REGRESSION LEARNING WITH CONTINUOUS AND DISCRETE DATA

CHENDI WANG

PhD

THE HONG KONG POLYTECHNIC UNIVERSITY

2021

The Hong Kong Polytechnic University

Department of Applied Mathematics

Regression Learning with Continuous and Discrete Data

Chendi Wang

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

May 2021

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

Machine learning has achieved enormous successes in many different application areas of data mining in the last twenty years. Regression is a big branch of learning problems. This thesis investigates several topics in regression learning problems from the perspective of learning theory and asymptotic theory.

First, we study a pairwise regularized least squares learning algorithm using the Kronecker product kernels. This pairwise learning model covers both score-based ranking problems and non-linear metric learning problems. A rank-independent non-asymptotic convergence rate of the obtained pairwise learning algorithm is derived. The pairwise learning algorithm achieves the minimax optimal learning rate, which is also derived in this thesis.

Second, we propose an empirical feature-based sparse approximation algorithm for privacy consideration. Instead of using sensitive private data, empirical features are computed with published unlabeled data (without privacy issues). Summary statistics instead of raw data are used to protect private information. This semi-supervised learning algorithm achieves both sparsity and approximation accuracy.

Third, we study the asymptotic theory of a modified Poisson estimator for discrete grouped and right-censored (GRC) count data. Asymptotic theoretical properties are derived under milder conditions on the information matrix of observations and results apply to both stochastic and fixed regressors. Results in this thesis improve existing results on modified Poisson estimators for GRC counts, where stochastic regressors

with strictly positive definite Fisher information matrices are studied, significantly. The big data performance of this estimator is investigated with data on drug use in America.

Publications Arising from the Thesis

1. C. Wang and X. Guo

   On the Minimax Optimality of Pairwise Learning with Kronecker Kernel Ridge Regression. Manuscript in Preparation.

   Sparse Semi-supervised Learning with Summary Statistics. Manuscript in Preparation.

2. C. Wang

   Modified Poisson Estimators for Grouped and Right-censored Counts. Accepted by Communications in Statistics - Theory and Methods.

Acknowledgements

My colorful Ph.D. journey is non-isolated. I would like to hereby acknowledge the assistance from all sides.

First and foremost, I would like to express my deep thanks to my supervisor Dr. Xin Guo who enlightens me in my early research career and provides me heuristic guidance on my own research interests in the later stage of my Ph.D. study. Moreover, I gratefully acknowledge his support for my future career.

Furthermore, I wish to thank my co-supervisor Prof. Xiaojun Chen for her suggestions on my research from the perspective of optimization.

I would like to thank Prof. Dingxuan Zhou for his guidance in our weekly group meeting and thank Prof. Qiang Wu for his assistance in our cooperative project.

I would like to acknowledge Dr. Xin Guo and Dr. Qiang Fu who introduce their work to me and provide me the real data.

I would like to express my thanks to my friends Mr. Jintian Zhu, Mr. Tianxing Mei, Dr. Lei Yang, Dr. Jin Yang, Dr. Zhiying Fang, Dr. Ben Dai, Dr. Yuze Zhang, Miss Changyu Liu, and Dr. Huihui Qin for their companion in my Ph.D. journey and their valuable advice on my career.

Especially, I would like to thank my family. I thank my parents Deni Fang and Lianhe Wang for their support in the past 26 years. I wish to express my thanks to my fiancee Peiran Yu whose love encourages me to go on.

Contents

Chapter 1

Introduction to Learning Theory

With a rapid development of computing hardware in the past decades, plentiful information can be obtained from massive data through automatic machine learning algorithms. As a result, data science, including image recognition, artificial intelligence, and sensitive data protection, becomes an indispensable part in modern society. Learning theory aims to provide a theoretical analysis of machine learning algorithms and to refine the learning efficiency of existing algorithms based on mathematical theory. An introduction to learning theory is given in this chapter.

1.1 Learning Problems

Consider an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Here, $\mathcal{X}$ is a compact metric space and $\mathcal{Y}$ is a subset of $\mathbb{R}$. The product space $\mathcal{X} \times \mathcal{Y}$ is equipped with a probability distribution ρ. ρ can be decomposed as a conditional distribution $\rho(y|x)$ on $\mathcal{Y}$ and a marginal distribution $\rho_{\mathcal{X}}$ on $\mathcal{X}$. Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a loss function. One objective of machine learning is to recover the target function $f_\rho^\ell : \mathcal{X} \to \mathcal{Y}$ that minimizes the risk

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

Specifically, a machine learning algorithm finds a function $f^D : \mathcal{X} \to \mathcal{Y}$ automatically from a class of functions $\mathcal{F}$ (hypothesis class), according to a sample

$D = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$ drawn independently from ρ, to approximate f_ρ^ℓ .

1.1.1 Least Squares Regression

One of the most fundamental problems in machine learning is the least squares regression problem with the least squares loss

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

The corresponding risk

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

is minimized by the regression function taking the form

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x).$$

In general, $\mathcal{E}(f_\rho)$, also known as the Bayes risk, is not 0. For example, if $y = f_\rho(x) + \epsilon$, where the noise ϵ has zero-mean Gaussian distribution with variance $\sigma^2 > 0$ and is independent of x, then $\mathcal{E}(f_\rho) = \sigma^2 > 0$. The excess risk $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ is widely used as a measurement of the accuracy of a machine learning algorithm. For least squares problems, it is not difficult to verify that

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

where $\|\cdot\|_\rho$ is the $L^2_{\rho_\mathcal{X}}$ norm on the space of all square-integrable functions on $\mathcal{X}$ with respect to $\rho_\mathcal{X}$.

1.1.2 Classification

Consider a binary classification problem with $\mathcal{Y} = \{\pm 1\}$. In classification problems, a frequently studied loss is the 0-1 loss

$$\ell_{0\text{-}1}(f(x), y) = \mathbb{1}(f(x) \neq y),$$

where $\mathbb{1}$ is the indicator function. Then corresponding risk of the 0-1 loss is the mis-classification error $\mathcal{R}(f) = \rho(y \neq f(x))$. The minimizer of $\mathcal{R}(f)$ is the Bayes classifier

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

Since 0-1 loss is non-convex and is intractable in practice, one may consider convex surrogate loss functions for 0-1 loss [10, 86], such as the hinge loss

$$\ell_{\text{Hinge}} = \max\{0, 1 - yf(x)\}.$$

For hinge loss, there holds

$$\mathcal{R}(\text{sgn}(f)) - \mathcal{R}(f_c) \leq \sqrt{\mathcal{E}_{\ell_{\text{Hinge}}}(f) - \mathcal{E}_{\ell_{\text{Hinge}}}(f_c)}, \tag{1.1}$$

for any measurable $f : \mathcal{X} \to \mathbb{R}$, where sgn is the sign function. (1.1) implies that the convergence of the excess risk of f with respect to the hinge loss leads to the convergence of the excess risk of $\text{sgn}(f)$ with respect to the 0-1 loss. (1.1) is a special case of the comparison theorem [10, 20, 86]. Another widely used loss in classification is the logistic loss

$$\ell_{\text{logistic}}(y, f(x)) = \frac{1}{\log 2} \log(1 + \exp(-yf(x))).$$

Logistic loss is a special case of the cross-entropy loss for multi-class classification problems (for example, [29]).

1.1.3 Pairwise Learning

Pairwise learning aims to learn a bivariate function $F : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$, that represents the relationship between two points $x, u \in \mathcal{X}$. Pairwise learning problems include, for example, ranking [3, 23, 27, 30], similarity and metric learning [16, 19, 52, 83], and AUC maximization [84, 89].

Scoring-based Ranking

Let $x, x' \in \mathcal{X}$, and let s and s' be the score of x and x', correspondingly. x is preferred over x' if $s > s'$. The target function $f_\rho : \mathcal{X} \to \mathcal{Y}$ is known as the scoring function. One may consider minimizing the probability of ranking mistake (also known as the ranking risk)

$$\text{Prob}\left[\left(\frac{s-s'}{2}\right)(f_\rho(x) - f_\rho(x')) < 0\right].$$

Similarity Learning

Similarity learning aims to learn the similarity between two points $x, x' \in \mathcal{X} \subset \mathbb{R}^d$. In bilinear similarity learning, the target function takes the form $x^T M_\rho x'$ with M_ρ being a $d \times d$ symmetric positive semi-definite matrix. We will generalize the bilinear similarity learning to the non-linear case in Chapter 2.

1.2 Kernel-based Least Squares Regression

Kernel methods [25, 76, 79], including kernel-based support vector machine [13, 24] and regularized least squares (e.g. [67]) draws much attention in the past two decades. Kernel-based learning algorithms are an important part of this dissertation. We give an introduction to kernel methods for least squares regression and reproducing kernel Hilbert spaces (RKHS) in this section.

1.2.1 Reproducing Kernel Hilbert Spaces

Consider a continuous symmetric positive semi-definite kernel (also known as a Mercer kernel) $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, that is

$$K(x, u) = K(u, x), \quad \text{for all } x, u \in \mathcal{X}$$

and

$$\sum_{i,j=1}^{m} c_i c_j K(u_i, u_j) \geq 0, \quad \text{for all } \{u_i\}_{i=1}^m \subset \mathcal{X}, \{c_i\}_{i=1}^m \subset \mathbb{R}, m \in \mathbb{N}.$$

Let $K_x : \mathcal{X} \to \mathbb{R}$ be a function defined by $K_x(u) = K(x, u)$ for any $x, u \in \mathcal{X}$. The inner product $\langle \cdot, \cdot \rangle_K$ is defined such that

$$\langle K_x, K_u \rangle = K(x, u), \quad \text{for any } x, u \in \mathcal{X}.$$

The corresponding reproducing kernel Hilbert space is given by

$$\mathcal{H}_K := \overline{\text{Span}\,\{K_x, x \in \mathcal{X}\}},$$

where the completion is taken with respect to the norm $\|\cdot\|_K$ induced by $\langle \cdot, \cdot \rangle_K$. For any $f \in \mathcal{H}_K$, there holds the reproducing property

$$f(x) = \langle f, K_x \rangle_K .$$

Denote $L^2_{\rho_\mathcal{X}}(\mathcal{X})$ the space of all the square integrable functions with respect to $\rho_\mathcal{X}$ equipped with the $L^2_{\rho_\mathcal{X}}$ norm, and introduce the integral operator

$$\begin{aligned} L_K : L^2_{\rho_\mathcal{X}}(\mathcal{X}) &\to L^2_{\rho_\mathcal{X}}(\mathcal{X}) \\ f &\mapsto \int_\mathcal{X} f(x) K_x d\rho_\mathcal{X}(x). \end{aligned} \tag{1.2}$$

L_K is a compact, symmetric, positive semi-definite, and Hilbert-Schmidt operator [67]. Moreover, we have $\mathcal{H}_K = L_K^{1/2}(L^2_{\rho_\mathcal{X}}(\mathcal{X}))$ as shown in [25]. Thus, we can write the eigensystem of L_K as $\{(\lambda_i, \phi_i)\}_{i=1}^\infty$. Here the non-negative eigenvalues $\lambda_i' s$ are arranged in non-increasing order and the eigenfunctions $\phi_i' s$ are normalized in $L^2_{\rho_\mathcal{X}}(\mathcal{X})$. There holds the following Mercer's expansion [54],

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

The convergence of the series of functions in (1.3) is absolute and uniform.

