r > 0. \tag{2.11}$$

Let

$$\bar{f}_\rho := f_\rho - \int_X f_\rho(x) d\rho_X(x)$$

be the centered regression function. In this work, we modify the regularization assumption to be

$$\bar{f}_\rho = L_{\bar{K}}^r \bar{h}_\rho, \qquad \text{for some } \bar{h}_\rho \text{ in } L^2_{\rho_X}(X) \text{ and } r > 0. \tag{2.12}$$

Now we demonstrate our main results with the proof deferred to the next section.

Theorem 2.1. *Assume that $|y| \leqslant M$ almost surely and (2.12) with $1/2 \leqslant r \leqslant 1$. Take $\lambda = N^{-\frac{1}{2r+1}}$. Then*

$$\mathbb{E}\|\hat{f}_\lambda^D + \hat{b}_\lambda^D - f_\rho\|_\rho \leqslant C N^{-\frac{r}{2r+1}},$$

where C is a constant independent of D, N, or λ, and it is specified in the proof.

In this work, we always assume $|y| \leqslant M$ almost surely for $(x, y) \sim \rho$.

2.3 Error Analysis for the CRK-based Learning Algorithm

For error analysis, we insert a sample free analogue of $(\hat{f}_\lambda^D, \hat{b}_\lambda^D)$ based on $\bar{K}$. Define

$$(\bar{f}_\lambda, \bar{b}_\lambda) = \arg \min_{f \in \mathcal{H}_{\bar{K}}, b \in \mathbb{R}} \left\{ \int_{X \times Y} (f(x) + b - y)^2 d\rho(x, y) + \lambda \|f\|_{\bar{K}}^2 \right\}.$$

Recall the centered regression function

$$\bar{f}_\rho := f_\rho - \int_X f_\rho(x) d\rho_X(x).$$

Parallel to $(\hat{f}_\lambda^D, \hat{b}_\lambda^D)$ in (2.7), as a consequence of

$$\int_X f(x) d\rho_X(x) = 0, \qquad \text{for all } f \text{ in } \mathcal{H}_{\bar{K}},$$

we obtain

$$\bar{b}_\lambda = \int_X f_\rho(x) d\rho_X(x),$$

$$\bar{f}_\lambda = (L_{\bar{K}} + \lambda I)^{-1} L_{\bar{K}} \bar{f}_\rho = (L_{\bar{K}} + \lambda I)^{-1} L_{\bar{K}} f_\rho.$$

Moreover, based on $\bar{K}$, we define the empirical analogue of $\bar{f}_\lambda$ as

$$\bar{f}_\lambda^D = \left(L_{\bar{K}}^{D(x)} + \lambda I \right)^{-1} \frac{1}{N} \sum_{i=1}^N y_i \bar{K}_{x_i},$$

where $L_{\bar{K}}^{D(x)}$ is defined similarly as $L_K^{D(x)}$, by replacing K by $\bar{K}$ in (1.6). It is easy to see that $\bar{f}_\lambda^D$ is the solution to the minimization problem

$$\min_{f \in \mathcal{H}_{\bar{K}}} \left\{ \frac{1}{N} \sum_{i=1}^N (f(x_i) - y_i)^2 + \lambda \|f\|_{\bar{K}}^2 \right\}.$$

We do the error analysis by using the decomposition

$$\left\| \hat{f}_\lambda^D + \hat{b}_\lambda^D - f_\rho \right\|_\rho = \left\| \hat{f}_\lambda^D + \hat{b}_\lambda^D - \bar{f}_\rho - \bar{b}_\lambda \right\|_\rho$$

$$\leqslant \left\| \hat{f}_\lambda^D - \bar{f}_\lambda^D \right\|_\rho + \|\bar{f}_\lambda^D - \bar{f}_\lambda\|_\rho + \|\bar{f}_\lambda - \bar{f}_\rho\|_\rho + \left| \hat{b}_\lambda^D - \bar{b}_\lambda \right|. \tag{2.13}$$

By the interlacing relationship (2.1) in Lemma 2.1 between $L_{\bar{K}}$ and L_K, we have the effective dimension of $L_{\bar{K}}$ satisfies that

$$\mathcal{N}_{L_{\bar{K}}}(\lambda) \leqslant \mathcal{N}_{L_K}(\lambda).$$

The norms $\|\bar{f}_\lambda^D - \bar{f}_\lambda\|_\rho$ and $\|\bar{f}_\lambda - \bar{f}_\rho\|_\rho$ have already been well bounded in literature under the regularization assumption on $\bar{f}_\rho$ (2.12) and the assumption on effectice dimension (1.10). The second term in the right-hand side of (2.13) was bounded by [11, 40] while the third term was bounded by [54]. Precisely, we have the following Lemma.

Lemma 2.3 (Smale & Zhou, [54])**.** *If we assume (2.11) with* $0 < r \leqslant 1$*, then*

$$\|f_\lambda - f_\rho\|_\rho \leqslant \lambda^r \|h_\rho\|_\rho.$$

Moreover, for $1/2 \leqslant r \leqslant 1$*, we have*

$$\|f_\lambda - f_\rho\|_K \leqslant \lambda^{r-1/2} \|h_\rho\|_\rho.$$

We prepare some notations for the next Lemma. Define the **sampling operator**

$$\begin{aligned} S_D : \mathcal{H}_K &\to \mathbb{R}^N \\ f &\mapsto (f(x_1), ..., f(x_N))^T. \end{aligned} \tag{2.14}$$

So the adjoint operator of S_D has the form

$$\begin{aligned} S_D^T : \mathbb{R}^N &\to \mathcal{H}_K \\ \mathbf{y} &\mapsto \sum_{i=1}^N y_i K_{x_i}. \end{aligned} \tag{2.15}$$

Let $\bar{S}_D$ and $\bar{S}_D^T$ denote the sampling operator and the adjoint sampling operator on $\mathcal{H}_{\bar{K}}$, respectively. The relation between the sampling operator and the empirical integral operator can be described as

$$L_K^{D(x)} = \frac{1}{N} S_D^T S_D.$$

In what follows, for any Mercer kernel G on X, we let $\|\cdot\|_{\mathrm{op}(G)}$ denote the operator norm on the RKHS $(\mathcal{H}_G, \langle\cdot,\cdot\rangle_G, \|\cdot\|_G)$ associated with G. That is

$$\|L\|_{\mathrm{op}(G)} = \sup_{f \in \mathcal{H}_G, \|f\|_G \leqslant 1} \|Lf\|_G,$$

for any bounded linear operator $L : \mathcal{H}_G \to \mathcal{H}_G$.

Chang et al. [11] studied the classical KRR (2.5) without the offset term b, with the help of f_λ, defined by

$$\begin{aligned} f_\lambda &= \arg\min_{f \in \mathcal{H}_K} \left\{ \int_{X \times Y} (f(x) - y)^2 \, d\rho(x,y) + \lambda \|f\|_K^2 \right\} \\ &= \arg\min_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \right\}, \end{aligned} \tag{2.16}$$

and obtained the following Lemma.

Lemma 2.4 (Chang et al., [11], Proposition 6)**.** *We have*

$$\max\left\{\left\|f_\lambda^D - f_\lambda\right\|_\rho, \sqrt{\lambda}\left\|f_\lambda^D - f_\lambda\right\|_K\right\} \leqslant \mathcal{Q}_{D,\lambda}^2(\mathcal{P}_{D,\lambda} + \mathcal{S}_{D,\lambda}\|f_\lambda\|_K), \tag{2.17}$$

where

$$\mathcal{P}_{D,\lambda} := \left\|(L_K + \lambda I)^{-1/2}\left(L_K f_\rho - \frac{1}{N}S_D^T\mathbf{y}\right)\right\|_K, \tag{2.18}$$

$$\mathcal{Q}_{D,\lambda} := \|(L_K + \lambda I)^{1/2}(L_K^{D(x)} + \lambda I)^{-1/2}\|_{\mathrm{op}(K)}, \tag{2.19}$$

$$\mathcal{S}_{D,\lambda} := \left\|(L_K + \lambda I)^{-1/2}(L_K - L_K^{D(x)})\right\|_{\mathrm{op}(K)}. \tag{2.20}$$

Moreover, from (2.16), with assumption (2.11) for $r \geqslant 1/2$,

$$\|f_\lambda - f_\rho\|_\rho^2 + \lambda\|f_\lambda\|_K^2 \leqslant 0 + \lambda\|f_\rho\|_K^2,$$

so

$$\|f_\lambda\|_K \leqslant \|f_\rho\|_K \leqslant \|L_K^r h_\rho\|_K \leqslant \kappa^{2r-1}\left\|L_K^{1/2}h_\rho\right\|_K \leqslant \kappa^{2r-1}\|h_\rho\|_\rho, \tag{2.21}$$

where the last equality comes form [23], see also Corollary 4.13 in [24], and the second inequality follows from the estimate $\|L_K\|_{\mathrm{op}(K)} \leqslant \kappa^2$, in fact,

$$\|L_K f\|_K = \left\|\int_X f(x)K_x d\rho_X(x)\right\|_K \leqslant \|f\|_K \sup_x \|K_x\|_K^2 = \kappa^2\|f\|_K.$$

Similar as Lemma 2.4, we define

$$\bar{\mathcal{P}}_{D,\lambda} = \left\|(L_{\bar{K}} + \lambda I)^{-1/2}\left(L_{\bar{K}} f_\rho - \frac{1}{N}\bar{S}_D^T\mathbf{y}\right)\right\|_{\bar{K}},$$

$$\bar{\mathcal{Q}}_{D,\lambda} = \|(L_{\bar{K}} + \lambda I)^{1/2}(L_{\bar{K}}^{D(x)} + \lambda I)^{-1/2}\|_{\mathrm{op}(\bar{K})},$$

$$\bar{\mathcal{S}}_{D,\lambda} = \left\|(L_{\bar{K}} + \lambda I)^{-1/2}(L_{\bar{K}} - L_{\bar{K}}^{D(x)})\right\|_{\mathrm{op}(\bar{K})}.$$

Lemma 2.4 implies that

$$\max\left\{\left\|\bar{f}_\lambda^D - \bar{f}_\lambda\right\|_\rho, \sqrt{\lambda}\left\|\bar{f}_\lambda^D - \bar{f}_\lambda\right\|_K\right\} \leqslant \bar{\mathcal{Q}}_{D,\lambda}^2(\bar{\mathcal{P}}_{D,\lambda} + \bar{\mathcal{S}}_{D,\lambda}\|\bar{f}_\lambda\|_{\bar{K}}). \tag{2.22}$$

We cite from [11] the following Lemma, which is a summary of several results proved in [10, 31, 40].

Lemma 2.5. *Assume that* $|y| \leqslant M$ *almost surely and* $0 < \delta < 1$. *Then each of the following inequality holds with probability at least* $1 - \delta$,

$$\mathcal{P}_{D,\lambda} \leqslant 2M(\kappa + 1)\mathcal{A}_{D,\lambda} \log(2/\delta), \tag{2.23}$$

$$\mathcal{Q}^2_{D,\lambda} \leqslant 2\left(\frac{2(\kappa^2 + \kappa)\mathcal{A}_{D,\lambda} \log(2/\delta)}{\sqrt{\lambda}}\right)^2 + 2, \tag{2.24}$$

$$\mathcal{S}_{D,\lambda} \leqslant 2(\kappa^2 + \kappa)\mathcal{A}_{D,\lambda} \log(2/\delta), \tag{2.25}$$

where

$$\mathcal{A}_{D,\lambda} = \frac{1}{N\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}_{L_K}(\lambda)}}{\sqrt{N}}. \tag{2.26}$$

As a result of

$$\mathcal{N}_{L_{\bar{K}}}(\lambda) \leqslant \mathcal{N}_{L_K}(\lambda),$$

we have

$$\frac{1}{N\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}_{L_{\bar{K}}}(\lambda)}}{\sqrt{N}} \leqslant \mathcal{A}_{D,\lambda}.$$

Moreover, since

$$\sqrt{\sup_{x \in X} \bar{K}(x, x)} \leqslant 2\kappa,$$

we obtain that each of the following inequality holds with probability at least $1 - \delta$,

$$\bar{\mathcal{P}}_{D,\lambda} \leqslant 2M(2\kappa + 1)\mathcal{A}_{D,\lambda} \log(2/\delta), \tag{2.27}$$

$$\bar{\mathcal{Q}}^2_{D,\lambda} \leqslant 2\left(\frac{4(2\kappa^2 + \kappa)\mathcal{A}_{D,\lambda} \log(2/\delta)}{\sqrt{\lambda}}\right)^2 + 2, \tag{2.28}$$

$$\bar{\mathcal{S}}_{D,\lambda} \leqslant 4(2\kappa^2 + \kappa)\mathcal{A}_{D,\lambda} \log(2/\delta). \tag{2.29}$$

To carry out the error analysis, we use the following **first order and second order decomposition**. In general, for two invertible operators on a Banach space, we have the following Lemma [40].