1.2.2 Empirical Risk Minimization

Recall the (population) risk $\mathcal{E}_\ell(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y) d\rho(x, y)$. In practice, the distribution ρ is unknown and one may consider minimizing the empirical risk

$$\mathcal{E}_\ell^D(f) = \frac{1}{N}\sum_{i=1}^N \ell(f(x_i), y_i).$$

In kernel-based learning schemes, a regularization functional $\Omega : \mathcal{H}_K \to \mathbb{R}$ is frequently adopted to prevent overfitting. A regularized learning algorithm has the form

$$f_{\ell,\lambda}^D = \arg\min_{f\in\mathcal{H}_K} \left\{\mathcal{E}_\ell^D(f) + \lambda\Omega(f)\right\},$$

where $\lambda > 0$ is a tuning parameter.

If $\Omega(f) = g(\|f\|)$ with some strictly increasing function $g : [0, +\infty) \to \mathbb{R}$, then the famous representer theorem [79] says that $f_{\ell,\lambda}^D$ belongs to the finite-dimensional space spanned by $\{K_{x_i}\}_{i=1}^N$.

1.2.3 Kernel-based Regularized Least Squares

Consider the regularized least squares learning algorithm

$$f_\lambda^D := \arg\min_{f\in\mathcal{H}_K} \left\{\frac{1}{N}\sum_{i=1}^N (f(x_i) - y_i)^2 + \lambda\|f\|_K^2\right\}. \tag{1.4}$$

Thanks to the representer theorem, f_λ^D is in the class of functions taking the form $\sum_{i=1}^N c_i K_{x_i}$, with coefficient vector $\mathbf{c} = (c_1, \cdots, c_N)^T \in \mathbb{R}^N$. Write $\mathbf{x} = \{x_i\}_{i=1}^N$ and $\mathbf{y} = (y_1, \cdots, y_N)^T \in \mathbb{R}^N$. Denote $K_\mathbf{x} = (K(x_i, x_j)) \in \mathbb{R}^{N\times N}$ the Gram matrix. Substitute

$$f = \sum_{i=1}^N c_i K_{x_i} \tag{1.5}$$

into (1.4) and we obtain

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

where $\mathbf{c}^D = (c_1^D, \cdots, c_N^D)^T$ is solution to the quadratic programming

$$\min_{c \in \mathbb{R}^{\mathbb{N}}} \left\{ \frac{1}{N} \|K_{\mathbf{x}}\mathbf{c} - \mathbf{y}\|_2^2 + \lambda \mathbf{c}^T K_{\mathbf{x}} \mathbf{c} \right\}. \tag{1.6}$$

The first order condition of (1.6) implies that $\mathbf{c}^D = (N\lambda I + K_{\mathbf{x}})^{-1}\mathbf{y}$. Here I is the identity matrix (or identity operator) whose dimension could be inferred from the context.

The regularized least squares algorithm (1.4) with a penalty term $\lambda \|f\|_K^2$ is also known as the kernel ridge regression (KRR). The convergence of f_λ^D to the regression function f_ρ has been studied in literature [25, 60, 66, 67, 70] for a long time and the convergence rate reaches the minimax optimal rate as shown in [17, 69]. In Chapter 2, we establish the learning theory of a novel regularized least squares learning algorithm for pairwise learning with a ridge-type penalty term based on the so-called Kronecker product kernels.

Define an empirical integral operator $L_K^{\mathbf{x}}$ by

$$L_K^{\mathbf{x}} : \mathcal{H}_K \to \mathcal{H}_K$$

$$f \mapsto \frac{1}{N} \sum_{i=1}^{N} f(x_i) K_{x_i}. \tag{1.7}$$

Let $\{\lambda_i^{\mathbf{x}}\}_{i=1}^{\infty}$ be the eigenvalues of $L_K^{\mathbf{x}}$ arranged in non-increasing order and let $\{\phi_i^{\mathbf{x}}\}_{i=1}^{\infty}$ be the associated eigenfunctions of $L_K^{\mathbf{x}}$ normalized in $\mathcal{H}_K$. $\phi_i^{\mathbf{x}}$'s are called the empirical features. Note that the rank (defined by the dimension of the image) of $L_K^{\mathbf{x}}$ is at most N and the top-N empirical features $\{\phi_i^{\mathbf{x}}\}_{i=1}^{N}$ can be computed through the eigendecomposition of $K_{\mathbf{x}}$ [38]. One may consider the hypothesis space $\left\{\sum_{i=1}^{N} c_i \phi_i^{\mathbf{x}} : \mathbf{c} \in \mathbb{R}^N\right\}$ [35, 38, 91–93]. The coefficients of the output function can be obtained by solving

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

which is an empirical feature-based least squares learning algorithm. Here $P : [0, +\infty) \to \mathbb{R}$ is a penalty function. The convergence of the output function of the empirical feature-based learning are studied in literature [35, 38] with l_1 penalty or a general folded concave penalty.

Note that the sequence $\{\phi_i^{\mathbf{x}}\}_{i=1}^{\infty}$ of empirical features is decided by the input part $\mathbf{x}$ of data and is not related with the output $\mathbf{y}$. In Chapter 3, we introduce a semi-supervised learning algorithm with empirical features $\{\phi_i^{\mathbf{u}}\}_{i=1}^{\infty}$ generated by another unlabeled data set $\mathbf{u} = \{u_i\}$ from $\rho_{\mathcal{X}}$. Convergence analysis of this semi-supervised learning algorithm is given in Chapter 3.

1.2.4 Neural Networks and Random Feature Kernels

In recent years, the research of data science develops rapidly due to the fast development of computing equipment and deep learning (e.g. [46]). Deep neural networks, which generate the hypothesis space used in deep learning, are nonlinear with respect to the parameters to be trained in learning algorithms, which is different from the kernel regime where functions in the hypothesis space are linear with respect to the trainable parameters (for example, in (1.5), f is linear with respect to $\mathbf{c}$).

The linearization of neural networks with respect to the parameters around a given point (e.g. around the initialization in the stochastic gradient descent algorithm [21]) is related to learning with a random feature kernel [6, 63] called the neural tangent kernel [43]. Kernel ridge regression based on random feature kernels is studied in [64]. It will be interesting to extend our theory in Chapter 2 and Chapter 3 to the random feature kernel setting. However, since the theory based on the random feature kernels is not our focus in this thesis, we will not expand the discussion here.

1.3 Generalized Linear Models and Maximum Likelihood Estimation

In Chapter 4, we study the asymptotic theory of maximum likelihood estimators for grouped and right-censored count data. In this section, we give an introduction to classical generalized linear models (GLM) and maximum likelihood estimation. The content of this section can be found in many textbooks, for example, [47, 53].

To be consistent with common notations in statistics, let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random variable. The sample points $\{(x_i, y_i)\}_{i=1}^N$ are i.i.d. copies of (X, Y). Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ with $d \in \mathbb{N}$. In generalized linear models, the conditional distribution of Y given X is assumed to be in the exponential family, that is, the conditional density $p(y|x)$ of Y given X has the form

$$p(y|x) = c(y) \exp\left(\theta(x)y - b(\theta(x))\right)$$

with $c(y) > 0$. Here $\theta \in \Theta \subset \mathbb{R}$, where Θ is a parameter space, θ is a parameter depending on X and $b : \mathbb{R} \to \mathbb{R}$ is a known function of θ. Assume that Θ is a natural parameter space with non-empty interior where all derivatives of $b(\theta)$ exist for θ in the interior of Θ. There holds

$$b'(\theta(x)) = \mathbb{E}[Y|X = x].$$

Assume further that $\mathbb{E}[Y|X = x] = \mu(\beta^T x)$, where $\beta \in \mathbb{R}^d$ is a parameter and $\mu : \mathbb{R} \to \mathbb{R}$ is known as the mean function. μ^{-1} is called the link function. Then, we have

$$\theta(x) = (b')^{-1}(\mu(\beta^T x)).$$

When $(b')^{-1} = \mu^{-1}$ and $\theta = \beta^T x$, μ^{-1} is named as the natural link function.

Define the likelihood function for the GLM by

$$L_N(\beta) = \prod_{i=1}^N p(y_i|x_i).$$

The maximum likelihood estimator $\hat{\beta}_N$ is the maximizer of $L_N(\beta)$ over a parameter space of β. Since logarithmic function is strictly increasing, one can maximize the log-likelihood function

$$l_N(\beta) = \log(L_N(\beta)) = \sum_{i=1}^{N} \log p(y_i|x_i)$$

to obtain $\hat{\beta}_N$. The asymptotic theory of $\hat{\beta}_N$ has been established in literature (e.g. [31, 47, 75]) for a long time.

Chapter 2

Pairwise Learning with Kronecker Kernel Ridge Regression

Pairwise learning, including ranking and similarity learning, has been widely used in many fields. Kronecker kernel ridge regression (KKRR) is a pairwise learning algorithm based on the so-called Kronecker product pairwise kernels. To our best knowledge, the theoretical analysis of KKRR is rare in literature. In this chapter, properties of the Kronecker product kernels and the capacity of the corresponding reproducing kernel Hilbert spaces are studied. Based on a sharp bound on the effective dimension of the Kronecker product integral operators, we establish an upper bound on the error of KKRR. The minimax lower bound for Kronecker product kernel based learning algorithms is investigated. The convergence rate of the output function of KKRR matches the lower bound and is optimal in the sense of minimax.

2.1 Pairwise Learning and Kronecker Kernel Ridge Regression

Consider a probability space $(\mathcal{X}^2 \times \mathcal{Y}, \rho)$, where ρ is a joint probability distribution that can be decomposed as $d\rho(x, u, y) = d\rho(y|(x, u))d\rho_{\mathcal{X}}(x)d\rho_{\mathcal{X}}(u)$ with $\rho(y|(x, u))$ being a conditional distribution on $\mathcal{Y}$ given $(x, u) \in \mathcal{X}^2$, and $\rho_{\mathcal{X}}$ being the marginal distribution on $\mathcal{X}$. Here we assume that $\rho_{\mathcal{X}}$ is a continuous distribution such that

$\rho_{\mathcal{X}}[x = u] = 0$, for any $x, u \in \mathcal{X}$. In pairwise learning, the sample usually has the form $\mathbf{z} = \{(x_i, x_j, y_{ij})\}_{i,j=1}^N \subset \mathcal{X}^2 \times \mathcal{Y}$, where the input part $\{x_i\}_{i=1}^N \subset \mathcal{X}$ is drawn independently from $\rho_{\mathcal{X}}$ and the output y_{ij} is drawn from $\rho(y|(x_i, x_j))$, for each $i, j = 1, \cdots, N$. Moreover, (x_i, x_j, y_{ij}) is independent with (x_s, x_t, y_{st}) for distinct positive integers i, j, s, t. In what follows, let $Y = (y_{ij})_{N\times N}$ be the output matrix.

The objective of pairwise regression is to recover a target function $F_\rho(x, u)$ that measures the relationship between two points x and u in $\mathcal{X}$, through the observations $\mathbf{z}$. In regression problems, the target function is the regression function defined by

$$F_\rho(x, u) = \arg\min_{\alpha\in\mathbb{R}} \int_{\mathcal{Y}} (y - \alpha)^2 d\rho(y|(x, u)) = \int_{\mathcal{Y}} y d\rho(y|(x, u)). \tag{2.1}$$

This model is quite ubiquitous in regression problems. For example, (2.1) holds true in the case $y = F_\rho(x, u) + \epsilon$ with noise ϵ satisfying $\mathbb{E}[\epsilon|(x, u)] = 0$. And the corresponding outputs can be written as $y_{ij} = F_\rho(x_i, x_j) + \epsilon_{ij}$ with $\mathbb{E}\left[\epsilon_{ij}|(x_i, x_j)\right] = 0, i, j = 1, 2, \cdots, N$.

2.1.1 Kronecker Kernel Ridge Regression

Kernel ridge regression (KRR) is a powerful tool for learning a univariate target function $f_\rho(x)$. For a given Mercer's kernel

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, (x, u) \mapsto K(x, u),$$

let $(\mathcal{H}_K, \langle\cdot,\cdot\rangle_K)$ be the corresponding reproducing kernel Hilbert space (RKHS), and let $\|\cdot\|_K$ be the norm induced by $\langle\cdot,\cdot\rangle_K$. In this chapter, we always assume that $\mathcal{X}$ is compact. As a result,

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

Recall the notation $K_x = K(x, \cdot)$ and the reproducing property

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

Based on the training set $\{(x_i, y_i)\}_{i=1}^N$, the kernel ridge regression is defined by (1.4). Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_i \geq ...$ be the eigenvalues of the integral operator L_K defined by (1.2) and let ϕ_i be the eigenfunction of L_K associated with λ_i. In this chapter, $\{\phi_i\}_{i=1}^\infty$ is selected to be the orthonormal basis of $\mathcal{H}_K$.

In pairwise learning, the target function is bivariate. As a straightforward generalization of the univariate case, we consider finding a function from a function class $\mathcal{F}$ consists of bivariate functions of the form

$$F(x,u) = \sum_{i,j=1}^{\infty} F_{ij}\phi_i(x)\phi_j(u) \text{ with } \sum_{i,j=1}^{\infty} F_{ij}^2 < \infty. \tag{2.2}$$

Define $\mathcal{H}_K \otimes \mathcal{H}_K$ as the completion of the space

$$\left\{ F : \mathcal{X}^2 \to \mathbb{R} \middle| \, F(x,u) = \sum_{i=1}^{m} f_i(x)g_i(u), f_i, g_i \in \mathcal{H}_K, m \in \mathbb{N} \right\}$$

completed with respected to the inner product such that

$$\left\langle F, \tilde{F} \right\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} \left\langle f_i, \tilde{f}_j \right\rangle_K \langle g_i, \tilde{g}_j \rangle_K$$

for any two given functions $F(x,u) = \sum_{i=1}^m f_i(x)g_i(u)$ and $\tilde{F}(x,u) = \sum_{j=1}^n \tilde{f}_j(x)\tilde{g}_j(u)$. $\mathcal{F}$ equipped with the inner product

$$\langle F, G \rangle_{\mathcal{F}} := \sum_{i,j=1}^{\infty} F_{ij}G_{ij}, \quad \forall F, G \in \mathcal{F}$$

is the same as $\mathcal{H}_K \otimes \mathcal{H}_K$ according to [5]. Moreover, according to the properties of the product between two reproducing kernels [5], $\mathcal{H}_K \otimes \mathcal{H}_K$ is an RKHS $(\mathcal{H}_{\mathscr{K}}, \langle \cdot, \cdot \rangle_{\mathscr{K}})$ spanned by a positive semi-definite pairwise kernel on $\mathcal{X}^2 \times \mathcal{X}^2$,

$$\mathscr{K}((x,u),(x',u')) := K(x,x')K(u,u'). \tag{2.3}$$

Similarly, we denote $\mathscr{K}_{(x,u)}(x',u') = \mathscr{K}((x,u),(x',u'))$ and there holds

$$F(x,u) = \left\langle F, \mathscr{K}_{(x,u)} \right\rangle_{\mathscr{K}}, \text{ for all } F \in \mathcal{H}_{\mathscr{K}} \text{ and } x, u \in \mathcal{X}.$$

Motivated by (1.4), the Kronecker kernel ridge regression (KKRR) is a pairwise learning algorithm defined by

$$F_\lambda^{\mathbf{z}} = \arg\min_{F\in\mathcal{H}_\mathscr{K}} \left\{ \frac{1}{N^2}\sum_{i,j=1}^{N} \left(F(x_i,x_j)-y_{ij}\right)^2 + \lambda \|F\|_\mathscr{K}^2 \right\}. \tag{2.4}$$

If we let $\mathcal{M}_N$ be the set of all $N\times N$ matrices and let $\|\cdot\|_F$ be the Frobenius norm on $\mathcal{M}_N$ induced by the Frobenius inner product $\langle A,B\rangle_F := \sum_{i,j=1}^{\infty} a_{ij}b_{ij}, \forall A = (a_{ij})_{N\times N}, B = (b_{ij})_{N\times N} \in \mathcal{M}_N$, then thanks to the representer theorem [79], one can obtain

$$F_\lambda^{\mathbf{z}} = \sum_{s,t=1}^{N} \hat{c}_{st}^{\lambda} \mathscr{K}_{(x_s,x_t)}, \tag{2.5}$$

where $\hat{c}_{st}^{\lambda}$ is the (s,t)-th element of $\hat{C}^\lambda \in \mathcal{M}_N$ such that

$$\hat{C}^\lambda = \arg\min_{C\in\mathcal{M}_N} \left\{ \frac{1}{N^2}\|K_{\mathbf{x}} C K_{\mathbf{x}} - Y\|_F^2 + \lambda \left\|K_{\mathbf{x}}^{1/2} C K_{\mathbf{x}}^{1/2}\right\|_F^2 \right\} \tag{2.6}$$

with $K_{\mathbf{x}} = (K(x_i,x_j))_{i,j=1}^N$ being the kernel matrix. To see the existence of the solution of (2.6), we introduce a symmetric (with respect to the Frobenius inner product) positive semi-definite operator

$$\begin{aligned} \mathscr{K}_{\mathbf{x}} : \mathcal{M}_N &\to \mathcal{M}_N \\ A &\mapsto K_{\mathbf{x}} A K_{\mathbf{x}}. \end{aligned}$$

By letting the gradient of (2.6) with respect to the Frobenius inner product vanish, there holds

$$\hat{C}^\lambda = \left(\mathscr{K}_{\mathbf{x}} + N^2\lambda I\right)^{-1} Y, \tag{2.7}$$

with I being the identity operator whose domain depends on the content. For two matrices $A = (a_{ij})_{N\times N}, B = (b_{ij})_{N\times N} \in \mathcal{M}_N$, define the Kronecker product between A and B as

$$A\otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1N}B \\ \vdots & \ddots & \vdots \\ a_{N1}B & \dots & a_{NN}B \end{bmatrix}.$$

If we let $\mathrm{vec}(A) := (a_{11}, \cdots, a_{N1}, a_{12}, \cdots, a_{N2}, a_{1N}, \cdots, a_{NN})^T$ be the vector of $A \in \mathcal{M}_N$, then we have

$$\mathrm{vec}(\mathscr{K}_{\mathbf{x}} A) = \mathrm{vec}(K_{\mathbf{x}} A K_{\mathbf{x}}) = (K_{\mathbf{x}} \otimes K_{\mathbf{x}}) \,\mathrm{vec}(A)$$

thanks to the symmetry of $K_{\mathbf{x}}$ and the properties of Kronecker product (Lemma 4.3.1, [41]).

In classical KRR, based on the universality of the kernel function K [68], there is a regularity assumption on the target function that $f_\rho \in L_K^r(L^2_{\rho_{\mathcal{X}}}(\mathcal{X}))$ for some $r > 0$. Specifically, for $r \geq 1/2$, this assumption is equivalent to $f_\rho \in L_K^{r-1/2}(\mathcal{H}_K)$ since for any $f, g \in L^2_{\rho_{\mathcal{X}}}(\mathcal{X})$, there holds $L_K^{1/2} f \in \mathcal{H}_K, L_K^{1/2} g \in \mathcal{H}_K$ and

$$\langle f, g \rangle_{L^2_{\rho_{\mathcal{X}}}} = \left\langle L_K^{1/2} f, L_K^{1/2} g \right\rangle_K .$$

When K is a universal kernel, the universality of $\mathscr{K}$ has already been ensured in literature, see, for example, [71, 72, 78]. Thus in this chapter, with the help of the pairwise integral operator

$$\begin{aligned} L_{\mathscr{K}} : \mathcal{H}_{\mathscr{K}} &\to \mathcal{H}_{\mathscr{K}}, \\ F &\mapsto \int_{\mathcal{X}\times\mathcal{X}} F(x,u) \mathscr{K}_{(x,u)} d\rho_{\mathcal{X}}(x) d\rho_{\mathcal{X}}(u), \end{aligned} \tag{2.8}$$

we modify the regularity assumption on the pairwise target function to be

$$F_\rho = L^r_{\mathscr{K}}(G_\rho), \quad \text{for some } G_\rho \in \mathcal{H}_{\mathscr{K}} \text{ and } r \geq 0. \tag{2.9}$$

When $r = 0$, (2.9) is reduced to $F_\rho \in \mathcal{H}_{\mathscr{K}}$. The assumption (2.9) will be discussed in Section 2.2.