Lemma 2.6 (Lin et al., [40]). *For A and B being two invertible operators on a Banach space, we have*

$$A^{-1} - B^{-1} = B^{-1}\left(B - A\right)A^{-1} = A^{-1}(B - A)B^{-1}. \tag{2.30}$$

Moreover, we have

$$A^{-1} - B^{-1} = B^{-1}(B - A)B^{-1} + B^{-1}(B - A)A^{-1}(B - A)B^{-1}. \tag{2.31}$$

As a direct use of the first and second order decomposition, we have the following Lemmas, which will be used to estimate the term $\hat{f}_\lambda^D - \bar{f}_\lambda^D$. Before giving the Lemmas, we first introduce some notations.

Define

$$\begin{aligned} &\hat{y}_i = y_i - f_\rho(x_i), \qquad \hat{\mathbf{y}} = (\hat{y}_1, ..., \hat{y}_N)^T, \\ &\tilde{\mathbf{y}} = \frac{1}{\sqrt{N}}\hat{\mathbf{y}}. \end{aligned} \tag{2.32}$$

We decompose $\bar{f}_\lambda^D$ as

$$\bar{f}_\lambda^D = \left(L_{\bar{K}}^{D(x)} + \lambda I\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \bar{y}_i \bar{K}_{x_i} + \bar{b}_\lambda \left(L_{\bar{K}}^{D(x)} + \lambda I\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \bar{K}_{x_i}. \tag{2.33}$$

By the representer theorem (1.9), we have

$$\left(L_{\bar{K}}^{D(x)} + \lambda I\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \bar{y}_i \bar{K}_{x_i} = \sum_{i=1}^{N} \bar{c}_i \bar{K}_{x_i},$$

where the coefficients

$$\bar{c} := (\bar{c}_1, ..., \bar{c}_N)^T = \left(\bar{K}_{[\mathbf{x}]} + N\lambda I_N\right)^{-1} \bar{\mathbf{y}}. \tag{2.34}$$

For simplicity, we define

$$\bar{G} = \bar{G}_{[\mathbf{x}]} = \frac{1}{N}\bar{K}_{[\mathbf{x}]}, \tag{2.35}$$

$$\hat{G} = \hat{G}_{[\mathbf{x}]} = \frac{1}{N}\hat{K}_{[\mathbf{x}]}. \tag{2.36}$$

Recall that $\hat{f}_\lambda^D$ is a part of the solution in (2.6), and we have defined the vector $\hat{c} = (\hat{c}_1, ..., \hat{c}_N)$ such that

$$\hat{f}_\lambda^D = \sum_{i=1}^N \hat{c}_i \hat{K}_{x_i}.$$

From (2.34) we have $\bar{c} = \frac{1}{N}\left(\bar{G} + \lambda I_N\right)^{-1}\bar{\mathbf{y}}$. Recall that $\hat{c} = \frac{1}{N}\left(I_N - P_N\right)\left(\hat{G} + \lambda I_N\right)^{-1}\bar{\mathbf{y}}$. According to (2.2), (2.10) and (2.33), we have

$$\begin{aligned}
\hat{f}_\lambda^D - \bar{f}_\lambda^D &= \sum_{i=1}^N \hat{c}_i \hat{K}_{x_i} - \sum_{i=1}^N \bar{c}_i \bar{K}_{x_i} - \bar{b}_\lambda \left(L_{\bar{K}}^{D(x)} + \lambda I\right)^{-1} \frac{1}{N}\sum_{i=1}^N \bar{K}_{x_i} \\
&= \sum_{i=1}^N \hat{c}_i \left(\bar{K}_{x_i} - \frac{1}{N}\sum_{j=1}^N \bar{K}_{x_j} - \frac{1}{N}\sum_{j=1}^N \bar{K}(x_i, x_j) + \frac{1}{N^2}\sum_{1\leqslant p,q\leqslant N} \bar{K}(x_p, x_q)\right) \\
&- \sum_{i=1}^N \bar{c}_i \bar{K}_{x_i} - \bar{b}_\lambda \left(L_{\bar{K}}^{D(x)} + \lambda I\right)^{-1} \frac{1}{N}\sum_{i=1}^N \bar{K}_{x_i} \\
&= \sum_{i=1}^N \left(\hat{c}_i \bar{K}_{x_i} - \bar{c}_i \bar{K}_{x_i}\right) - \frac{1}{N}\sum_{1\leqslant i,j\leqslant N} \hat{c}_i \bar{K}(x_i, x_j) - \bar{b}_\lambda \left(L_{\bar{K}}^{D(x)} + \lambda I\right)^{-1} \frac{1}{N}\sum_{i=1}^N \bar{K}_{x_i} \\
&=: J_1 + J_2 + J_3. \qquad (2.37)
\end{aligned}$$

We bound the three terms separately below.

To bound J_1, we need the following concentration inequality.

Lemma 2.7 (Pinelis-Hoeffding [45])**.** *Let $\{\xi_i\}$ be a sequence of independent random variables in a Hilbert space $(H, \|\cdot\|_H)$ with $\mathbb{E}\xi_i = 0$ and $\|\xi_i\|_H \leqslant c_i < \infty$ almost surely for every i. Then for any $\epsilon > 0$ and $N \geqslant 1$,*

$$\operatorname{Prob}\left\{\left\|\sum_{i=1}^N \xi_i\right\|_H \geqslant \epsilon\right\} \leqslant 2\exp\left\{-\frac{\epsilon^2}{8\sum_{i=1}^N c_i^2}\right\}. \qquad (2.38)$$

As a direct application of the Pinelis-Hoeffding inequality, we have the following lemma.

Lemma 2.8. *For any vector* $\eta = (\eta_1, ..., \eta_N)^T \in \mathbb{R}^N$ *of coefficients, and* $\epsilon > 0$, *we have*

$$\text{Prob}\left\{\left|\eta^T \tilde{\mathbf{y}}\right| \geqslant \epsilon | D(x)\right\} \leqslant 2 \exp\left\{-\frac{N\epsilon^2}{2M^2 \left\|\eta\right\|_2^2}\right\}, \tag{2.39}$$

or equivalently, for $0 < \delta < 1$,

$$\text{Prob}\left\{\left|\eta^T \tilde{\mathbf{y}}\right| \geqslant 2M \left\|\eta\right\|_2 \sqrt{\frac{2\log(2/\delta)}{N}} \middle| D(x)\right\} \leqslant \delta. \tag{2.40}$$

Proof. Recall that $\tilde{\mathbf{y}} = (\tilde{y}_1, ..., \tilde{y}_N)^T$ with

$$\tilde{y}_i = \frac{1}{\sqrt{N}} \left(y_i - f_\rho(x_i)\right), \qquad 1 \leqslant i \leqslant N.$$

So $\mathbb{E}\eta_i\tilde{y}_i = 0$ and $|\eta_i\tilde{y}_i| \leqslant \frac{2M}{\sqrt{N}}|\eta_i|$ almost surely. One applies Lemma 2.7 to obtain that for $\epsilon > 0$,

$$\text{Prob}\left\{\left|\eta^T \tilde{\mathbf{y}}\right| \geqslant \epsilon | D(x)\right\} \leqslant 2 \exp\left\{-\frac{N\epsilon^2}{8M^2 \left\|\eta\right\|_2^2}\right\}.$$

This proves (2.39), and (2.40) is obtained by letting the right-hand side of (2.39) equal δ. □

The following Lemma estimates J_1 in $\mathcal{H}_{\bar{K}}$.

Lemma 2.9. *Assume* $\bar{f}_\rho \in \mathcal{H}_{\bar{K}}$. *We have*

$$\left\|\sum_{i=1}^N \left(\hat{c}_i - \bar{c}_i\right) \bar{K}_{x_i}\right\|_{\bar{K}} \leqslant \frac{2\left\|\bar{G}^{1/2} e_N\right\|_2}{\sqrt{\lambda}} \left(\frac{1}{\sqrt{\lambda}} \left| e_N^T \left(\frac{1}{\lambda}\bar{G} + I_N\right)^{-1} \tilde{\mathbf{y}}\right| + \left\|\bar{f}_\rho\right\|_{\bar{K}}\right). \tag{2.41}$$

Proof. As a result of (2.3), we get

$$\left(\hat{G} + \lambda I\right)^{-1} (I_N - P_N) = (I_N - P_N) \left(\hat{G} + \lambda I\right)^{-1}.$$

By (2.30), one has

$$\begin{aligned}\left\|\sum_{i=1}^{N}\left(\hat{c}_i-\bar{c}_i\right)\bar{K}_{x_i}\right\|_{\bar{K}}^2 &= \left(\hat{c}-\bar{c}\right)^T\bar{K}_{[\mathbf{x}]}\left(\hat{c}-\bar{c}\right)=\left\|\bar{K}_{[\mathbf{x}]}^{1/2}\left(\hat{c}-\bar{c}\right)\right\|_2^2\\ &= \left\|\bar{G}^{1/2}\left[\left(I_N-P_N\right)\left(\hat{G}+\lambda I_N\right)^{-1}-\left(\bar{G}+\lambda I_N\right)^{-1}\right]\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right\|_2^2\\ &= \left\|\bar{G}^{1/2}\left(\hat{G}+\lambda I_N\right)^{-1}\left[\left(I_N-P_N\right)\bar{G}P_N-\lambda P_N\right]\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right\|_2^2.\end{aligned} \tag{2.42}$$

Recall that $\frac{1}{\sqrt{N}}\bar{\mathbf{y}}=\tilde{\mathbf{y}}+\frac{1}{\sqrt{N}}\bar{S}_D\bar{f}_\rho$ (see (2.14), (2.15) and the paragraph below for the definition of $\bar{S}_D$). We decompose the right-hand side of (2.42) above, to give

$$\left\|\sum_{i=1}^{N}\left(\hat{c}_i-\bar{c}_i\right)\bar{K}_{x_i}\right\|_{\bar{K}}\leqslant S_1+S_2,$$

with

$$\begin{aligned}S_1 &:= \left\|\bar{G}^{1/2}\left(\hat{G}+\lambda I_N\right)^{-1}\left[\left(I_N-P_N\right)\bar{G}P_N-\lambda P_N\right]\left(\bar{G}+\lambda I_N\right)^{-1}\tilde{\mathbf{y}}\right\|_2,\\ S_2 &:= \left\|\bar{G}^{1/2}\left(\hat{G}+\lambda I_N\right)^{-1}\left[\left(I_N-P_N\right)\bar{G}P_N-\lambda P_N\right]\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{S}_D\bar{f}_\rho\right\|_2.\end{aligned}$$

Below we abuse the notations by letting $\|\cdot\|_2$ also denote the spectral norm of a matrix (i.e. the maximum singular value). This will not introduce ambiguity. Recall that $\left\|AA^T\right\|_2=\left\|A^TA\right\|_2$ for any matrix A, and that $\hat{G}=\left(I_N-P_N\right)\bar{G}\left(I_N-P_N\right)$. We have

$$\begin{aligned}&\left\|\bar{G}^{1/2}\left(\hat{G}+\lambda I_N\right)^{-1}\left(I_N-P_N\right)\bar{G}^{1/2}\right\|_2\\ =&\left\|\bar{G}^{1/2}\left(I_N-P_N\right)\left(\hat{G}+\lambda I_N\right)^{-1/2}\left(\hat{G}+\lambda I_N\right)^{-1/2}\left(I_N-P_N\right)\bar{G}^{1/2}\right\|_2\\ =&\left\|\left(\hat{G}+\lambda I_N\right)^{-1/2}\hat{G}\left(\hat{G}+\lambda I_N\right)^{-1/2}\right\|_2\leqslant 1.\end{aligned} \tag{2.43}$$