In many applications, the training set is $\mathbf{z}' = \{(x_i, u_i, y_i')\}_{i=1}^n \subset \mathcal{X}^2 \times \mathcal{Y}$ drawn independently from $\rho(x, u, y')$, instead of $\mathbf{z}$. Based on $\mathbf{z}'$, the KKRR is defined as

$$F_\lambda^{\mathbf{z}'} = \arg\min_{F \in \mathcal{H}_{\mathscr{K}}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(F(x_i, u_i) - y_i'\right)^2 + \lambda \|F\|^2_{\mathscr{K}} \right\}. \tag{2.10}$$

According to the widely used *average of "sums-of-i.i.d." blocks* technique in U-statistics [23], later on, we will see that the convergence rate of $F_\lambda^{\mathbf{z}}$ is dominated by that of $F_\lambda^{\mathbf{z}'}$ with $n = \lfloor \frac{N}{2} \rfloor$, $u_i = x_{n+i}$ and $y_i' = y_{i,n+i}$. Since n has the same order of N, this technique can reduce the analysis of $\mathbf{z}$ with correlated data points to analyzing the i.i.d. sample $\mathbf{z}'$ without spoiling the learning rate.

2.1.2 Related Works

Pairwise learning covers several machine learning problems including ranking [3,23,30], similarity and metric learning [16, 19, 52]. In existing literature, the case considered most frequently for ranking is scoring-based ranking [2, 23, 30] via a univariate scoring function f_ρ. Kernel methods for pairwise learning are studied in literature [22, 85]. In Section 2.2, we will see that both ranking and bilinear similarity learning can be formulated as learning problems with target functions belonging to $\mathcal{H}_{\mathscr{K}}$.

Kernel ridge regression is a classical learning algorithm that has been widely studied in literature [17, 25, 67, 69]. Pairwise-kernel-based learning algorithms are used in chemistry [48], bio-informatics [77] and other subjects related to data science. Pairwise learning via KKRR has also been established in the past several years [9, 58, 71]. However, to our best knowledge, comparing with its extensive application, results on the properties of the Kronecker product kernels and the corresponding Kronecker product integral operators are rare in existing literature. These properties, especially the effective dimension of the integral operator [87], plays a notably significant role in deriving the minimax rate of classical KRR [17]. As a result, the theoretical guarantee of KKRR is scarce.

To reduce the high computational cost in classical KRR, a distributed learning scheme of kernel ridge regression has been introduced and analyzed by [49, 88]. There are some other techniques to reduce the computation cost of KRR such as carrying out randomized sketches to kernel matrices [82]. We believe that both distributed

learning methods and randomized sketches can be applied to KKRR for computational efficiency.

Waegeman et al. [78] introduces the symmetric part of $\mathscr{K}$,

$$\mathscr{K}^{\mathrm{S}}((x,u),(x',u')) = \frac{1}{2}\left[\mathscr{K}((x,u),(x',u')) + \mathscr{K}((u,x),(x',u'))\right], \tag{2.11}$$

and the skew-symmetric part of $\mathscr{K}$,

$$\mathscr{K}^{\mathrm{SS}}((x,u),(x',u')) = \frac{1}{2}\left[\mathscr{K}((x,u),(x',u')) - \mathscr{K}((u,x),(x',u'))\right], \tag{2.12}$$

that can be applied to similarity learning and ranking with symmetric and skew-symmetric target function, respectively. Pahikkala et al. [58] prove that the output function of KKRR based on $\mathscr{K}$ is equivalent to that based on $\mathscr{K}^{\mathrm{S}}$ ($\mathscr{K}^{\mathrm{SS}}$) when the outputs of the observations are symmetric (skew-symmetric).

In this chapter, we study some further properties of $\mathscr{K}$, $\mathscr{K}^{\mathrm{S}}$ and $\mathscr{K}^{\mathrm{SS}}$ based on the theory of RKHS and the pairwise integral operators. Furthermore, we reveal that $\mathscr{K}$-based pairwise learning is a generalization of two specific ubiquitous pairwise learning problems, the scoring-based ranking and bilinear similarity learning, in theory with the help of centered reproducing kernels [37, 80] and linear kernels. In the view of learning theory, we derive both an upper bound for the error of KKRR and a minimax lower bound for $\mathscr{K}$-based learning algorithms. The upper bound and the lower bound match each other which means that the rate of KKRR established in this chapter is minimax optimal. To our best knowledge, prior to his work, there is no learning theory estimate about $\mathscr{K}$-based pairwise regression problems.

2.1.3 Structure of this Chapter

In Section 2.2, properties of the Kronecker product kernels and the KKRR learning scheme are introduced. The convergence results of KKRR and the minimax lower bound of learning algorithms based on $\mathscr{K}$ are established in Section 2.3. For the sake of completeness, the proof is provided in Section 2.4-Section 2.7.

2.2 Properties and Applications of Kronecker Product Kernels

In this section, some properties of $\mathscr{K}$, especially the eigenvalues and eigenvectors of $L_{\mathscr{K}}$, are studied. Moreover, we show in this section that $\mathscr{K}$-based pairwise learning is a generalization of two existing models in ranking and similarity learning.

2.2.1 Properties of Kronecker Product Pairwise Kernels

For the Kronecker product pairwise kernel $\mathscr{K}$, it's obvious that

$$\sup_{(x,u)\in\mathcal{X}^2} \left(\mathscr{K}((x,u),(x,u))\right)^{1/2} = \sup_{x,u\in\mathcal{X}} \sqrt{K(x,x)K(u,u)} = \kappa^2$$

and

$$\sup_{(x,u)\in\mathcal{X}^2} \left\|\mathscr{K}_{(x,u)}\right\|_{\mathscr{K}} = \sup_{(x,u)\in\mathcal{X}^2} \left(\mathscr{K}((x,u),(x,u))\right)^{1/2} = \kappa^2.$$

Thus, for any $F \in \mathcal{H}_{\mathscr{K}}$, there holds

$$\|F\|_\infty = \sup_{(x,u)\in\mathcal{X}^2} |F(x,u)| \leq \sup_{(x,u)\in\mathcal{X}^2} \left\|\mathscr{K}_{(x,u)}\right\|_{\mathscr{K}} \|F\|_{\mathscr{K}} = \kappa^2 \|F\|_{\mathscr{K}}.$$

Moreover, since $\rho_{\mathcal{X}}$ is a probability measure, we have

$$\|L_{\mathscr{K}} F\|_{\mathscr{K}} \leq \sup_{(x,u)\in\mathcal{X}^2} \left\|K_{(x,u)}\right\|_{\mathscr{K}}^2 \|F\|_{\mathscr{K}} \leq \kappa^4 \|F\|_{\mathscr{K}}$$

and

$$\|L_{\mathscr{K}}\|_{\mathrm{op}} \leq \kappa^4, \tag{2.13}$$

where $\|\cdot\|_{\mathrm{op}}$ is the spectral norm of an operator.

Recall the symmetric part $\mathscr{K}^{\mathrm{S}}$ and the skew-symmetric part $\mathscr{K}^{\mathrm{SS}}$ of $\mathscr{K}$ defined in (2.11) and (2.12). Note that

$$\mathscr{K}^{\mathrm{SS}}((x,u),(x',u')) = \frac{1}{4}\left\langle \mathscr{K}_{(x,u)} - \mathscr{K}_{(u,x)}, \mathscr{K}_{(x',u')} - \mathscr{K}_{(u',x')} \right\rangle_{\mathscr{K}}$$

and

$$\mathscr{K}^{\mathrm{S}}((x,u),(x',u')) = \frac{1}{4}\left\langle \mathscr{K}_{(x,u)} + \mathscr{K}_{(u,x)}, \mathscr{K}_{(x',u')} + \mathscr{K}_{(u',x')} \right\rangle_{\mathscr{K}}.$$

It is not difficult to verify that both $\mathscr{K}^{\mathrm{S}}$ and $\mathscr{K}^{\mathrm{SS}}$ are Mercer's kernels on $\mathcal{X} \times \mathcal{X}$. According to the properties of the sum of reproducing kernels [5], $\mathcal{H}_{\mathscr{K}} = \mathcal{H}_{\mathscr{K}^{\mathrm{S}}} \oplus \mathcal{H}_{\mathscr{K}^{\mathrm{SS}}}$. We list some properties of $\mathscr{K}$, $\mathscr{K}^{\mathrm{S}}$, and $\mathscr{K}^{\mathrm{SS}}$ in the following Proposition that will be proved in Section 2.5.

Proposition 2.1. *Let $\{(\lambda_i, \phi_i)\}_{i=1}^{\infty}$ be the eigensystem of L_K normalized in $\mathcal{H}_K$ and define $\Phi_{ij}(x,u) = \phi_i(x)\phi_j(u)$. Then the following properties of $\mathscr{K}$ hold.*

(a) $\{\Phi_{ij}\}_{i,j=1}^{\infty}$ is the orthonormal basis of $\mathcal{H}_{\mathscr{K}}$.

(b) Mercer's expansion of $\mathscr{K}$ is given by

$$\mathscr{K}((x,u),(x',u')) = \sum_{i,j=1}^{\infty} \Phi_{ij}(x,u)\Phi_{ij}(x',u'), \tag{2.14}$$

where the convergence is absolute and uniform on $\mathcal{X} \times \mathcal{X}$.

(c) Φ_{ij} and Φ_{ji} are the eigenvectors of $L_{\mathscr{K}}$ associated with the eigenvalue $\lambda_i\lambda_j$. Thus $L_{\mathscr{K}}$ is positive semi-definite with $\mathrm{Tr}(L_{\mathscr{K}}) = \mathrm{Tr}^2(L_K) < \infty$.

(d) The subspace $\mathcal{H}_{\mathscr{K}^{\mathrm{S}}}$ of $\mathcal{H}_{\mathscr{K}}$ consists of all symmetric functions in $\mathcal{H}_{\mathscr{K}}$ with the orthonormal basis $\left\{\frac{1}{\sqrt{2}}(\Phi_{ij} + \Phi_{ji})\right\}_{i<j} \cup \{\Phi_{ii}\}_{i=1}^{\infty}$.

(e) The subspace $\mathcal{H}_{\mathscr{K}^{\mathrm{SS}}}$ of $\mathcal{H}_{\mathscr{K}}$ consists of all skew-symmetric functions in $\mathcal{H}_{\mathscr{K}}$ with the orthonormal basis $\left\{\frac{1}{\sqrt{2}}(\Phi_{ij} - \Phi_{ji})\right\}_{i<j}$.

(f) $\frac{1}{\sqrt{2}}(\Phi_{ij} + \Phi_{ji})$ and $\frac{1}{\sqrt{2}}(\Phi_{ij} - \Phi_{ji})$ are the eigenvectors of $L_{\mathscr{K}^{\mathrm{S}}}$ and $L_{\mathscr{K}^{\mathrm{SS}}}$, respectively, associated with the eigenvalue $\lambda_i\lambda_j$. Thus, $L_{\mathscr{K}^{\mathrm{S}}}$ and $L_{\mathscr{K}^{\mathrm{SS}}}$ are the constraints of $L_{\mathscr{K}}$ on $\mathcal{H}_{\mathscr{K}^{\mathrm{S}}}$ and $\mathcal{H}_{\mathscr{K}^{\mathrm{SS}}}$, respectively.