Now we apply (2.43) to obtain

$$\begin{aligned}
S_1 \leqslant & \left\| \bar{G}^{1/2} \left(\hat{G} + \lambda I_N \right)^{-1} \left(I_N - P_N \right) \bar{G} P_N \left(\bar{G} + \lambda I_N \right)^{-1} \tilde{\mathbf{y}} \right\|_2 \\
& + \lambda \left\| \bar{G}^{1/2} \left(\hat{G} + \lambda I_N \right)^{-1} P_N \left(\bar{G} + \lambda I_N \right)^{-1} \tilde{\mathbf{y}} \right\|_2 \\
\leqslant & \left\| \bar{G}^{1/2} P_N \left(\bar{G} + \lambda I_N \right)^{-1} \tilde{\mathbf{y}} \right\|_2 + \lambda \left\| \bar{G}^{1/2} e_N \right\|_2 \left| e_N^T \left(\hat{G} + \lambda I_N \right)^{-1} e_N \right| \left| e_N^T \left(\bar{G} + \lambda I_N \right)^{-1} \tilde{\mathbf{y}} \right| \\
\leqslant & \, 2 \left\| \bar{G}^{1/2} e_N \right\|_2 \left| e_N^T \left(\bar{G} + \lambda I_N \right)^{-1} \tilde{\mathbf{y}} \right|, \qquad (2.44)
\end{aligned}$$

and

$$\begin{aligned}
S_2 \leqslant & \left\| \bar{G}^{1/2} \left(\hat{G} + \lambda I_N \right)^{-1} \left(I_N - P_N \right) \bar{G} P_N \left(\bar{G} + \lambda I_N \right)^{-1} \frac{1}{\sqrt{N}} \bar{S}_D \bar{f}_\rho \right\|_2 \\
& + \lambda \left\| \bar{G}^{1/2} \left(\hat{G} + \lambda I_N \right)^{-1} P_N \left(\bar{G} + \lambda I_N \right)^{-1} \frac{1}{\sqrt{N}} \bar{S}_D \bar{f}_\rho \right\|_2 \\
\leqslant & \left\| \bar{G}^{1/2} P_N \left(\bar{G} + \lambda I_N \right)^{-1} \frac{1}{\sqrt{N}} \bar{S}_D \bar{f}_\rho \right\|_2 \\
& + \lambda \left\| \bar{G}^{1/2} e_N \right\|_2 \left| e_N^T \left(\hat{G} + \lambda I_N \right)^{-1} e_N \right| \left| e_N^T \left(\bar{G} + \lambda I_N \right)^{-1} \frac{1}{\sqrt{N}} \bar{S}_D \bar{f}_\rho \right|.
\end{aligned}$$

Note that

$$\begin{aligned}
& \left\| \left(\bar{G} + \lambda I \right)^{-1/2} \frac{1}{\sqrt{N}} \bar{S}_D \bar{f}_\rho \right\|_2^2 = \left\langle \left(\bar{G} + \lambda I \right)^{-1/2} \frac{1}{\sqrt{N}} \bar{S}_D f_\rho, \left(\bar{G} + \lambda I \right)^{-1/2} \frac{1}{\sqrt{N}} \bar{S}_D \bar{f}_\rho \right\rangle_2 \\
& = \frac{1}{N} \left\langle \bar{S}_D^T \left(\bar{G} + \lambda I \right)^{-1} \bar{S}_D \bar{f}_\rho, \bar{f}_\rho \right\rangle_{\bar{K}} = \frac{1}{N} \left\langle \bar{S}_D^T \left(\frac{1}{N} \bar{S}_D \bar{S}_D^T + \lambda I \right)^{-1} \bar{S}_D \bar{f}_\rho, \bar{f}_\rho \right\rangle_{\bar{K}} \\
& = \frac{1}{N} \left\langle \left(\frac{1}{N} \bar{S}_D^T \bar{S}_D + \lambda I \right)^{-1} \left(\frac{1}{N} \bar{S}_D^T \bar{S}_D + \lambda I \right) \bar{S}_D^T \left(\frac{1}{N} \bar{S}_D \bar{S}_D^T + \lambda I \right)^{-1} \bar{S}_D \bar{f}_\rho, \bar{f}_\rho \right\rangle_{\bar{K}} \\
& = \left\langle \left(\frac{1}{N} \bar{S}_D^T \bar{S}_D + \lambda I \right)^{-1} \frac{1}{N} \bar{S}_D^T \bar{S}_D \bar{f}_\rho, \bar{f}_\rho \right\rangle_{\bar{K}} \leqslant \left\| \bar{f}_\rho \right\|_{\bar{K}}^2 .
\end{aligned}$$

So,

$$S_2 \leqslant 2 \left\| \bar{G}^{1/2} e_N \right\|_2 \left\| \left(\bar{G} + \lambda I_N\right)^{-1/2} e_N \right\|_2 \left\| \left(\bar{G} + \lambda I_N\right)^{-1/2} \frac{1}{\sqrt{N}} \bar{S}_D \bar{f}_\rho \right\|_2$$

$$\leqslant \frac{2 \left\| \bar{G}^{1/2} e_N \right\|_2}{\sqrt{\lambda}} \left\| \bar{f}_\rho \right\|_{\bar{K}} . \tag{2.45}$$

This completes the proof. □

Lemma 2.10. *Recall* $\bar{G} = \frac{1}{N} \bar{K}_{[\mathbf{x}]}$. *We have*

$$\mathbb{E} e_N^T \bar{G} e_N \leqslant \frac{4\kappa^2}{N}, \qquad \mathbb{E} \left(e_N^T \bar{G} e_N\right)^2 \leqslant \frac{48\kappa^2}{N^2}.$$

So, according to Hölder's inequality, $\mathbb{E}[\left\| \bar{G}^{1/2} e_N \right\|_2^r] \leqslant \left(\frac{2\kappa}{\sqrt{N}}\right)^r$, *for any* $r \in (0, 2]$.

Proof. Since

$$e_N^T \bar{G} e_N = \frac{1}{N^2} \sum_{i,j=1}^N \bar{K}(x_i, x_j) = \frac{1}{N^2} \sum_{i=1}^N \bar{K}(x_i, x_i) + \frac{1}{N^2} \sum_{i \neq j} \bar{K}(x_i, x_j),$$

and

$$\mathbb{E} \left[\bar{K}(x_i, x_j) | x_j \right] = 0, \qquad \text{for } i \neq j, \tag{2.46}$$

we have

$$\mathbb{E} e_N^T \bar{G} e_N = \frac{1}{N} \mathbb{E} \bar{K}(x_i, x_i) \leqslant \frac{(2\kappa)^2}{N}.$$

Moreover, we have

$$\left(e_N^T \bar{G} e_N\right)^2 = \left(\frac{1}{N^2} \sum_{i,j=1}^N \bar{K}(x_i, x_j) \right)^2 = \left(\frac{1}{N^2} \sum_{i=1}^N \bar{K}(x_i, x_i) + \frac{1}{N^2} \sum_{i \neq j} \bar{K}(x_i, x_j) \right)^2$$

$$= \frac{1}{N^4} \left\{ \left(\sum_{i=1}^N \bar{K}(x_i, x_i) \right)^2 + 2 \left(\sum_{i=1}^N \bar{K}(x_i, x_i) \right) \left(\sum_{k \neq l} \bar{K}(x_k, x_l) \right) + \left(\sum_{i \neq j} \bar{K}(x_i, x_j) \right)^2 \right\}$$

Also by the degenerate property (2.46), we obtain

$$\mathbb{E}\left(\sum_{i=1}^{N}\bar{K}(x_i,x_i)\right)\left(\sum_{k\neq l}\bar{K}(x_k,x_l)\right)=0,$$

and

$$\mathbb{E}\left(\sum_{i\neq j}\bar{K}(x_i,x_j)\right)^2=\mathbb{E}\sum_{i\neq j,k\neq l}\bar{K}(x_i,x_j)\bar{K}(x_k,x_l)=2\mathbb{E}\sum_{i\neq j}\bar{K}(x_i,x_j)^2\leqslant 32\kappa^4N(N-1).$$

Since

$$\left(\sum_{i=1}^{N}\bar{K}(x_i,x_i)\right)^2\leqslant 16N^2\kappa^4,$$

we get

$$\mathbb{E}\left(e_N^T\bar{G}e_N\right)^2\leqslant\frac{48\kappa^4}{N^2}.$$

□

Now we estimate the error between $\hat{f}_\lambda^D$ and $\bar{f}_\lambda^D$.

Proposition 2.1. *Assume $|y|\leqslant M$ almost surely and $\bar{f}_\rho\in\mathcal{H}_{\bar{K}}$. We have*

$$\mathbb{E}\left\|\hat{f}_\lambda^D-\bar{f}_\lambda^D\right\|_\rho\leqslant C_1'\left(\frac{1}{\sqrt{N\lambda}}+\frac{1}{N\lambda}+\frac{\mathcal{A}_{D,\lambda}}{\lambda\sqrt{N}}\right),$$

where C_1' is a constant independent of D,N, or λ, and it will be specified in the proof.

Proof. Recall the decomposition (2.37). We have

$$\begin{aligned}\|J_1\|_\rho=\left\|L_{\bar{K}}{}^{1/2}J_1\right\|_{\bar{K}}&\leqslant 2\kappa\|J_1\|_{\bar{K}}\\&\leqslant\frac{4\kappa\left\|\bar{G}^{1/2}e_N\right\|_2}{\sqrt{\lambda}}\left(\frac{1}{\sqrt{\lambda}}\left|e_N^T\left(\frac{1}{\lambda}\bar{G}+I_N\right)^{-1}\tilde{\mathbf{y}}\right|+\|\bar{f}_\rho\|_{\bar{K}}\right).\end{aligned}\tag{2.47}$$

Obviously $\left\|\left(\frac{1}{\lambda}\bar{G}+I_N\right)^{-1}e_N\right\|_2 \leqslant \|e_N\|_2 = 1$. We apply Lemma 2.8 to obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(e_N^T\left(\frac{1}{\lambda}\bar{G}+I_N\right)^{-1}\tilde{\mathbf{y}}\right)^2\middle| D(x)\right]\\
=&\int_0^\infty \operatorname{Prob}\left(\left|e_N^T\left(\frac{1}{\lambda}\bar{G}+I_N\right)^{-1}\tilde{\mathbf{y}}\right|>\sqrt{t}\middle| D(x)\right)dt\\
\leqslant&2\int_0^\infty \exp\left\{-\frac{Nt}{2M^2}\right\}dt=\frac{4M^2}{N}.
\end{aligned}$$

We apply Hölder's inequality to give

$$\mathbb{E}\left[\left|e_N^T\left(\frac{1}{\lambda}\bar{G}+I_N\right)^{-1}\tilde{\mathbf{y}}\right|\middle| D(x)\right]\leqslant\frac{2M}{\sqrt{N}}.$$

According to the decomposition $\mathbb{E}[\|J_1\|_{\bar{K}}] = \mathbb{E}[\mathbb{E}[\|J_1\|_{\bar{K}}|D(x)]]$, we use (2.47) and Lemma 2.10 to obtain

$$\mathbb{E}\left[\|J_1\|_\rho\right]\leqslant\frac{8\kappa^2}{\sqrt{N\lambda}}\left(\frac{2M}{\sqrt{N\lambda}}+\|\bar{f}_\rho\|_{\bar{K}}\right).$$

Obviously,

$$\begin{aligned}
-J_2&=\frac{1}{\sqrt{N}}e_N^T\bar{K}_{[\mathbf{x}]}\frac{1}{N}\left(I_N-P_N\right)\left(\hat{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}}\\
&=e_N^T\bar{G}\left(I_N-P_N\right)\left(\hat{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}.
\end{aligned}$$

The fact $\left\|\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right\|_2\leqslant 2M$ implies that

$$\begin{aligned}
|J_2|&\leqslant\left\|\bar{G}^{1/2}e_N\right\|_2\left\|\bar{G}^{1/2}\left(I_N-P_N\right)\left(\hat{G}+\lambda I_N\right)^{-1/2}\right\|_2\frac{1}{\sqrt{\lambda}}\left\|\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right\|_2\\
&\leqslant\frac{2M\left\|\bar{G}^{1/2}e_N\right\|_2}{\sqrt{\lambda}}\left\|\left(\hat{G}+\lambda I_N\right)^{-1/2}\left(I_N-P_N\right)\bar{G}\left(I_N-P_N\right)\left(\hat{G}+\lambda I_N\right)^{-1/2}\right\|_2^{1/2}\\
&\leqslant\frac{2M}{\sqrt{\lambda}}\left\|\bar{G}^{1/2}e_N\right\|_2.
\end{aligned}$$

We use Lemma 2.10 to get

$$\mathbb{E}[|J_2|] \leqslant \frac{4M\kappa}{\sqrt{N\lambda}}.$$

Now we estimate $\|J_3\|_\rho$. Since $J_3 \in \mathcal{H}_{\bar{K}}$ and $|\bar{b}_\lambda| \leqslant M$,

$$\begin{aligned}
\|J_3\|_\rho &\leqslant \left\| (L_{\bar{K}} + \lambda I)^{1/2} \left(L_{\bar{K}}^{D(x)} + \lambda I \right)^{-1} \frac{\bar{b}_\lambda}{N} \sum_{i=1}^{N} \bar{K}_{x_i} \right\|_{\bar{K}} \\
&\leqslant \bar{\mathcal{Q}}_{D,\lambda} \left\| \left(L_{\bar{K}}^{D(x)} + \lambda I \right)^{-1/2} \bar{b}_\lambda \frac{1}{N} \sum_{i=1}^{N} \bar{K}_{x_i} \right\|_{\bar{K}} \\
&\leqslant \frac{M \bar{\mathcal{Q}}_{D,\lambda}}{\sqrt{\lambda}} \left\| \frac{1}{N} \sum_{i=1}^{N} \bar{K}_{x_i} \right\|_{\bar{K}} = \frac{M \bar{\mathcal{Q}}_{D,\lambda}}{\sqrt{\lambda}} \left\| \bar{G}^{1/2} e_N \right\|_2. \qquad (2.48)
\end{aligned}$$

Let $\vartheta = \frac{2}{\lambda} \left(4(2\kappa^2 + \kappa)\mathcal{A}_{D,\lambda} \right)^2$. We rewrite (2.28) as

$$\text{Prob} \left(\bar{\mathcal{Q}}_{D,\lambda}^2 \geqslant 2 + \vartheta \log^2 \frac{2}{\delta} \right) \leqslant \delta.$$

Let $x = \vartheta \log^2 \frac{2}{\delta} + 2 \in [2 + \vartheta \log^2 2, +\infty)$ to obtain a solution $\delta = 2\exp(-\sqrt{(x-2)/\vartheta})$. By letting $u = \sqrt{\frac{x-2}{\vartheta}}$, we have

$$\begin{aligned}
\mathbb{E}[\bar{\mathcal{Q}}_{D,\lambda}^2] &\leqslant \int_0^{2+\vartheta \log^2 2} dx + 2 \int_{2+\vartheta \log^2 2}^{\infty} \exp\left\{ -\sqrt{\frac{x-2}{\vartheta}} \right\} dx \\
&= 2 + \vartheta \log^2 2 + 4\vartheta \int_{\log 2}^{\infty} u e^{-u} du \\
&= 2 + \vartheta \log^2 2 + 2\vartheta (\log 2 + 1) \leqslant 4\vartheta + 2.
\end{aligned}$$

We apply Hölder's inequality to (2.48) to give

$$\begin{aligned}
\mathbb{E}[\|J_3\|_\rho] &\leqslant \frac{M}{\sqrt{\lambda}} \sqrt{\frac{8}{\lambda} (4(2\kappa^2 + \kappa)\mathcal{A}_{D,\lambda})^2 + 2} \frac{2\kappa}{\sqrt{N}} \\
&\leqslant \frac{2\kappa M}{\sqrt{N\lambda}} \left(\sqrt{2} + \frac{8\sqrt{2}(2\kappa^2 + \kappa)}{\sqrt{\lambda}} \mathcal{A}_{D,\lambda} \right).
\end{aligned}$$

The proof is completed by letting

$$C_1^{'} = \max\left\{8\kappa^2 \left\|\bar{f}_\rho\right\|_{\bar{K}} + 4M\kappa + 2\sqrt{2}\kappa M, 16\kappa^2 M, 16\sqrt{2}\kappa M(2\kappa^2+\kappa)\right\}.$$

□

Proof of Theorem 2.1. We decompose the norm $\left\|\hat{f}_\lambda^D + \hat{b}_\lambda^D - f_\rho\right\|_\rho$ according to (2.13), of which the four terms at the right-hand side are estimated one by one below. First, take $\lambda = N^{-\frac{1}{2r+1}}$. Recall the definition of $\mathcal{A}_{D,\lambda}$ and the fact that $\mathcal{N}_{L_{\bar{K}}}(\lambda) \leqslant \mathcal{N}_{L_K}(\lambda) \leqslant \frac{1}{\lambda}$. We have

$$\mathcal{A}_{D,\lambda} = \frac{1}{N\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}_{L_K}(\lambda)}}{\sqrt{N}} \leqslant N^{-1+\frac{1/2}{2r+1}} + N^{-\frac{1}{2}+\frac{1/2}{2r+1}} \leqslant 2N^{-\frac{r}{2r+1}}.$$

Proposition 2.1 implies

$$\mathbb{E}\left\|\hat{f}_\lambda^D - \bar{f}_\lambda^D\right\|_\rho \leqslant C_1^{'}\left(N^{-\frac{r}{2r+1}} + N^{-\frac{2r}{2r+1}} + 2N^{-\frac{2r-1/2}{2r+1}}\right) \leqslant 4C_1^{'}N^{-\frac{r}{2r+1}}.$$

Similar as (2.21), we have $\left\|\bar{f}_\lambda\right\|_{\bar{K}} \leqslant (2\kappa)^{2r-1}\|\bar{h}_\rho\|_\rho$. We substitue the estimates (2.27), (2.28), and (2.29) into (2.22) to give that with confidence $1-\delta$,

$$\begin{aligned}&\left\|\bar{f}_\lambda^D - \bar{f}_\lambda\right\|_\rho \leqslant \bar{\mathcal{Q}}_{D,\lambda}^2(\bar{\mathcal{P}}_{D,\lambda} + \bar{\mathcal{S}}_{D,\lambda}\|\bar{f}_\lambda\|_{\bar{K}})\\ \leqslant&\left(2(4(2\kappa^2+\kappa))^2\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}\log^2\frac{6}{\delta}+2\right)\times\\ &\left(2M(2\kappa+1)\mathcal{A}_{D,\lambda}\log\frac{6}{\delta} + 4(2\kappa^2+\kappa)\mathcal{A}_{D,\lambda}\left(\log\frac{6}{\delta}\right)(2\kappa)^{2r-1}\|\bar{h}_\rho\|_\rho\right)\\ \leqslant& C_1^* N^{-\frac{r}{2r+1}}\log^3\frac{6}{\delta},\end{aligned} \tag{2.49}$$

where we have used the fact that $\frac{\mathcal{A}_{D,\lambda}^2}{\lambda} \leqslant 4N^{-\frac{2r-1}{2r+1}} \leqslant 4$ thanks to the assumption $r \geqslant 1/2$, and the constant C_1^* is defined by

$$C_1^* = (8(4(2\kappa^2+\kappa)))^2 + 2)(4M(2\kappa+1) + 8(2\kappa^2+\kappa)(2\kappa)^{2r-1}\|\bar{h}_\rho\|_\rho).$$

We write (2.49) equivalently as

$$\mathrm{Prob}\left(\left\|\bar{f}_\lambda^D - \bar{f}_\lambda\right\|_\rho \geqslant C_1^* N^{-\frac{r}{2r+1}} \log^3 \frac{6}{\delta}\right) \leqslant \delta.$$

Let $x = C_1^* N^{-\frac{r}{2r+1}} \log^3 \frac{6}{\delta} \in [C_1^* N^{-\frac{r}{2r+1}} \log^3 6, +\infty)$ to obtain a solution $\delta = 6 \exp\left\{-\sqrt[3]{x N^{\frac{r}{2r+1}} (C_1^*)^{-1}}\right\}$. We let $u = \sqrt[3]{x N^{\frac{r}{2r+1}} (C_1^*)^{-1}}$ to obtain

$$\begin{aligned}
\mathbb{E}\left\|\bar{f}_\lambda^D - \bar{f}_\lambda\right\|_\rho &\leqslant C_1^* N^{-\frac{r}{2r+1}} \log^3 6 + 6\int_{C_1^* N^{-\frac{r}{2r+1}} \log^3 6}^{\infty} e^{-u} du \\
=& C_1^* N^{-\frac{r}{2r+1}} \log^3 6 + 18 N^{-\frac{r}{2r+1}} C_1^* \int_{\log 6}^{\infty} e^{-u} u^2 du \\
\leqslant& 33 C_1^* N^{-\frac{r}{2r+1}}.
\end{aligned} \tag{2.50}$$

For the third term at the right-hand side of (2.13), we use Lemma 2.3 to give

$$\left\|\bar{f}_\lambda - \bar{f}_\rho\right\|_\rho \leqslant \lambda^r \left\|\bar{h}_\rho\right\|_\rho = \left\|\bar{h}_\rho\right\|_\rho N^{-\frac{r}{2r+1}}.$$

The fourth term at the right-hand side of (2.13) is estimated by considering

$$\hat{b}_\lambda^D - \bar{b}_\lambda^D = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \int_X f_\rho(x) d\rho_X(x)\right)$$

as the average of N i.i.d. zero-mean random numbers, each of which has the variance

$$\mathrm{Var}\left(y_i - \int_X f_\rho(x) d\rho_X(x)\right) = \mathrm{Var}(y_i) \leqslant \mathbb{E} y_i^2 \leqslant M^2.$$

So

$$\mathbb{E}|\hat{b}_\lambda^D - \bar{b}_\lambda^D| \leqslant \sqrt{\mathrm{Var}(y_i)/N} \leqslant \frac{M}{\sqrt{N}} \leqslant M N^{-\frac{r}{2r+1}}.$$

We summarize the above analysis and complete the proof by letting

$$C = 4C_1^{'} + 33 C_1^* + \|\bar{h}_\rho\|_\rho + M.$$

□

2.4 A Simulation under Gaussian RBF kernel

Theorem 2 in [43] tells us that the RKHS generated by the Gaussian RBF kernel

$$K(s,t) = \exp\left(-\frac{(s-t)^2}{2}\right)$$

contains no non-zero polynomial on X, including the non-zero constant. This phenomenon inspires us to consider the case where the regression funcion is added by a constant.

To verify the intuition, we run a simulation under the gaussian RBF kernel as following. First, we generated a regression function f_ρ using the empirical feature methods by Remark 1 in [30] by 3000 points uniformly distributed on $[0,1]$. The samples $\{x_i\}_{i=1}^N$ are generated randomly according to the uniform distribution on $[0,1]$. We set the output y_i as

$$y_i = f_\rho(x_i) + \epsilon_i$$

where noise ϵ_i is Gaussian with variance $\sigma^2 = 0.01$ and is independent of all $\{x_i\}_{i=1}^N$.

Then we use another regression function

$$f_\rho^{\text{const}} = f_\rho + 10$$

and the same inputs $\{x_i\}_{i=1}^N$ and the outputs

$$y_i^{\text{const}} = y_i + 10.$$

The errors of each case is defined as

$$\text{error} = \left(\frac{1}{3000}\sum_{j=1}^{3000}(f_\lambda^D(x_j^{\text{error}}) - f_\rho(x_j^{\text{error}}))^2\right)^{1/2}$$

or

$$\text{error.const} = \left(\frac{1}{3000}\sum_{j=1}^{3000}(f_\lambda^{D.const}(x_j^{\text{error}}) - f_\rho(x_j^{\text{error}}))^2\right)^{1/2},$$

respectively, to arrpoximate the $L^2_{\rho_X}(X)$ norm with $\{x_j^{\text{error}}\}_{j=1}^{3000}$ generated by the uniform distribution on $[0,1]$. Moreover, we take the regularization parameter $\lambda = N^{-4/5}$. The results are displayed in the following table.

N	error	error.const
100	0.06679019	0.3580925
300	0.05064902	0.2502581
500	0.03951036	0.2118137
1000	0.03089287	0.1768217

Through the simulation results, we find that the convergence rate becomes much slower when the regression function is replaced by the original one plus a constant.

Chapter 3

Distributed KRR

In this chapter, we get convergent results for the output function of distributed kernel ridge regression with high probability and get the almost sure convergence for distributed KRR as a result of the Borel-Cantelli Lemma.