Remark. According to Proposition 2.1, $L_{\mathscr{K}}$ is a positive semi-definite operator from $\mathcal{H}_{\mathscr{K}}$ to $\mathcal{H}_{\mathscr{K}}$ since $\lambda_i\lambda_j \geq 0, i, j = 1, 2, \cdots$. Moreover, for any function $F \in \mathcal{H}_{\mathscr{K}}$, we can represent

$$F = \sum_{i,j=1}^{\infty} F_{ij}\Phi_{ij}, \text{ with } F_{ij} = \langle F, \Phi_{ij}\rangle_{\mathscr{K}}, \text{ and } \|F\|_{\mathscr{K}}^2 = \sum_{i,j=1}^{\infty} F_{ij}^2.$$

Furthermore, for any given continuous function f of $L_{\mathscr{K}}$, Φ_{ij} is the eigenvector of $f(L_{\mathscr{K}})$ corresponding to the eigenvalue $f(\lambda_i\lambda_j)$. Thus the regularity assumption (2.9) becomes

$$F_\rho = \sum_{i,j=1}^{\infty} \lambda_i^r\lambda_j^r G_{\rho ij}\Phi_{ij}, \text{ for some } G_\rho = \sum_{i,j=1}^{\infty} G_{\rho ij}\Phi_{ij} \in \mathcal{H}_{\mathscr{K}}.$$

Note that $L_K^r(\mathcal{H}_K)$ is a subspace of $\mathcal{H}_K$ and $L_K^r(\mathcal{H}_K) \otimes L_K^r(\mathcal{H}_K)$ is well-defined. It is obvious that the following proposition holds.

Proposition 2.2. *$L_K^r(\mathcal{H}_K) \otimes L_K^r(\mathcal{H}_K)$ is a subspace of $\overline{L_{\mathscr{K}}^r(\mathcal{H}_{\mathscr{K}})}$, where the completion is taken with respect to the RKHS inner product.*

Isometry between $\mathcal{H}_{\mathscr{K}}$ and $\mathrm{HS}(\mathcal{H}_K)$. We call an operator L from $\mathcal{H}_K$ to $\mathcal{H}_K$ a Hilbert-Schmidt operator if $\sum_{i=1}^{\infty} \|L\phi_i\|_K^2 < \infty$. The Hilbert space $\mathrm{HS}(\mathcal{H}_K)$ is the space of all Hilbert-Schmidt operators on $\mathcal{H}_K$ equipped with the inner product

$$\langle L_1, L_2\rangle_{\mathrm{HS}(\mathcal{H}_K)} := \sum_{i=1}^{\infty} \langle L_1\phi_i, L_2\phi_i\rangle_K, \forall L_1, L_2 \in \mathrm{HS}(\mathcal{H}_K).$$

If we define the rank-1 operator $f \otimes g$ from $\mathcal{H}_K$ to $\mathcal{H}_K$ by $(f \otimes g)\, h = \langle g, h\rangle_K f$, for any $f, g, h \in \mathcal{H}_K$, then $\{\phi_i \otimes \phi_j\}_{i,j=1}^{\infty}$ is the orthonormal basis of $\mathrm{HS}(\mathcal{H}_K)$. In fact, it is easy to verify that for any $L \in \mathrm{HS}(\mathcal{H}_K)$, $L = \sum_{i,j=1}^{\infty} L_{ij}\phi_i \otimes \phi_j$ with $L_{ij} = \langle \phi_i, L\phi_j\rangle_K = \langle L, \phi_i \otimes \phi_j\rangle_{\mathrm{HS}(\mathcal{H}_K)}$. Thus the map $\Phi_{ij} \mapsto \phi_i \otimes \phi_j, i, j = 1, 2, \cdots$ defined through the base vectors is obviously an isometry between $\mathcal{H}_{\mathscr{K}}$ and $\mathrm{HS}(\mathcal{H}_K)$.

Connections between $L_{\mathscr{K}}$ and L_K. Consider the operator $L_K\tilde{\otimes}L_K$ defined by

$$L_K\tilde{\otimes}L_K : \mathrm{HS}(\mathcal{H}_{\mathscr{K}}) \to \mathrm{HS}(\mathcal{H}_{\mathscr{K}}),$$
$$B \mapsto L_K B L_K.$$

Note that

$$L_K\tilde{\otimes}L_K(\phi_i \otimes \phi_j) = L_K(\phi_i \otimes \phi_j)L_K = (L_K\phi_i) \otimes (L_K\phi_j) = \lambda_i\lambda_j\phi_i \otimes \phi_j.$$

So $\{(\lambda_i\lambda_j, \phi_i \otimes \phi_j)\}_{i,j=1}^{\infty}$ is the eigensystem of $L_K\tilde{\otimes}L_K$. In other words, if we regard $\mathcal{H}_{\mathscr{K}}$ and $\mathrm{HS}(\mathcal{H}_K)$ as the same space, then we can rewrite $L_{\mathscr{K}} = L_K\tilde{\otimes}L_K$.

In classical KRR, the error between the output function and the target function is usually measured by the $L^2_{\rho_{\mathcal{X}}}$ norm

$$\|f\|_{L^2_{\rho_{\mathcal{X}}}} := \left(\int_{\mathcal{X}} f(x)^2 d\rho_{\mathcal{X}}(x)\right)^{1/2}, \text{ for all } f \in L^2_{\rho_{\mathcal{X}}}(\mathcal{X}).$$

Moreover, it's well-known [25, 26] that

$$\|f\|_{L^2_{\rho_{\mathcal{X}}}} = \left\|L_K^{1/2} f\right\|_K, \text{ for all } f \in L^2_{\rho_{\mathcal{X}}}(\mathcal{X}). \tag{2.15}$$

In this chapter, the learning efficiency of $F^{\mathbf{z}}_{\lambda}$ and $F^{\mathbf{z}'}_{\lambda}$ is measured by the $L^2_{\rho_{\mathcal{X}}\times\rho_{\mathcal{X}}}$ norm $\|\cdot\|_\rho$ defined by

$$\|F\|_\rho^2 = \int_{\mathcal{X}\times\mathcal{X}} F(x,u)^2 d\rho_{\mathcal{X}}(x) d\rho_{\mathcal{X}}(u), \text{ for any } F \in \mathcal{H}_{\mathscr{K}}. \tag{2.16}$$

Proposition 2.3. *For $F \in \mathcal{H}_{\mathscr{K}}$, there holds*

$$\|F\|_\rho = \left\|L_{\mathscr{K}}^{1/2} F\right\|_{\mathscr{K}}. \tag{2.17}$$

Proposition 2.3 will be proved in Section 2.5.

2.2.2 Applications of Kronecker Product Pairwise Kernels

The target function belonging to $\mathcal{H}_{\mathscr{K}}$ can cover and generalize several models in AUC-based ranking and similarity learning. In this chapter, we only study the least

squares loss for regression. In fact, for pairwise SVM classifier based on the hinge loss, one can easily obtain the bound of its generalization error (e.g. [16]) by a general Rademacher complexity argument since the pairwise kernel is bounded.

Bipartite ranking

In bipartite ranking, the target function F_ρ is used to compare two items $x_i, x_j \in \mathcal{X}$. Precisely, if $F_\rho(x_i, x_j) > 0$, then we say that x_i is preferred over x_j.

Bipartite ranking via scoring functions. As a special case of bipartite ranking, the target of scoring-based ranking is to recover a scoring function f_ρ via the sample $\{(x_i, s_i)\}_{i=1}^N$ with s_i being the score of x_i such that $s_i > s_j$ if x_i is preferred over x_j.

Consider the scores s_i and s_j pairwisely, and we reformulate the scoring-based ranking as $y_{ij} = s_i - s_j$ and $F_\rho(x, u) = f_\rho(x) - f_\rho(u)$. Then the bipartite ranking problem can be studied under the framework of pairwise learning [85]. For $f_\rho \in \mathcal{H}_K$, there holds $f_\rho(x) = \sum_{i=1}^\infty f_{\rho_i}\phi_i(x)$. In this setting,

$$F_\rho(x, u) = \sum_{i=1}^\infty f_{\rho_i}(\phi_i(x) - \phi_i(u)) = \sum_{i=1}^\infty f_{\rho_i}(\phi_i(x)\mathbb{1}(u) - \phi_i(u)\mathbb{1}(x)) \tag{2.18}$$

with $\mathbb{1}(x) \equiv 1$ for any $x \in \mathcal{X}$. To make sure that F_ρ is still in $\mathcal{H}_\mathscr{K}$, the constant function $\mathbb{1}$ should be the eigenfunction of L_K, which is not always the case. For this reason, [37] introduced a centered reproducing kernel

$$\bar{K}(x, u) := K(x, u) - \int_\mathcal{X} K(x, u)d\rho_\mathcal{X}(x) - \int_\mathcal{X} K(x, u)d\rho_\mathcal{X}(u) + \int_{\mathcal{X}\times\mathcal{X}} K(x, u)d\rho_\mathcal{X}(x)d\rho_\mathcal{X}(u).$$

In practice, $\bar{K}$ can be approximated by

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

$\mathcal{H}_{\bar{K}}$ is perpendicular to constant functions in the sense of $L^2_{\rho_\mathcal{X}}$ inner product, that is, for any $f \in \mathcal{H}_{\bar{K}}$, $\int_\mathcal{X} f(x)d\rho_\mathcal{X}(x) = 0$. Note that $K^1(x, u) \equiv 1$ is a reproducing

kernel with $(\mathcal{H}_{K^1}, \langle\cdot,\cdot\rangle_{K^1}) = (\mathbb{R}, \cdot)$, where $\cdot$ is the product between two real numbers. Consider a new kernel function $\tilde{K}(x,u) = \bar{K}(x,u) + K^1(x,u)$. Then $\mathcal{H}_{\tilde{K}} = \mathcal{H}_{\bar{K}} \oplus \mathbb{R}$ and $\phi_0 = \mathbb{1}$ is the eigenfunction of $L_{\tilde{K}}$ with respect to the eigenvalue $\lambda_0 = 1$. In fact,

$$L_{\tilde{K}}(\mathbb{1}) = \int_{\mathcal{X}} \bar{K}_x + \mathbb{1} d\rho_{\mathcal{X}}(x) = 0 + \mathbb{1} = \mathbb{1}.$$

Let $\bar{\phi}_i$ be the eigenfunction of $L_{\bar{K}}$ associated with the i-th eigenvalue $\bar{\lambda}_i$ of $L_{\bar{K}}$. We have

$$L_{\tilde{K}}\bar{\phi}_i = \int_{\mathcal{X}} \bar{\phi}_i(x)\bar{K}_x + \bar{\phi}_i(x) d\rho_{\mathcal{X}}(x) = L_{\bar{K}}\bar{\phi}_i + 0 = \bar{\lambda}_i\bar{\phi}_i.$$

As a consequence, $\bar{\phi}_i$ is the eigenfunction of $L_{\tilde{K}}$ with respect to $\bar{\lambda}_i$. Then we replace (2.18) with

$$\begin{aligned} F_\rho(x,u) &= \left(f_\rho(x) - \int_{\mathcal{X}} f_\rho(x) d\rho_{\mathcal{X}}(x)\right) - \left(f_\rho(u) - \int_{\mathcal{X}} f_\rho(u) d\rho_{\mathcal{X}}(u)\right) \\ &= \sum_{i=1}^{\infty} \bar{f}_{\rho_i} \left(\bar{\phi}_i(x)\phi_0(u) - \phi_0(x)\bar{\phi}_i(u)\right), \end{aligned} \tag{2.19}$$

which is a function in $\mathcal{H}_{\tilde{\mathscr{K}}^{\mathrm{SS}}}$ with $\tilde{\mathscr{K}}((x,u),(x',u')) = \tilde{K}(x,x')\tilde{K}(u,u')$.