3.1 The Convergence with High Probability of DKRR

Let $D = \{(x_i, y_i)\}_{i=1}^{|D|}$ be a labeled training sample. For the purpose of distributed learning, we divide D evenly into m disjoint subsets $D = \bigcup_{j=1}^{m} D_j$ (so that $D_i \bigcap D_j = \varnothing$ whenever $i \neq j$). Without loss of (too much) generality, in this thesis, we assume $|D_1| = ... = |D_m|$. Recall the KRR on a single machine

$$f_\lambda^{D_j} = \arg\min_{f \in \mathcal{H}_K} \frac{1}{|D_j|} \sum_{(x,y)\in D_j} (f(x) - y)^2 + \lambda \|f\|_K^2$$

$$= \left(L_K^{D_j(x)} + \lambda I\right)^{-1} \frac{1}{|D_j|} \sum_{(x,y)\in D_j} y K_x.$$

To approximate the regression function f_ρ, we synthesize these output functions by

$$\tilde{f}_\lambda^D = \sum_{j=1}^{m} \frac{|D_j|}{|D|} f_\lambda^{D_j}.$$

The convergent results with high probability for DKRR could be described as the following Theorem and its Corollary.

Theorem 3.1. *Assume that*

$$f_\rho = L_K^r h_\rho, \qquad \text{for some } h_\rho \in L^2_{\rho_X}(X) \text{ and } 1/2 \leqslant r \leqslant 1,$$

and that $|y| \leqslant M$ *almost surely. For* $0 < \delta < \frac{1}{20m}$*, the following estimate holds true with confidence* $1 - \delta$*.*

$$\left\| \tilde{f}_\lambda^D - f_\rho \right\|_\rho \leqslant \tilde{C}_1 \max_{1 \leqslant j \leqslant m} \left\{ \left(1 + \frac{\mathcal{A}_{D_j,\lambda}}{\sqrt{\lambda}}\right) \frac{\mathcal{A}^2_{D_j,\lambda}}{\sqrt{\lambda}} \log^3(16m/\delta) \right\}$$

$$+ \tilde{C}_2 \left(1 + \frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}^2_{D,\lambda}}{\lambda}\right) \mathcal{A}_{D,\lambda} \log^3(20/\delta) + \|h_\rho\|_\rho \lambda^r, \qquad (3.1)$$

where $\tilde{C}_1$ *and* $\tilde{C}_2$ *are constants independent of* $m, |D|, |D_j|, \lambda,$ *or* δ*, and they will be specified in the proof.*

Corollary 3.1. *Let* $|y| \leqslant M$ *almost surely and*

$$\mathcal{N}_{L_K}(\lambda) \leqslant C_0 \lambda^{-s}, \qquad \text{for some } s > 0.$$

Moreover, assume that

$$f_\rho = L_K^r h_\rho, \qquad \text{for some } h_\rho \in L^2_{\rho_X}(X) \text{ and } 1/2 < r \leqslant 1,$$

$|D_1| = |D_2| = \ldots = |D_m|$*, and*

$$m \leqslant \frac{|D|^{\frac{r-1/2}{2r+s}}}{\log^3 |D| + 1}. \qquad (3.2)$$

If we take $\lambda = |D|^{-\frac{1}{2r+s}}$*, then for any* $0 < \delta < \frac{1}{20m}$*, we have*

$$\|\tilde{f}_\lambda^D - f_\rho\|_\rho \leqslant \tilde{C} |D|^{-\frac{r}{2r+s}} \log^3 \frac{20}{\delta}$$

with probability at least $1 - \delta$*, where* $\tilde{C}$ *is a constant independent of* $m, |D|, |D_j|, \lambda,$ *or* δ*, and it will be specified in the proof.*

3.2 Error Analysis for DKRR

Introduce the sample free analogue of f_λ^D

$$\begin{aligned} f_\lambda &:= \arg\min_{f\in\mathcal{H}_K}\left\{\int_{X\times Y}(f(x)-y)^2\,d\rho+\lambda\|f\|_K^2\right\} \\ &= (L_K+\lambda I)^{-1}L_K f_\rho. \end{aligned}$$

We decompose the difference between $\tilde{f}_\lambda^D$ and f_ρ as

$$\tilde{f}_\lambda^D - f_\rho = \tilde{f}_\lambda^D - f_\lambda^D + f_\lambda^D - f_\lambda + f_\lambda - f_\rho.$$

The approximation error $\|f_\lambda - f_\rho\|_\rho$ has already been bounded in Lemma 2.3. Now we are going to bound the sampling error $\|\tilde{f}_\lambda^D - f_\lambda\|_\rho$ for distributed KRR.

Let

$$Q_{D(x)} = (L_K^{D(x)}+\lambda I)^{-1} - (L_K+\lambda I)^{-1}. \tag{3.3}$$

Moreover, by first order decomposition (2.30), we have

$$Q_{D(x)} = (L_K+\lambda I)^{-1}\left(L_K - L_K^{D(x)}\right)\left(L_K^{D(x)}+\lambda I\right)^{-1}, \tag{3.4}$$

and by the second order decomposition (2.31), we have

$$\begin{aligned} Q_{D(x)} &= (L_K+\lambda I)^{-1}\left(L_K - L_K^{D(x)}\right)(L_K+\lambda I)^{-1} \\ &\quad + (L_K+\lambda I)^{-1}\left(L_K - L_K^{D(x)}\right)\left(L_K^{D(x)}+\lambda I\right)^{-1}\left(L_K - L_K^{D(x)}\right)(L_K+\lambda I)^{-1}. \end{aligned} \tag{3.5}$$

Let

$$\Delta_j^{'} = \frac{1}{|D_j|} S_{D_j}^T \mathbf{y}_j - L_K^{D_j(x)} f_\rho = \frac{1}{|D_j|} \sum_{(x,y)\in D_j} \left(y - f_\rho(x)\right) K_x,$$

$$\Delta_D^{'} = \frac{1}{|D|} S_D^T \mathbf{y} - L_K^{D(x)} f_\rho = \frac{1}{|D|} \sum_{(x,y)\in D} \left(y - f_\rho(x)\right) K_x = \sum_{j=1}^m \frac{|D_j|}{|D|} \Delta_j^{'},$$

$$\Delta_j^{''} = \left(L_K^{D_j(x)} - L_K\right)(f_\rho - f_\lambda),$$

$$\Delta_D^{''} = \left(L_K^{D(x)} - L_K\right)(f_\rho - f_\lambda) = \sum_{j=1}^m \frac{|D_j|}{|D|} \Delta_j^{''},$$

$$\Delta_j = \Delta_j^{'} + \Delta_j^{''},$$

$$\Delta_D = \Delta_D^{'} + \Delta_D^{''} = \sum_{j=1}^m \frac{|D_j|}{|D|} \Delta_j.$$

Then we have the following error decomposition.

Lemma 3.1 (Lin et al. [40]). *If $\mathbb{E}y^2 < +\infty$, then we have*

$$\begin{aligned} \tilde{f}_\lambda^D - f_\lambda^D &= \sum_{j=1}^m \frac{|D_j|}{|D|} \left[\left(L_K^{D_j(x)} + \lambda I\right)^{-1} - \left(L_K^{D(x)} + \lambda I\right)^{-1} \right] \Delta_j \\ &= \sum_{j=1}^m \frac{|D_j|}{|D|} Q_{D_j(x)} \Delta_j - Q_{D(x)} \Delta_D \\ &= \sum_{j=1}^m \frac{|D_j|}{|D|} Q_{D_j(x)} \Delta_j^{'} + \sum_{j=1}^m \frac{|D_j|}{|D|} Q_{D_j(x)} \Delta_j^{''} - Q_{D(x)} \Delta_D, \end{aligned} \tag{3.6}$$

and

$$\begin{aligned} f_\lambda^D - f_\lambda &= \left(L_K^{D(x)} + \lambda I\right)^{-1} \Delta_D \\ &= Q_{D(x)} \Delta_D + (L_K + \lambda I)^{-1} \Delta_D. \end{aligned} \tag{3.7}$$

Proposition 3.1. *Assume $|y| \leqslant M$ almost surely. For any $\lambda > 0$, we have*

$$\left\|\tilde{f}_\lambda^D - f_\lambda^D\right\|_\rho \leqslant \max_{1\leqslant j\leqslant m}\left\{\left(\frac{\mathcal{S}_{D_j,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}^2_{D_j,\lambda}}{\lambda}\right)\left(\mathcal{R}_{D_j,\lambda,f_\rho} + \mathcal{P}_{D_j,\lambda} + \mathcal{R}_{D_j,\lambda,f_\lambda - f_\rho}\right)\right\}$$

$$+ \left(\frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}^2_{D,\lambda}}{\lambda}\right)\left(\mathcal{R}_{D,\lambda,f_\rho} + \mathcal{P}_{D,\lambda} + \mathcal{R}_{D,\lambda,f_\lambda - f_\rho}\right), \tag{3.8}$$

where

$$\mathcal{R}_{D,\lambda,g} := \left\|(L_K + \lambda I)^{-1/2}(L_K - L_K^{D(x)})g\right\|_K, \qquad \text{for any } g \text{ in } \mathcal{H}_K,$$

and $\mathcal{P}_{D,\lambda}$, $\mathcal{Q}_{D,\lambda}$, $\mathcal{S}_{D,\lambda}$ are defined by (2.18), (2.19) and (2.20), respectively.

Proof. Using Lemma 3.1, we bound each term in the right-hand side of (3.6), respectively.

To bound $Q_{D(x)}\Delta_D^{'}$, use the second order decomposition (3.5) techniques in [40] and we get

$$\left\|Q_{D(x)}\Delta_D^{'}\right\|_\rho = \left\|L_K^{1/2}Q_{D(x)}\Delta_D^{'}\right\|_K = \left\|L_K^{1/2}\left((L_K^{D(x)} + \lambda I)^{-1} - (L_K + \lambda I)^{-1}\right)\Delta_D^{'}\right\|_K$$

$$\leqslant \left\|L_K^{1/2}\left(L_K + \lambda I\right)^{-1}\left(L_K - L_K^{D(x)}\right)\left(L_K + \lambda I\right)^{-1}\Delta_D^{'}\right\|_K$$

$$+ \left\|L_K^{1/2}\left(L_K + \lambda I\right)^{-1}\left(L_K - L_K^{D(x)}\right)\left(L_K^{D(x)} + \lambda I\right)^{-1}\left(L_K - L_K^{D(x)}\right)\left(L_K + \lambda I\right)^{-1}\Delta_D^{'}\right\|_K$$

$$\leqslant \left\|L_K^{1/2}\left(L_K + \lambda I\right)^{-1/2}\right\|_{\mathrm{op}(K)}\mathcal{S}_{D,\lambda}\left\|\left(L_K + \lambda I\right)^{-1/2}\right\|_{\mathrm{op}(K)}\left\|\left(L_K + \lambda I\right)^{-1/2}\Delta_D^{'}\right\|_K$$

$$+ \left\|L_K^{1/2}\left(L_K + \lambda I\right)^{-1/2}\right\|_{\mathrm{op}(K)}\mathcal{S}_{D,\lambda}\left\|\left(L_K^{D(x)} + \lambda I\right)^{-1}\right\|_{\mathrm{op}(K)}\mathcal{S}_{D,\lambda}\left\|\left(L_K + \lambda I\right)^{-1/2}\Delta_D^{'}\right\|_K$$

$$\leqslant \frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}}\left\|\left(L_K + \lambda I\right)^{-1/2}\Delta_D^{'}\right\|_K + \frac{\mathcal{S}^2_{D,\lambda}}{\lambda}\left\|\left(L_K + \lambda I\right)^{-1/2}\Delta_D^{'}\right\|_K$$

$$= \left(\frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}^2_{D,\lambda}}{\lambda}\right)\left\|\left(L_K + \lambda I\right)^{-1/2}\Delta_D^{'}\right\|_K, \tag{3.9}$$

where we have used the following properties since both L_K and $L_K^{D(x)}$ are positive semi-difinite,

$$\left\| L_K^{1/2} \left(L_K + \lambda I\right)^{-1/2} \right\|_{\mathrm{op}(K)} \leqslant 1,$$

$$\left\| (L_K + \lambda I)^{-1/2} \right\|_{\mathrm{op}(K)} \leqslant \frac{1}{\sqrt{\lambda}},$$

$$\left\| \left(L_K^{D(x)} + \lambda I \right)^{-1} \right\|_{\mathrm{op}(K)} \leqslant \frac{1}{\lambda}.$$

Moreover, since

$$\Delta_D^{'} = \frac{1}{|D|} S_D^T \mathbf{y} - L_K^{D(x)} f_\rho = \frac{1}{|D|} S_D^T \mathbf{y} - L_K f_\rho + L_K f_\rho - L_K^{D(x)} f_\rho,$$

we obtain

$$\begin{aligned} \left\| (L_K + \lambda I)^{-1/2} \Delta_D^{'} \right\|_K &\leqslant \left\| (L_K + \lambda I)^{-1/2} (L_K - L_K^{D(x)}) f_\rho \right\|_K \\ &\quad + \left\| (L_K + \lambda I)^{-1/2} (L_K f_\rho - \frac{1}{|D|} S_D^T \mathbf{y}) \right\|_K \\ &= \mathcal{R}_{D,\lambda,f_\rho} + \mathcal{P}_{D,\lambda}. \end{aligned}$$

Thus

$$\left\| Q_{D(x)} \Delta_D^{'} \right\|_\rho \leqslant \left(\frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}_{D,\lambda}^2}{\lambda} \right) \left(\mathcal{R}_{D,\lambda,f_\rho} + \mathcal{P}_{D,\lambda} \right). \tag{3.10}$$

Similarly, we have

$$\begin{aligned} \left\| Q_{D(x)} \Delta_D^{''} \right\|_\rho = \left\| L_K^{1/2} Q_{D(x)} \Delta_D^{''} \right\|_K &\leqslant \left(\frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}_{D,\lambda}^2}{\lambda} \right) \left\| (L_K + \lambda I)^{-1/2} \Delta_D^{''} \right\|_K \\ &= \left(\frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}_{D,\lambda}^2}{\lambda} \right) \left\| (L_K + \lambda I)^{-1/2} (L_K - L_K^{D(x)}) (f_\lambda - f_\rho) \right\|_K \\ &= \left(\frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}_{D,\lambda}^2}{\lambda} \right) \mathcal{R}_{D,\lambda,f_\lambda - f_\rho}. \end{aligned} \tag{3.11}$$

Note that

$$\Delta_D = \Delta_D^{'} + \Delta_D^{''}.$$

By (3.10) and (3.11), we have

$$\begin{aligned}\left\|Q_{D(x)}\Delta_D\right\|_\rho &\leqslant \left\|Q_{D(x)}\Delta_D^{'}\right\|_\rho + \left\|Q_{D(x)}\Delta_D^{''}\right\|_\rho \\ &\leqslant \left(\frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}_{D,\lambda}^2}{\lambda}\right)\left(\mathcal{R}_{D,\lambda,f_\rho} + \mathcal{P}_{D,\lambda} + \mathcal{R}_{D,\lambda,f_\lambda - f_\rho}\right). \qquad (3.12)\end{aligned}$$

The same is true if we replace D with D_j, i.e.,

$$\left\|Q_{D_j(x)}\Delta_j\right\|_\rho \leqslant \left(\frac{\mathcal{S}_{D_j,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}_{D_j,\lambda}^2}{\lambda}\right)\left(\mathcal{R}_{D_j,\lambda,f_\rho} + \mathcal{P}_{D_j,\lambda} + \mathcal{R}_{D_j,\lambda,f_\lambda - f_\rho}\right). \qquad (3.13)$$

Thus by Lemma 3.1, (3.10), (3.11), (3.12) and (3.13) , we have

$$\begin{aligned}\left\|\tilde{f}_\lambda^D - f_\lambda^D\right\|_\rho &= \left\|L_K^{1/2}\left(\tilde{f}_\lambda^D - f_\lambda^D\right)\right\|_K \\ &\leqslant \sum_{j=1}^m \frac{|D_j|}{|D|}\left(\frac{\mathcal{S}_{D_j,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}_{D_j,\lambda}^2}{\lambda}\right)\left(\mathcal{R}_{D_j,\lambda,f_\rho} + \mathcal{P}_{D_j,\lambda} + \mathcal{R}_{D_j,\lambda,f_\lambda - f_\rho}\right) \\ &+ \left(\frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}_{D,\lambda}^2}{\lambda}\right)\left(\mathcal{R}_{D,\lambda,f_\rho} + \mathcal{P}_{D,\lambda} + \mathcal{R}_{D,\lambda,f_\lambda - f_\rho}\right) \\ &\leqslant \max_{1\leqslant j\leqslant m}\left(\frac{\mathcal{S}_{D_j,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}_{D_j,\lambda}^2}{\lambda}\right)\left(\mathcal{R}_{D_j,\lambda,f_\rho} + \mathcal{P}_{D_j,\lambda} + \mathcal{R}_{D_j,\lambda,f_\lambda - f_\rho}\right) \\ &+ \left(\frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}_{D,\lambda}^2}{\lambda}\right)\left(\mathcal{R}_{D,\lambda,f_\rho} + \mathcal{P}_{D,\lambda} + \mathcal{R}_{D,\lambda,f_\lambda - f_\rho}\right).\end{aligned}$$

□

The estimation of $\mathcal{P}_{D,\lambda}$ and $\mathcal{S}_{D,\lambda}$ has already been given in Lemma 2.5. Estimating $\mathcal{R}_{D,\lambda,g}$ is similar and was given in [40].

Lemma 3.2 (Lin et al. [40], Lemma 18)**.** *For any $0 < \delta < 1$, it holds with probability at least $1 - \delta$ that*

$$\mathcal{R}_{D,\lambda,g} \leqslant 2\|g\|_\infty (\kappa + 1)\,\mathcal{A}_{D,\lambda}\log(2/\delta). \qquad (3.14)$$

Proof of Theorem 3.1. We use Proposition 3.1, Lemma 2.4, and Lemma 2.3 to obtain

$$\begin{aligned}\left\|\tilde{f}_\lambda^D - f_\rho\right\|_\rho &\leqslant \left\|\tilde{f}_\lambda^D - f_\lambda^D\right\|_\rho + \left\|f_\lambda^D - f_\lambda\right\|_\rho + \|f_\lambda - f_\rho\|_\rho \\ &\leqslant J_1^{'} + J_2^{'}, \end{aligned} \tag{3.15}$$

where

$$\begin{aligned} J_1^{'} &= \max_{1\leqslant j\leqslant m}\left\{\left(\frac{\mathcal{S}_{D_j,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}^2_{D_j,\lambda}}{\lambda}\right)\left(\mathcal{R}_{D_j,\lambda,f_\rho} + \mathcal{P}_{D_j,\lambda} + \mathcal{R}_{D_j,\lambda,f_\lambda - f_\rho}\right)\right\} \\ J_2^{'} &= \left(\frac{\mathcal{S}_{D,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}^2_{D,\lambda}}{\lambda}\right)\left(\mathcal{R}_{D,\lambda,f_\rho} + \mathcal{P}_{D,\lambda} + \mathcal{R}_{D,\lambda,f_\lambda - f_\rho}\right) \\ &\quad + \mathcal{Q}^2_{D,\lambda}(\mathcal{P}_{D,\lambda} + \mathcal{S}_{D,\lambda}\|f_\lambda\|_K) + \lambda^r\|h_\rho\|_\rho. \end{aligned}$$

Since $1/2 \leqslant r \leqslant 1$, we use the estimate $\|f_\lambda\|_K \leqslant \kappa^{2r-1}\|h_\rho\|_\rho$ from (2.21). With the assumption that $|y| \leqslant M$ almost surely, we have

$$\|f_\rho\|_\infty = \operatorname*{ess\,sup}_{x\in X}\left|\int_Y y d\rho(y|x)\right| \leqslant M,$$

so (recall (1.2)),

$$\|f_\lambda - f_\rho\|_\infty \leqslant \kappa\|f_\lambda\|_K + \|f_\rho\|_\infty \leqslant \kappa^{2r}\|h_\rho\|_\rho + M.$$

With the above estimates, Lemma 2.5, and Lemma 3.2, we have that with probability at least $1 - 4m\delta$,

$$\begin{aligned} J_1^{'} \leqslant \max_{1\leqslant j\leqslant m}\Bigg\{&\left(\frac{2(\kappa^2+\kappa)}{\sqrt{\lambda}}\mathcal{A}_{D_j,\lambda}\log\frac{2}{\delta} + \frac{4(\kappa^2+\kappa)^2}{\lambda}\mathcal{A}^2_{D_j,\lambda}\log^2\frac{2}{\delta}\right) \times \\ &\left(2\|f_\rho\|_\infty(\kappa+1)\mathcal{A}_{D_j,\lambda}\log\frac{2}{\delta} + 2\|f_\lambda - f_\rho\|_\infty(\kappa+1)\mathcal{A}_{D_j,\lambda}\log\frac{2}{\delta} + 2M(\kappa+1)\mathcal{A}_{D_j,\lambda}\log\frac{2}{\delta}\right)\Bigg\}. \end{aligned}$$

We scale δ to $\frac{\delta}{8m}$ to see that with confidence $1 - \frac{\delta}{2}$,

$$J_1^{'} \leqslant \tilde{C}_1 \max_{1\leqslant j\leqslant m}\left\{\left(1 + \frac{\mathcal{A}_{D_j,\lambda}}{\sqrt{\lambda}}\right)\frac{\mathcal{A}^2_{D_j,\lambda}}{\sqrt{\lambda}}\log^3\frac{16m}{\delta}\right\}, \tag{3.16}$$

with

$$\tilde{C}_1 = 2(\kappa^2+\kappa)(1+2(\kappa^2+\kappa))(2M(\kappa+1)+2(\kappa+1)(\kappa^{2r}\|h_\rho\|_\rho+M)+2M(\kappa+1)).$$

Similarly, with probability at least $1-5\delta$,

$$\begin{aligned} J_2' \leqslant & \left(\frac{2(\kappa^2+\kappa)}{\sqrt{\lambda}}\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}+\frac{4(\kappa^2+\kappa)^2}{\lambda}\mathcal{A}_{D,\lambda}^2\log^2\frac{2}{\delta}\right) \\ & \times\left(2\|f_\rho\|_\infty(\kappa+1)\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}+2\|f_\lambda-f_\rho\|_\infty(\kappa+1)\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}+2M(\kappa+1)\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}\right) \\ & +\left(2\left(\frac{2(\kappa^2+\kappa)}{\sqrt{\lambda}}\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}\right)^2+2\right)\left(2M(\kappa+1)\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}\right. \\ & \left.+2\|f_\lambda\|_K(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}\right)+\lambda^r\|h_\rho\|_\rho. \end{aligned}$$

We scale δ to $\frac{\delta}{10}$ to see that with confidence $1-\frac{\delta}{2}$,

$$J_2' \leqslant \tilde{C}_2\left(1+\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}+\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}\right)\mathcal{A}_{D,\lambda}\log^3\frac{20}{\delta}+\lambda^r\|h_\rho\|_\rho, \tag{3.17}$$

with

$$\tilde{C}_2=\max\{\tilde{C}_1,(2+2(2(\kappa^2+\kappa))^2)(2M(\kappa+1)+2(\kappa^2+\kappa)\kappa^{2r-1}\|h_\rho\|_\rho)\}.$$

□

Proof of Corollary 3.1. By taking

$$\lambda=|D|^{-\frac{1}{2r+s}}$$

and the assumption

$$\mathcal{N}_{L_K}(\lambda)\leqslant C_0\lambda^{-s},$$

we have

$$\mathcal{A}_{D_j,\lambda}\leqslant\frac{m}{|D|\sqrt{\lambda}}+\sqrt{\frac{C_0m}{|D|\lambda^s}}=m|D|^{-\frac{2r+s-1/2}{2r+s}}+\sqrt{C_0m}|D|^{-\frac{r}{2r+s}},$$

and

$$\mathcal{A}_{D,\lambda} \leqslant |D|^{-\frac{4r+2s-1}{4r+2s}} + \sqrt{C_0}|D|^{-\frac{r}{2r+s}} \leqslant (1+\sqrt{C_0})|D|^{-\frac{r}{2r+s}}. \tag{3.18}$$

Since (3.2) and $r > 1/2$, we obtain

$$m|D|^{-\frac{2r+s-1/2}{2r+s}} \leqslant \sqrt{m}|D|^{-\frac{r}{2r+s}}.$$

As a result,

$$\mathcal{A}_{D_j,\lambda} \leqslant (\sqrt{C_0}+1)\sqrt{m}|D|^{-\frac{r}{2r+s}} \tag{3.19}$$

and

$$\frac{\mathcal{A}_{D_j,\lambda}}{\sqrt{\lambda}} \leqslant (\sqrt{C_0}+1)\sqrt{m}|D|^{-\frac{r-1/2}{2r+s}} \leqslant \sqrt{C_0}+1. \tag{3.20}$$