For bipartite ranking, we can assume that the target function has the form

$$\begin{aligned} F_\rho(x,u) &= \sum_{i=1}^{\infty}\sum_{j<i} F_{\rho_{ij}} \left(\phi_i(x)\phi_j(u) - \phi_i(u)\phi_j(x)\right) \\ &= \sum_{i=1}^{\infty}\sum_{j<i} F_{\rho_{ij}} \left(\Phi_{ij}(x,u) - \Phi_{ji}(x,u)\right) \in \mathcal{H}_{\mathscr{K}^{\mathrm{SS}}}, \end{aligned}$$

which is a generalization of the scoring-based ranking (2.19).

Similarity learning

The target function of bilinear similarity learning [19, 52] is the similarity function

$$F_\rho(x,u) = x^T M_\rho u, \forall x, u \in \mathcal{X} = \mathbb{R}^d, \tag{2.20}$$

with M_ρ being a $d \times d$ symmetric positive semi-definite matrix.

We are now going to show that (2.20) can be formulated under the framework of $\mathcal{H}_{\mathscr{K}}$ with respect to the linear kernel $K^L(x,u) = x^T u$, whose corresponding RKHS is isometric to $\mathbb{R}^d$ equipped with the Euclidean inner product $\langle x, u\rangle_d := x^T u$. Precisely, for any $f \in \mathcal{H}_{K^L}$, there is one and only one $y_f \in \mathbb{R}^d$ such that $f(x) = y_f^T x = \langle y_f, x\rangle_d$. Moreover, for the integral operator L_{K^L},

$$L_{K^L} f(u) = \int_{\mathcal{X}} y_f^T x x^T u d\rho_{\mathcal{X}}(x) = (My_f)^T u, \text{ where } M = \int_{\mathcal{X}} xx^T d\rho_{\mathcal{X}}(x).$$

As a consequence, $\{(\lambda_i, y_{\phi_i})\}_{i=1}^d$ is the eigensystem of M and $\{y_{\phi_i}\}_{i=1}^d$ is the orthonormal basis of $\mathbb{R}^d$, where $\{(\lambda_i, \phi_i)\}_{i=1}^d$ is the eigensystem of L_{K^L}. Since M_ρ is symmetric, by letting $M_{\rho_{ij}} := y_{\phi_i}^T M_\rho y_{\phi_j} = y_{\phi_j}^T M_\rho y_{\phi_i} = M_{\rho_{ji}}$, we can rewrite M_ρ as

$$M_\rho = \sum_{i,j=1}^d M_{\rho_{ij}} y_{\phi_i} y_{\phi_j}^T = \sum_{i<j} M_{\rho_{ij}} \left(y_{\phi_i} y_{\phi_j}^T + y_{\phi_j} y_{\phi_i}^T \right) + \sum_{i=1}^d M_{\rho_{ii}} y_{\phi_i} y_{\phi_i}^T.$$

Thus (2.20) is equivalent to

$$F_\rho(x,u) = \sum_{i,j=1}^d M_{\rho_{ij}} x^T y_{\phi_i} y_{\phi_j}^T u = \sum_{i,j=1}^d M_{\rho_{ij}} \phi_i(x)\phi_j(u)$$

$$= \sum_{i<j} M_{\rho_{ij}} \left(\phi_i(x)\phi_j(u) + \phi_j(x)\phi_i(u)\right) + \sum_{i=1}^d M_{\rho_{ii}} \phi_i(x)\phi_i(u) \in \mathcal{H}_{\mathscr{K}^S}.$$

By replacing the linear kernel with general universal kernels, we can generalize the bilinear similarity learning model to non-linear models.

2.2.3 Properties of KKRR

To study the solution of (2.4), like the classical KRR, we introduce the sampling operator $S_{\mathbf{x}} : \mathcal{H}_{\mathscr{K}} \to \mathcal{M}_N$, $S_{\mathbf{x}} F := \left(F(x_i, x_j)\right)_{N\times N}$, for all $F \in \mathcal{H}_{\mathscr{K}}$. Then its adjoint is $S_{\mathbf{x}}^T : \mathcal{M}_N \to \mathcal{H}_{\mathscr{K}}, S_{\mathbf{x}}^T A = \sum_{i,j=1}^N a_{ij} \mathscr{K}_{(x_i,x_j)}, \forall A = (a_{ij})_{N\times N} \in \mathcal{M}_N$. Moreover,

define the empirical integral operator from $\mathcal{H}_{\mathscr{K}}$ to $\mathcal{H}_{\mathscr{K}}$ by

$$L_{\mathscr{K}}^{\mathbf{x}} F = \frac{1}{N^2} S_{\mathbf{x}}^T S_{\mathbf{x}} F = \frac{1}{N^2} \sum_{i,j=1}^{N} F(x_i, x_j) \mathscr{K}_{(x_i, x_j)}, \text{ for any } F \in \mathcal{H}_{\mathscr{K}}.$$

It is obvious that $L_{\mathscr{K}}^{\mathbf{x}}$ is symmetric. To see the positive semi-definiteness, we introduce two operators, $L_K^{\mathbf{x}} : \mathcal{H}_K \to \mathcal{H}_K, f \mapsto \frac{1}{N} \sum_{i=1}^{N} f(x_i) K_{x_i}$, which is positive semi-definite according to the theory of classical KRR, and $L_K^{\mathbf{x}} \tilde{\otimes} L_K^{\mathbf{x}} : \mathrm{HS}(\mathcal{H}_K) \to \mathrm{HS}(\mathcal{H}_K), B \mapsto L_K^{\mathbf{x}} B L_K^{\mathbf{x}}$. We donte $\lambda_i^{\mathbf{x}}, i = 1, 2, \cdots$, the eigenvalues of $L_K^{\mathbf{x}}$ arranged in non-increasing order. Then, like the relations between $L_{\mathscr{K}}$ and L_K, it is not difficult to verify that $L_{\mathscr{K}}^{\mathbf{x}} = L_K^{\mathbf{x}} \tilde{\otimes} L_K^{\mathbf{x}}$ by regarding $\mathrm{HS}(\mathcal{H}_K)$ and $\mathcal{H}_{\mathscr{K}}$ as a same space and $\lambda_i^{\mathbf{x}} \lambda_j^{\mathbf{x}} \geq 0, i, j = 1, 2, \cdots$, are eigenvalues of $L_{\mathscr{K}}^{\mathbf{x}}$. As a result, $L_{\mathscr{K}}^{\mathbf{x}} + \lambda I$ is invertible and we can write

$$F_\lambda^{\mathbf{z}} = \left(L_{\mathscr{K}}^{\mathbf{x}} + \lambda I\right)^{-1} \frac{1}{N^2} S_{\mathbf{x}}^T Y. \tag{2.21}$$

For (2.10) based on $\mathbf{z}' = \{(x_i, u_i, y_i')\}_{i=1}^n$, similarly, if we denote $\mathbf{x}' = \{(x_i, u_i)\}_{i=1}^n, Y' = (y_1', ..., y_n')^T$, and introduce the sampling operator $S_{\mathbf{x}'} : \mathcal{H}_{\mathscr{K}} \to \mathbb{R}^n$ such that $S_{\mathbf{x}'} F = (F(x_i, u_i)))_{i=1}^n$, then $S_{\mathbf{x}'}^T c = \sum_{i=1}^n c_i \mathscr{K}_{(x_i, u_i)}, \forall c \in \mathbb{R}^n$ and

$$F_\lambda^{\mathbf{z}'} = \left(L_{\mathscr{K}}^{\mathbf{x}'} + \lambda I\right)^{-1} \frac{1}{n} S_{\mathbf{x}'}^T Y', \tag{2.22}$$

where $L_{\mathscr{K}}^{\mathbf{x}'} F := \frac{1}{n} \sum_{i=1}^n F(x_i, u_i) \mathscr{K}_{(x_i, u_i)}$ is a symmetric positive semi-definite operator since $\mathscr{K}$ is a Mercer's kernel.

2.3 Convergence Results

We state the main convergence theorems in this section. Since the convergence rate depends on a quantity known as the effective dimension of $L_{\mathscr{K}}$, the bounds of this quantity and some corresponding convergence results are studied. In the end of this section, the minimax rate for $\mathscr{K}$-based pairwise regression is provided. Since both the minimax rate and the final error bounds are determined by the eigenvalues of L_K,

according to Proposition 2.1, we study only the learning theory of $\mathscr{K}$-based KKRR and one can easily generalize the results to KKRR based on $\mathscr{K}^{\mathrm{S}}$ and $\mathscr{K}^{\mathrm{SS}}$.

2.3.1 Main Theorems

In classical learning theory, the capacity of $\mathcal{H}_K$ is usually measured by the effective dimension of L_K defined by

$$\mathcal{N}_K(\lambda) = \mathrm{Tr}\left(L_K(L_K+\lambda I)^{-1}\right) = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i+\lambda}.$$

As an analogy, we introduce the effective dimension of $L_{\mathscr{K}}$ as

$$\mathcal{N}_{\mathscr{K}}(\lambda) := \mathrm{Tr}\left(L_{\mathscr{K}}\left(L_{\mathscr{K}}+\lambda I\right)^{-1}\right) = \sum_{i,j=1}^{\infty} \frac{\lambda_i\lambda_j}{\lambda+\lambda_i\lambda_j}. \tag{2.23}$$

As we can see in the following two theorems, $\mathcal{N}_{\mathscr{K}}(\lambda)$ is essential in the final error bounds.

We first state the convergence of $F_\lambda^{\mathbf{z}'}$.

Theorem 2.1. *Assume $|y'| \le M$. Under the regularity assumption (2.9) with $0 \le r < 1/2$, there holds*

$$\mathbb{E}\left[\left\|F_\lambda^{\mathbf{z}'} - F_\rho\right\|_\rho\right] \le C'\left(\mathcal{A}_{n,\lambda}\left(\frac{\mathcal{A}_{n,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{n,\lambda}^2}{\lambda} + 1\right) + \lambda^{r+1/2}\right), \tag{2.24}$$

where $\mathcal{A}_{n,\lambda} = \frac{1}{n\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}_{\mathscr{K}}(\lambda)}{n}}$, and C' is a constant independent of n or λ that will be specified in the proof.

For the output function $F_\lambda^{\mathbf{z}}$, we have the following error bound.

Theorem 2.2. *Assume $|y| \le M$. Under the regularity assumption (2.9) with $0 \le r < 1/2$, for $N \ge 4$, there holds*

$$\mathbb{E}\left[\left\|F_\lambda^{\mathbf{z}} - F_\rho\right\|_\rho\right] \le C\left(\mathcal{A}_{N,\lambda}\left(\frac{\mathcal{A}_{N,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{N,\lambda}^2}{\lambda} + 1\right) + \lambda^{r+1/2}\right), \tag{2.25}$$

where $\mathcal{A}_{N,\lambda} = \frac{1}{N\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}_{\mathscr{K}}(\lambda)}{N}}$, *and* C *is a constant independent of* N *or* λ *that will be specified in the proof.*

2.3.2 Capacity Dependent Error Analysis

From the definition of $\mathcal{N}_{\mathscr{K}}(\lambda)$, we see that $\mathcal{N}_{\mathscr{K}}$ and $\mathcal{N}_K$ has the relationship

$$\mathcal{N}_{\mathscr{K}}(\lambda) = \sum_{i,j=1}^{\infty} \frac{\lambda_j}{\lambda_j + (\lambda/\lambda_i)} = \sum_{i=1}^{\infty} \mathcal{N}_K\left(\frac{\lambda}{\lambda_i}\right).$$

Moreover, there holds

$$\mathcal{N}_K^2\left(\sqrt{\lambda}\right) = \sum_{i,j=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i + \sqrt{\lambda}}\right)\left(\frac{\lambda_j}{\lambda_j + \sqrt{\lambda}}\right) \leq \sum_{i,j=1}^{\infty} \frac{\lambda_i\lambda_j}{\lambda_i\lambda_j + \lambda} = \mathcal{N}_{\mathscr{K}}(\lambda).$$

It is well understood that $\mathcal{N}_K(\lambda)$ is closely related to the decay of $\{\lambda_i\}_{i=1}^{\infty}$. For $\mathcal{N}_{\mathscr{K}}(\lambda)$, we have similar results. The accurate bounds of $\mathcal{N}_{\mathscr{K}}(\lambda)$ can be obtained if we assume that the eigenvalues decay polynomially or exponentially.

Proposition 2.4. *There hold the following estimates of* $\mathcal{N}_{\mathscr{K}}(\lambda)$.

(i) (Upper Bound.) If we assume $\lambda_i \leq D_2 i^{-1/s_2}, i = 1, 2, \cdots$, *for some* $D_2 < \infty$ *and* $0 < s_2 < 1$, *then there is a constant* $C_0 < \infty$ *such that*

$$\mathcal{N}_{\mathscr{K}}(\lambda) \leq C_0 \lambda^{-s_2} \log(1/\lambda), \text{ for } 0 < \lambda \leq e^{-1}. \tag{2.26}$$

(ii) (Lower Bound.) If we assume $\lambda_i \geq D_1 i^{-1/s_1}, i = 1, 2, \cdots$, *for* $D_1 > 0$ *and* $0 < s_1 < 1$, *then there is a constant* $D_0 > 0$ *such that*

$$\mathcal{N}_{\mathscr{K}}(\lambda) \geq D_0 \lambda^{-s_1} \log(1/\lambda), \text{ for } 0 < \lambda < 1. \tag{2.27}$$

Remark. The bound of $\mathcal{N}_{\mathscr{K}}(\lambda)$ is sharp when $s_1 = s_2 = s$, that is, the upper bound of eigenvalues matches the lower bound.

Remark. In some literature [49], there is another assumption on $\mathcal{N}_K(\lambda)$ that

$$\mathcal{N}_K(\lambda) \leq \tilde{C}_0 \lambda^{-s} \text{ for some } \tilde{C}_0 < \infty \text{ and } s > 0. \tag{2.28}$$

When $s = 1$, this assumption is trivial since L_K is of trace class. According to [17], if (2.26) holds for $s_2 = s$, then there is a constant $\tilde{C}_0 < \infty$ such that (2.28) holds. Thus, the upper bound (2.26) of $\mathcal{N}_{\mathscr{K}}(\lambda)$ is equivalent to that of $\mathcal{N}_K(\lambda)$ subject to the $\log(1/\lambda)$ factor. On the other hand, the following proposition tells us that the polynomial upper bound of λ_i can be derived from (2.28).

Proposition 2.5. *Suppose that there are constants $0 < \tilde{C}_0 < \infty$ and $s > 0$ such that $\mathcal{N}_K(\lambda) \leq \tilde{C}_0 \lambda^{-s}$ for any $0 < \lambda \leq \lambda_1$. Then there holds $\lambda_i \leq \left(\frac{i}{2\tilde{C}_0}\right)^{-1/s}$ for each i.*

Corollary 2.1. *Assume $|y|, |y'| \leq M$. Assume that $\lambda_i \leq D_2 i^{-1/s}, i = 1, 2, \cdots$, with $0 < D_2 < \infty$ and $0 < s < 1$. Under the regularity assumption (2.9) with $0 \leq r < 1/2$, we have the following error bounds.*

(i) For $N \geq \max\{4, e^{2r+1+s}\}$, by taking $\lambda = (N/\log N)^{-\frac{1}{2r+1+s}}$, we have

$$\mathbb{E} \left\| F_\lambda^{\mathbf{z}} - F_\rho \right\|_\rho \leq C^* \left(\frac{N}{\log N} \right)^{-\frac{r+1/2}{2r+1+s}}, \tag{2.29}$$

where C^* is a universal constant that will be specified in the proof.

(ii) For $n \geq e^{2r+1+s}$, by taking $\lambda = (n/\log n)^{-\frac{1}{2r+1+s}}$, we have

$$\mathbb{E} \left\| F_\lambda^{\mathbf{z}'} - F_\rho \right\|_\rho \leq C^{*\prime} \left(\frac{n}{\log n} \right)^{-\frac{r+1/2}{2r+1+s}}, \tag{2.30}$$

where $C^{*\prime}$ is a universal constant that will be specified in the proof.

Remark. The convergence rate of KKRR given in Corollary 2.1 matches the minimax rate $n^{-\frac{r+1/2}{2r+1+s}}$ for classical KRR [17] up to a logarithmic factor $\log^{\frac{r+1/2}{2r+1+s}} n$. This is reasonable since we now learn a function equivalent to an operator, whose

rank can be infinity, in a more complicated space $\mathrm{HS}(\mathcal{H}_K)$. From the minimax lower bound to be derived later, we shall see that this logarithmic factor is intrinsic and cannot be eliminated.

Proposition 2.6. *There hold the following estimates of* $\mathcal{N}_{\mathscr{K}}(\lambda)$.

(i) (Upper Bound.) Assume $\lambda_i \le \hat{D}_2 \exp(-t_2 i), i = 1, 2, \cdots$, *with some constants* $0 < \hat{D}_2 < \infty$ *and* $t_2 > 0$. *Then there is a constant* $\hat{C}_0 < +\infty$ *such that*

$$\mathcal{N}_{\mathscr{K}}(\lambda) \le \hat{C}_0 \log^2(1/\lambda), \text{ for } 0 < \lambda < e^{-1}. \tag{2.31}$$

(ii) (Lower Bound.) Assume $\lambda_i \ge \hat{D}_1 \exp(-t_1 i), i = 1, 2, \cdots$, *with some positive constants* $\hat{D}_1$ *and* $t_1 < +\infty$. *Then, for* $0 < \lambda < 1$ *and*

$$\lambda \le \min\left\{\hat{D}_1^4, \hat{D}_1^2 \exp(-8t_1)\right\},$$

there holds

$$\mathcal{N}_{\mathscr{K}}(\lambda) \ge \hat{D}_0 \log^2(1/\lambda) \tag{2.32}$$

with a universal constant $\hat{D}_0 > 0$.

Remark. For the above scenarios where the eigenvalues $\{\lambda_i\}$ decay exponentially, the bound derived in Proposition 2.6 is still sharp. It is easy to verify that $\mathcal{N}_K(\lambda) = O(\log(1/\lambda))$. Thus, the bound of $\mathcal{N}_{\mathscr{K}}(\lambda)$ is greater than that of $\mathcal{N}_K(\lambda)$ due to the $\log(1/\lambda)$ term that will tends to infinity as λ tends to 0.

Corollary 2.2. *Assume* $|y|, |y'| \le M$. *Under the regularity assumption (2.9) with* $0 \le r < 1/2$ *and the assumption* $\lambda_i \le \hat{D}_2 \exp(-t_2 i), i = 1, 2, \cdots$, *with some constants* $0 < \hat{D}_2 < \infty$ *and* $t_2 > 0$, *we have the following error bounds.*

(i) If we take $\lambda = N^{-1}$, *then for* $N \ge 4$, *there holds*

$$\mathbb{E}\left[\left\|F_\lambda^{\mathbf{z}} - F_\rho\right\|_\rho\right] \le D^* N^{-1/2} \log N, \tag{2.33}$$

where D^* *is a universal constant that will be specified in the proof.*

(ii) If we take $\lambda = n^{-1}$, then for $n \geq 3$, there holds

$$\mathbb{E}\left[\left\|F_\lambda^{\mathbf{z}'} - F_\rho\right\|_\rho\right] \leq D^{*\prime} n^{-1/2} \log n, \tag{2.34}$$

where $D^{*\prime}$ is a universal constant that will be specified in the proof.

Remark. The rate $\lambda_i = O(\exp(-ti^2))$ of decay for some $t > 0$, also known as the Gaussian-type decay, has been considered in some existing works of kernel ridge regression [28, 88]. For this Gaussian-type decay, we have $\mathcal{N}_K(\lambda) = O(\log^{1/2}(1/\lambda))$ and $\mathcal{N}_{\mathscr{K}}(\lambda) = O(\log(1/\lambda))$. Since the technical proof (with the help of the polar coordinate system) is similar to the case where eigenvalues decay exponentially, we omit the details and the corresponding corollary here for conciseness.