Also we have

$$1 + \frac{\mathcal{A}_{D_j,\lambda}}{\sqrt{\lambda}} \leqslant \sqrt{C_0}+2.$$

Since for $0 < \delta < \frac{1}{20m}$, $\log(16/\delta) > 1$, by the inequality

$$(a+b)^3 \leqslant 8(\max\{a,b\})^3 \leqslant 8(a^3+b^3), \text{ for } a, b > 0,$$

we have

$$\begin{aligned}\log^3(16m/\delta) &\leqslant 8\left(\log^3(16/\delta) + \log^3 m\right)\\ &\leqslant 8\log^3(16/\delta)(\log^3 m + 1) \leqslant 8\log^3(16/\delta)(\log^3|D|+1).\end{aligned} \tag{3.21}$$

Substitute (3.2), (3.19), (3.20), (3.18) and (3.21) into (3.1), we obtain

$$\begin{aligned}\left\|\tilde{f}_\lambda^D - f_\rho\right\|_\rho &\leqslant \tilde{C}_1 \max_{1\leqslant j\leqslant m}\left\{\left(1+\frac{\mathcal{A}_{D_j,\lambda}}{\sqrt{\lambda}}\right)\frac{\mathcal{A}^2_{D_j,\lambda}}{\sqrt{\lambda}}\log^3(16m/\delta)\right\}\\ &\quad + \tilde{C}_2\left(1+\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}+\frac{\mathcal{A}^2_{D,\lambda}}{\lambda}\right)\mathcal{A}_{D,\lambda}\log^3(20/\delta) + \|h_\rho\|_\rho\lambda^r\\ &\leqslant 8\tilde{C}_1\left(\sqrt{C_0}+2\right)(1+\sqrt{C_0})^2|D|^{-\frac{r}{2r+s}}m|D|^{-\frac{r-1/2}{2r+s}}(\log^3|D|+1)\log^3(16/\delta)\\ &\quad + \tilde{C}_2\left(2(\sqrt{C_0}+2)^2\right)(1+\sqrt{C_0})|D|^{-\frac{r}{2r+s}}\log^3(20/\delta) + \|h_\rho\|_\rho|D|^{-\frac{r}{2r+s}}\\ &\leqslant \tilde{C}|D|^{-\frac{r}{2r+s}}\log^3(20/\delta),\end{aligned}$$

with

$$\tilde{C} = \max\left\{8\tilde{C}_1\left(\sqrt{C_0}+2\right)(1+\sqrt{C_0})^2, \tilde{C}_2\left(2(\sqrt{C_0}+2)^2+1)\right)(1+\sqrt{C_0}), \|h_\rho\|_\rho\right\}.$$

□

3.3 The Semi-supervised Learning For DKRR

To analyze the almost sure convergence of the semi-supervised learning scheme, we adopt the method developed in [11] which introduced a new training set as following.

If on each local processor, we have labeled data $D_j = \{(x_i^j, y_i^j)\}_{i=1}^{|D_j|}$ and unlabeled data $\tilde{D}_j(x) = \{\tilde{x}_i^j\}_{i=1}^{|\tilde{D}_j|}$, we introduce a new training set based on $D_j \cup \tilde{D}_j(x)$. Specifically, let $D_j^* = \{(x_i^*, y_i^*)\}_{i=1}^{|D_j^*|}$ with

$$x_i^* = \begin{cases} x_i, & \text{if} \quad 1 \leqslant i \leqslant |D_j|, \\ \tilde{x}_{i-|D_j|}, & \text{if} \quad |D_j|+1 \leqslant i \leqslant |D_j^*|, \end{cases} \quad \text{and} \quad y_i^* = \begin{cases} \frac{|D_j^*|}{|D_j|} y_i, & \text{if} \quad 1 \leqslant i \leqslant |D_j|, \\ 0, & \text{if} \quad |D_j|+1 \leqslant i \leqslant |D_j^*|, \end{cases} \tag{3.22}$$

be the training data set on each machine. The whole training set is written as

$$D^* = \bigcup_{j=1}^m D_j^*.$$

The output function of the KRR based on each training set is

$$f_\lambda^{D_j^*} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{|D_j^*|} \sum_{(x,y)\in D_j^*} (f(x)-y)^2 + \lambda \|f\|_K^2 \right\}.$$

Moreover, we synthesize these functions by

$$\tilde{f}_\lambda^{D^*} = \sum_{j=1}^m \frac{|D_j^*|}{|D^*|} f_\lambda^{D_j^*}.$$

The almost sure convergence of the output function of semi-supervised distributed KRR are given as follows.

Theorem 3.2. *Assume that*

$$f_\rho = L_K^r h_\rho, \qquad \textit{for some } h_\rho \in L^2_{\rho_X}(X) \textit{ and } 1/2 \leqslant r \leqslant 1,$$

and that $|y| \leqslant M$ *almost surely. For* $0 < \delta < \frac{1}{20m}$*, the following estimate holds true with confidence at least* $1-\delta$*.*

$$\begin{aligned} \left\| \tilde{f}_\lambda^{D*} - f_\rho \right\|_\rho \leqslant \tilde{C}_1 \max_{1\leqslant j \leqslant m} \left\{ \left(1 + \frac{\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right) \frac{\mathcal{A}_{D_j,\lambda}\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}} \log^3(16m/\delta) \right\} \\ + \tilde{C}_2 \left(1 + \frac{\mathcal{A}_{D*,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}^2_{D*,\lambda}}{\lambda}\right) \mathcal{A}_{D,\lambda} \log^3(20/\delta) + \|h_\rho\|_\rho \lambda^r, \end{aligned} \tag{3.23}$$

where $\tilde{C}_1$ *and* $\tilde{C}_2$ *are the same constants as in Theorem 3.1.*

Proof. By (3.22), we have

$$\frac{1}{|D^*|} S_{D^*}^T \mathbf{y}^* = \frac{1}{|D^*|} \sum_{(x^*,y^*)\in D^*} y^* K_{x^*} = \frac{1}{|D|} S_D^T \mathbf{y}.$$

Thus (3.9) becomes

$$\left\| L_K^{1/2} \left[Q_{D^*(x)}\right] \Delta'_{D^*} \right\|_K \leqslant \left(\frac{\mathcal{S}_{D^*,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}^2_{D^*,\lambda}}{\lambda} \right) \left\| (L_K + \lambda I)^{-1/2} \Delta'_{D^*} \right\|_K .$$

Moreover,

$$\begin{aligned} \left\| (L_K + \lambda I)^{-1/2} \Delta'_{D^*} \right\|_K &\leqslant \left\| (L_K + \lambda I)^{-1/2} (L_K - L_K^{D^*(x)}) f_\rho \right\|_K \\ &\quad + \left\| (L_K + \lambda I)^{-1/2} (L_K f_\rho - \frac{1}{|D|} S_D^T \mathbf{y}) \right\|_K \\ &= \mathcal{R}_{D^*,\lambda,f_\rho} + \mathcal{P}_{D,\lambda}. \end{aligned}$$

Note that the label of data doesn't appear in Δ''_D. Thus by (3.11), we have

$$\left\| L_K^{1/2} \left[Q_{D^*(x)}\right] \Delta_{D^*} \right\|_K \leqslant \left(\frac{\mathcal{S}_{D^*,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{S}^2_{D^*,\lambda}}{\lambda} \right) (\mathcal{R}_{D^*,\lambda,f_\rho} + \mathcal{P}_{D,\lambda} + \mathcal{R}_{D^*,\lambda,f_\lambda - f_\rho}).$$

Since

$$\mathcal{A}_{D^*,\lambda} \leqslant \mathcal{A}_{D,\lambda}$$

similarly to (3.16), by replacing D^* with D_j^*, we have

$$\left\|\tilde{f}_\lambda^{D^*} - f_\lambda^{D^*}\right\|_\rho \leqslant \tilde{C}_1 \max_{1\leqslant j\leqslant m}\left\{\left(1+\frac{\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)\frac{\mathcal{A}_{D_j^*,\lambda}\mathcal{A}_{D_j,\lambda}}{\sqrt{\lambda}}\log(16m/\delta)\right\}$$

with probability at least $1-\frac{\delta}{2}$. Moreover, similarly to (3.17), we get

$$\|f_\lambda^{D^*} - f_\lambda\|_\rho \leqslant \tilde{C}_2\left(1+\frac{\mathcal{A}_{D^*,\lambda}}{\sqrt{\lambda}}+\frac{\mathcal{A}_{D^*,\lambda}^2}{\lambda}\right)\mathcal{A}_{D,\lambda}\log^3(20/\delta)$$

with probability at least $1-\frac{\delta}{2}$. In conclusion, it holds with confidence $1-\delta$ that

$$\begin{aligned}\left\|\tilde{f}_\lambda^{D} - f_\rho\right\|_\rho \leqslant{}& \tilde{C}_1\left\{\max_{1\leqslant j\leqslant m}\left(1+\frac{\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)\frac{\mathcal{A}_{D_j^*,\lambda}\mathcal{A}_{D_j,\lambda}}{\sqrt{\lambda}}\log^3(16m/\delta)\right\}\\ &+\tilde{C}_2\left(1+\frac{\mathcal{A}_{D^*,\lambda}}{\sqrt{\lambda}}+\frac{\mathcal{A}_{D^*,\lambda}^2}{\lambda}\right)\mathcal{A}_{D,\lambda}\log^3(20/\delta)+\lambda^r\|h_\rho\|_\rho. \end{aligned}\tag{3.24}$$

□

Corollary 3.2. *Let $|y| \leqslant M$ almost surely and*

$$\mathcal{N}_{L_K}(\lambda) \leqslant C_0\lambda^{-s} \text{ for some } s > 0.$$

Moreover, assume that

$$f_\rho = L_K^r h_\rho, \qquad \text{for some } h_\rho \in L^2_{\rho_X}(X) \text{ and } 1/2 < r \leqslant 1,$$

$|D_1| = |D_2| = \ldots = |D_m|$, $|D_1^*| = |D_2^*| = \ldots = |D_m^*|$, *and*

$$m \leqslant \min\left\{\frac{|D^*|^{1/2}|D|^{-\frac{s+1}{4r+2s}}}{\log^3|D|+1}, \frac{|D^*|^{1/3}|D|^{-\frac{2r+s-2}{6r+3s}}}{\log^3|D|+1}, |D|\right\}. \tag{3.25}$$

If we take $\lambda = |D|^{-\frac{1}{2r+s}}$, then for any $0 < \delta < \frac{1}{20m}$, we have

$$\|\tilde{f}_\lambda^{D^*} - f_\rho\|_\rho \leqslant \tilde{C}|D|^{-\frac{r}{2r+s}}\log^3\frac{20}{\delta}$$

with probability at least $1-\delta$, where $\tilde{C}$ is the same constant as in Corollary 3.1 which is independent of $m, |D|, |D_j|, |D^*|, |D_j^*|, \lambda,$ or δ.

Proof. By taking $\lambda = |D|^{-\frac{1}{2r+s}}$, we get

$$\mathcal{A}_{D_j^*,\lambda} \leqslant m|D^*|^{-1}|D|^{\frac{1}{4r+2s}} + \sqrt{C_0 m}|D^*|^{-1/2}|D|^{\frac{s}{4r+2s}},$$

and

$$\mathcal{A}_{D_j,\lambda} \leqslant m|D|^{-\frac{2r+s-1/2}{2r+s}} + \sqrt{C_0 m}|D|^{-\frac{r}{2r+s}}.$$

Moreover, since

$$m \leqslant \min\left\{\frac{|D^*|^{1/2}|D|^{-\frac{s+1}{4r+2s}}}{\log^3|D|+1}, \frac{|D^*|^{1/3}|D|^{-\frac{2r+s-2}{6r+3s}}}{\log^3|D|+1}, |D|\right\},$$

we obtain

$$m|D^*|^{-1}|D|^{\frac{1}{4r+2s}} \leqslant \sqrt{m}|D^*|^{-1/2}|D|^{\frac{s}{4r+2s}}.$$

Thus

$$\mathcal{A}_{D_j^*,\lambda} \leqslant (\sqrt{C_0}+1)\sqrt{m}|D^*|^{-1/2}|D|^{\frac{s}{4r+2s}},$$

$$\frac{\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}} \leqslant (\sqrt{C_0}+1)\sqrt{m}|D^*|^{-1/2}|D|^{\frac{1+s}{4r+2s}}.$$

Note that

$$\mathcal{A}_{D^*,\lambda} \leqslant \mathcal{A}_{D,\lambda}, \qquad \frac{\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}} \leqslant \sqrt{C_0}+1$$

and

$$\begin{aligned}\log^3(16m/\delta) &\leqslant 8\left(\log^3(16/\delta) + \log^3 m\right)\\ &\leqslant 8\log^3(16/\delta)(\log^3 m + 1) \leqslant 8\log^3(16/\delta)(\log^3|D|+1).\end{aligned}$$

Thus we obtain

$$\begin{aligned}\left\|\tilde{f}_{\lambda}^{D*}-f_{\rho}\right\|_{\rho} &\leqslant \tilde{C}_1\left\{\max_{1\leqslant j\leqslant m}\left(1+\frac{\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)\frac{\mathcal{A}_{D_j^*,\lambda}\mathcal{A}_{D_j,\lambda}}{\sqrt{\lambda}}\log^3(16m/\delta)\right\}\\
&+\tilde{C}_2\left(1+\frac{\mathcal{A}_{D*,\lambda}}{\sqrt{\lambda}}+\frac{\mathcal{A}_{D*,\lambda}^2}{\lambda}\right)\mathcal{A}_{D,\lambda}\log^3(20/\delta)+\lambda^r\|h_\rho\|_\rho\\
&\leqslant 8\tilde{C}_1\left(\sqrt{C}_0+2\right)\left(1+\sqrt{C_0}\right)\left(m\sqrt{m}|D|^{-\frac{2r+s-1/2}{2r+s}}|D^*|^{-1/2}|D|^{\frac{s+1}{4r+2s}}\right.\\
&\left.+\sqrt{C_0}m|D^*|^{-1/2}|D|^{\frac{1+s}{4r+2s}}|D|^{-\frac{r}{2r+s}}\right)(\log^3|D|+1)\log^3(16/\delta)\\
&+\tilde{C}_2\left(2\left(\sqrt{C_0}+2\right)^2\right)\left(1+\sqrt{C_0}\right)|D|^{-\frac{r}{2r+s}}\log^3(20/\delta)+|D|^{-\frac{r}{2r+s}}\\
&\leqslant \tilde{C}|D|^{-\frac{r}{2r+s}}\log^3(20/\delta).\end{aligned}$$

□

3.4 The Almost Sure Convergence of DKRR

Based on the convergent results with high probability in the previous section, in this section, we utilize the Borel-Cantelli Lemma to get the almost sure convergence of distributed kernel ridge regression whose convergence rate could be arbitrarily close to the optimal minimax rate. The tenique of using the Borel-Cantelli Lemma to get the almost sure convergence via the convergence with high probability is adopted in [41].

Recall the Borel-Cantelli Lemma.

Lemma 3.3 (Borel-Cantelli)**.** *Let $\{\xi_n\}$ be a sequence of random variables and $\{\mu_n\}$ be a sequence satisfying $\lim_{n\to\infty}\mu_n=0$. If*

$$\sum_{n=1}^{\infty}\mathbb{P}[|\xi_n-\xi|>\mu_n]<\infty,$$

then $\xi_n\to\xi$ almost surely.

Using this Lemma, we get the following almost sure convergence.

Corollary 3.3. *Assume that* $|y| \leqslant M$ *almost surely and*

$$\mathcal{N}_{L_K}(\lambda) \leqslant C_0 \lambda^{-s}.$$

Moreover, assume that

$$f_\rho = L_K^r h_\rho, \qquad \text{for some } h_\rho \in L^2_{\rho_X}(X) \text{ and } 1/2 < r \leqslant 1.$$

If $|D_1| = |D_2| = ... = |D_m|$ *and (3.2) holds, by taking*

$$\lambda = |D|^{-\frac{1}{2r+s}},$$

we have

$$\lim_{|D| \to +\infty} |D|^{\frac{r(1-\epsilon)}{2r+s}} \|\tilde{f}_\lambda^D - f_\rho\|_\rho = 0 \tag{3.26}$$

almost surely for arbitrary $\epsilon > 0$.

Furthermore, if $|D_1| = |D_2| = ... = |D_m|, |D_1^*| = |D_2^*| = ... = |D_m^*|$ *and (3.25) holds, we also have*

$$\lim_{|D| \to +\infty} |D|^{\frac{r(1-\epsilon)}{2r+s}} \|\tilde{f}_\lambda^{D*} - f_\rho\|_\rho = 0 \tag{3.27}$$

almost surely for arbitrary $\epsilon > 0$.

Proof. By Theorem 3.1, for $N = |D|$ and $\delta = N^{-2}$, we have

$$\mathbb{P}\left[N^{-\frac{r(1-\epsilon)}{2r+s}} \|\tilde{f}_\lambda^D - f_\rho\|_\rho > \tilde{C} N^{-\frac{r\epsilon}{2r+s}} \left(\log \frac{20}{N^{-2}}\right)\right] \leqslant N^{-2}.$$

Thus

$$\sum_{N=1}^{\infty} \mathbb{P}\left[N^{-\frac{r(1-\epsilon)}{2r+s}} \|\tilde{f}_\lambda^D - f_\rho\|_\rho > \tilde{C} N^{-\frac{r\epsilon}{2r+s}} \left(\log \frac{20}{N^{-2}}\right)\right] \leqslant \sum_{N=1}^{\infty} N^{-2} < \infty.$$

By Borel-Cantelli's Lemma, we get (3.26).

Similarly, by Theorem 3.2, we obtain (3.27) . □

Bibliography

[1] Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *The Journal of Machine Learning Research*, 15(1):1653–1674, 2014.

[2] Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, and Dan Roth. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6(Apr):393–425, 2005.

[3] Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(Feb):441–474, 2009.

[4] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[5] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781, 2013.

[6] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[7] G. Blanchard, P. Massart, R. Vert, and L. Zwald. Kernel projection machine: a new tool for pattern recognition. *Proc. NIPS*, pages 1649–1656, 2004.

[8] G. Blanchard and L. Zwald. Finite-dimensional projection for classification and statistical learning. *IEEE Transactions on Information Theory*, 54:4169–4182, 2008.

[9] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[10] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[11] Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *J. Mach. Learn. Res*, 2017.

[12] Di-Rong Chen, Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5(Sep):1143–1175, 2004.

[13] Sungmoon Cheong, Sang Hoon Oh, and Soo-Young Lee. Support vector machines with binary tree architecture for multi-class classification. *Neural Information Processing-Letters and Reviews*, 2(3):47–51, 2004.

[14] Andreas Christmann and Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, pages 799–819, 2007.

[15] Andreas Christmann and Ingo Steinwart. How svms can estimate quantiles and the median. In *Advances in neural information processing systems*, pages 305–312, 2008.

[16] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, pages 844–874, 2008.

[17] Stéphan Clémençon and Nicolas Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8(Dec):2671–2699, 2007.

[18] Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. In *Advances in neural information processing systems*, pages 313–320, 2004.

[19] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[20] David Cossock and Tong Zhang. Subset ranking using regression. In *International Conference on Computational Learning Theory*, pages 605–619. Springer, 2006.

[21] David Cossock and Tong Zhang. Statistical analysis of bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11):5140–5154, 2008.

[22] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[23] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

[24] Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

[25] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[26] Aymeric Dieuleveut, Francis Bach, et al. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

[27] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[28] Xin Guo, Jun Fan, and Ding-Xuan Zhou. Sparsity and error analysis of empirical feature-based regularization schemes. *Journal of Machine Learning Research*, 17:1–34, 2016.

[29] Xin Guo, Ting Hu, and Qiang Wu. Centered reproducing kernel for variable and interaction selection. *Manuscript in Preparation*, 2018.

[30] Xin Guo and Ding-Xuan Zhou. An empirical feature-based learning algorithm producing sparse approximations. *Applied and Computational Harmonic Analysis*, 32:389–400, 2012.

[31] Zheng Chu Guo, Shao Bo Lin, and Ding Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 2017.

[32] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[33] A. J. Hoffman and H. W. Wielandt. The vatration of the spectrum of a normal matrix. *Duke Mathematical Journal*, 20:37–39, 1953.

[34] Ting Hu, Dao-Hong Xiang, and Ding-Xuan Zhou. Online learning for quantile regression and support vector regression. *Journal of Statistical Planning and Inference*, 142(12):3107–3122, 2012.

[35] Peter J. Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, pages 73–101, 1964.

[36] Changha Hwang and Jooyong Shim. A simple quantile regression via support vector machine. In *International Conference on Natural Computation*, pages 512–520. Springer, 2005.

[37] BA J Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Phil. Trans. R. Soc. Lond. A*, 209(441-458):415–446, 1909.

[38] Jyrki Kivinen, Alex J Smola, and Robert C Williamson. Online learning with kernels. In *Advances in neural information processing systems*, pages 785–792, 2002.

[39] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[40] Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

[41] Shao-Bo Lin and Ding -Xuan Zhou. Distributed kernel-based gradient descent algorithms. *Constructive Approximation*, pages 1–28, 2017.

[42] Gábor Lugosi and Nicolas Vayatis. On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, pages 30–55, 2004.

[43] Ha Quang Minh. Some properties of gaussian reproducing kernel hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2010.

[44] T.M. Mitchell. *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill, 1997.

[45] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

[46] Tomaso Poggio and Steve Smale. The mathematics of learning: Dealing with data. *Notices of the AMS*, 50(5):537–544, 2003.

[47] Wojciech Rejchel. On ranking and generalization bounds. *Journal of Machine Learning Research*, 13(May):1373–1392, 2012.

[48] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[49] Saharon Rosset. Bi-level path following for cross validated solution of kernel quantile regression. *Journal of Machine Learning Research*, 10(Nov):2473–2505, 2009.

[50] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

[51] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

[52] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[53] Steve Smale and Yuan Yao. Online learning algorithms. *Foundations of computational mathematics*, 6(2):145–170, 2006.

[54] Steve Smale and Ding-Xuan Zhou. Shannon sampling ii: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.

[55] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[56] Ingo Steinwart, Andreas Christmann, et al. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

[57] Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, 2009.

[58] Ingo Steinwart and Clint Scovel. Mercers theorem on general domains: on the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35(3):363–417, 2012.

[59] Ichiro Takeuchi, Quoc V Le, Timothy D Sears, and Alexander J Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(Jul):1231–1264, 2006.

[60] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

[61] Vladimir Naumovich Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[62] Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

[63] Grace Wahba et al. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. *Advances in Kernel Methods-Support Vector Learning*, 6:69–87, 1999.

[64] Yuyang Wang, Roni Khardon, Dmitry Pechyony, and Rosie Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *Conference on Learning Theory*, pages 13–1, 2012.

[65] Kilian Q Weinberger, Fei Sha, and Lawrence K Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the twenty-first international conference on Machine learning*, page 106. ACM, 2004.

[66] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[67] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005, 2004.

[68] Dao-Hong Xiang, Ting Hu, and Ding-Xuan Zhou. Learning with varying insensitive loss. *Applied Mathematics Letters*, 24(12):2107–2109, 2011.

[69] Gui-Bo Ye and Ding-Xuan Zhou. Fully online classification by regularization. *Applied and Computational Harmonic Analysis*, 23(2):198–214, 2007.

[70] Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

[71] Yiming Ying and D-X Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788, 2006.

[72] Yiming Ying and Ding-Xuan Zhou. Online pairwise learning algorithms. *Neural computation*, 28(4):743–777, 2016.

[73] Yiming Ying and Ding-Xuan Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224–244, 2017.

[74] Haizhang Zhang, Yuesheng Xu, and Jun Zhang. Reproducing kernel banach spaces for machine learning. *Journal of Machine Learning Research*, 10(Dec):2741–2775, 2009.

[75] Tong Zhang. Effective dimension and generalization of kernel learning. In *NIPS*, 4(1):454–461, 2002.

[76] Tong Zhang. On the dual formulation of regularized linear systems with convex risks. *Machine Learning*, 46(1-3):91–129, 2002.

[77] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

[78] Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16:3299–3340, 2015.

[79] Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

[80] L. Zwald. *Performances statistiques d'algorithmes d'apprentissage:" Kernel projection machine" et analyse en composantes principales à noyau.* PhD thesis, Universit Paris-Sud, 11 2005.

[81] L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysis, In *NIPS*, 1649-1656, 2006.