2.3.3 Minimax Lower Bound for $\mathscr{K}$-based Pairwise Regression

To our best knowledge, there is no result on the minimax rate of pairwise learning via the Kronecker product kernels. Thus to evaluate the rate derived in (2.30), motivated by [17,35], in what follows, we study the minimax rate for $\mathscr{K}$-based pairwise learning. The minimax rate is derived under the training set $\mathbf{z}' = \{(x_i, u_i, y_i')\}_{i=1}^n$ and is given over all distributions in the following two classes.

Let $\mathcal{P}(s_1, s_2, r)$ be the set of Borel probability measures on $\mathcal{X}^2 \times \mathcal{Y}$ such that:

1. $d\rho(x, u, y') = d\rho_{\mathcal{X}}(x) d\rho_{\mathcal{X}}(u) d\rho(y'|(x, u))$,
2. $|y'| \leq M$ almost surely,
3. $F_\rho = L_{\mathscr{K}}^r(G_\rho)$ for some $G_\rho \in \mathcal{H}_{\mathscr{K}}$ with $\|G_\rho\|_{\mathscr{K}} \leq R$, where $R > 0$ is a constant.
4. $D_1 i^{-1/s_1} \leq \lambda_i \leq D_2 i^{-1/s_2}$ for each i with universal constants $D_1 > 0, D_2 < \infty$ and $0 < s_1, s_2 < 1$.

Let $\mathcal{P}(t_1, t_2, r)$ be another class of Borel probability measures satisfying the first three conditions in the definition of $\mathcal{P}(s_1, s_2, r)$ and the condition that $\hat{D}_1 \exp(-t_1 i) \leq \lambda_i \leq \hat{D}_2 \exp(-t_2 i)$ for each i with $\hat{D}_1 > 0, \hat{D}_2 < \infty$ and $0 < t_1, t_2 < +\infty$.

Theorem 2.3. *Let $F_{\mathbf{z}'} \in \mathcal{H}_{\mathscr{K}}$ be the output of any learning algorithm according to the observations $\mathbf{z}' = \{(x_i, u_i, y_i')\}_{i=1}^n$. Then, for $0 < \delta < 1$,we have*

$$\lim_{n\to\infty} \inf_{F_{\mathbf{z}'}} \sup_{\rho\in\mathcal{P}(s_1,s_2,r)} \mathbb{P}_{\mathbf{z}'\sim\rho^n} \left\{ \|F_{\mathbf{z}'} - F_\rho\|_\rho^2 \geq \frac{1}{4}\tau_\delta \left(\frac{n}{\delta}\right)^{-\frac{s_2(2r+1)}{s_1(2r+1)+s_1 s_2}} \log^{\frac{s_2(2r+1)}{s_1(2r+1+s_2)}} \frac{n}{\delta} \right\}$$
$$\geq 1 - \delta \tag{2.35}$$

and

$$\lim_{n\to\infty} \inf_{F_{\mathbf{z}'}} \sup_{\rho\in\mathcal{P}(t_1,t_2,r)} \mathbb{P}_{\mathbf{z}'\sim\rho^n} \left\{ \|F_{\mathbf{z}'} - F_\rho\|_\rho^2 \geq \hat{\tau}_\delta \left(\frac{\log^2 \frac{n}{\delta}}{n/\delta}\right) \right\} \geq 1 - \delta, \tag{2.36}$$

where $\tau_\delta, \hat{\tau}_\delta$ are constants independent of n (but they may depend on δ).

Remark. Comparing with the minimax rate given in (2.35) and (2.36), the rates derived in (2.30) and (2.34) are optimal when $s_1 = s_2 = s$.

2.4 Proofs of Convergence Results

The results given in section 2.3 are proved in this section. We state some technical lemmas in this section for the proof of the main theorems and provide a detailed proof of these lemmas in Section 2.7.

2.4.1 Statement of Technical Lemmas

The sample-free analogy of (2.4),

$$F_\lambda := \arg\min_{F\in\mathcal{H}_{\mathscr{K}}} \left\{ \|F - F_\rho\|_\rho^2 + \lambda \|F\|_{\mathscr{K}}^2 \right\}, \tag{2.37}$$

has been widely used in learning theory. Equation (2.17) implies that

$$F_\lambda = \left(L_{\mathscr{K}} + \lambda I\right)^{-1} L_{\mathscr{K}} F_\rho. \tag{2.38}$$

We use the following decompositions,

$$F_\lambda^{\mathbf{z}} - F_\rho = \left(F_\lambda^{\mathbf{z}} - F_\lambda\right) + \left(F_\lambda - F_\rho\right), \tag{2.39}$$

$$F_\lambda^{\mathbf{z}'} - F_\rho = \left(F_\lambda^{\mathbf{z}'} - F_\lambda\right) + \left(F_\lambda - F_\rho\right). \tag{2.40}$$

The error analysis of $F_\lambda^{\mathbf{z}} - F_\lambda$ and $F_\lambda^{\mathbf{z}'} - F_\lambda$ is based on the so-called *first and second order decomposition* of the difference between the inverse of two operators proposed by [49].

Lemma 2.1. *Let A and B be two invertible operators on a Banach space. We have*

$$A^{-1} - B^{-1} = B^{-1}\left(B - A\right)A^{-1} = A^{-1}(B - A)B^{-1}. \tag{2.41}$$

Moreover,

$$A^{-1} - B^{-1} = B^{-1}(B - A)B^{-1} + B^{-1}(B - A)A^{-1}(B - A)B^{-1}. \tag{2.42}$$

To bound $F_\lambda^{\mathbf{z}'} - F_\lambda$, we need the help of the following three quantities. Define

$$\mathcal{Q}_{\mathbf{z}',\lambda} = \left\| \left(L_{\mathscr{K}} + \lambda I\right)^{1/2} \left(L_{\mathscr{K}}^{\mathbf{x}'} + \lambda I\right)^{-1} \left(L_{\mathscr{K}} + \lambda I\right)^{1/2} \right\|_{\mathrm{op}},$$

$$\mathcal{P}_{\mathbf{z}',\lambda} = \left\| \left(L_{\mathscr{K}} + \lambda I\right)^{-1/2} \left(L_{\mathscr{K}} - L_{\mathscr{K}}^{\mathbf{x}'}\right) \right\|_{\mathrm{op}},$$

$$\mathcal{S}_{\mathbf{z}',\lambda} = \left\| \left(L_{\mathscr{K}} + \lambda I\right)^{-1/2} \left(\frac{1}{n} S_{\mathbf{x}'}^T Y' - L_{\mathscr{K}} F_\rho\right) \right\|_{\mathscr{K}},$$

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm. We provide their estimates in the following Lemma.

Lemma 2.2. *Assume $|y'| \le M$. For any $\theta > 0$, one has*

$$\mathbb{E}\mathcal{S}_{\mathbf{z}',\lambda}^\theta \le C_{\theta,\mathcal{S}'} \mathcal{A}_{n,\lambda}^\theta, \tag{2.43}$$

$$\mathbb{E}\mathcal{P}_{\mathbf{z}',\lambda}^\theta \le C_{\theta,\mathcal{P}'} \mathcal{A}_{n,\lambda}^\theta, \text{ and} \tag{2.44}$$

$$\mathbb{E}\mathcal{Q}_{\mathbf{z}',\lambda}^\theta \le C_{\theta,\mathcal{Q}'} \left(\frac{\mathcal{A}_{n,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{n,\lambda}^2}{\lambda} + 1\right)^\theta, \tag{2.45}$$

where $C_{\theta,\mathcal{S}'}, C_{\theta,\mathcal{P}'}$ and $C_{\theta,\mathcal{Q}'}$ are universal constants to be specified in the proof.

Similarly, we introduce

$$\mathcal{Q}_{\mathbf{z},\lambda} = \left\| (L_{\mathscr{K}} + \lambda I)^{1/2} \left(L^{\mathbf{x}}_{\mathscr{K}} + \lambda I\right)^{-1} (L_{\mathscr{K}} + \lambda I)^{1/2} \right\|_{\mathrm{op}},$$

$$\mathcal{P}_{\mathbf{z},\lambda} = \left\| (L_{\mathscr{K}} + \lambda I)^{-1/2} \left(L_{\mathscr{K}} - L^{\mathbf{x}}_{\mathscr{K}}\right) \right\|_{\mathrm{op}}, \text{ and}$$

$$\mathcal{S}_{\mathbf{z},\lambda} = \left\| (L_{\mathscr{K}} + \lambda I)^{-1/2} \left(\frac{1}{N^2} S^T_{\mathbf{x}} Y - L_{\mathscr{K}} F_\rho \right) \right\|_{\mathscr{K}},$$

for the error analysis of $F^{\mathbf{z}}_\lambda - F_\lambda$.

Lemma 2.3. *Assume that $|y| \le M$. Then there are constants $C_{\mathcal{Q}}, C_{\mathcal{P}}, C_{\mathcal{S}}$ such that*

$$\mathbb{E}\mathcal{Q}^2_{\mathbf{z},\lambda} \le C_{\mathcal{Q}} \left(\frac{\mathcal{A}^2_{N,\lambda}}{\lambda} + \frac{\mathcal{A}^4_{N,\lambda}}{\lambda^2} \right) + 2, \tag{2.46}$$

$$\mathbb{E}\mathcal{P}^2_{\mathbf{z},\lambda} \le C_{\mathcal{P}} \mathcal{A}^2_{N,\lambda}, \text{ and} \tag{2.47}$$

$$\mathbb{E}\mathcal{S}^2_{\mathbf{z},\lambda} \le C_{\mathcal{S}} \mathcal{A}^2_{N,\lambda}. \tag{2.48}$$

The proofs of Lemma 2.2 and Lemma 2.3 are given in Section 2.7.

2.4.2 Proofs of the Upper Bounds

For the error bound of $F_\lambda - F_\rho$, motivated by [66], we have the following estimate.

Proposition 2.7. *If (2.9) holds for $0 \le r < 1/2$, then*

$$\|F_\lambda - F_\rho\|_\rho \le \|G_\rho\|_{\mathscr{K}} \lambda^{r+1/2}. \tag{2.49}$$

Proof. By (2.38) and (2.9), we have

$$F_\rho = \sum_{i,j=1}^{\infty} F_{\rho_{ij}} \Phi_{ij} = \sum_{i,j=1}^{\infty} (\lambda_i \lambda_j)^r G_{\rho_{ij}} \Phi_{ij},$$

$$F_\lambda = (L_{\mathscr{K}} + \lambda I)^{-1} L_{\mathscr{K}} \sum_{i,j=1}^{\infty} F_{\rho_{ij}} \Phi_{ij} = \sum_{i,j=1}^{\infty} \frac{(\lambda_i \lambda_j)^{1+r}}{\lambda_i \lambda_j + \lambda} G_{\rho_{ij}} \Phi_{ij},$$

and

$$F_\lambda - F_\rho = \sum_{i,j=1}^{\infty} \frac{(\lambda_i \lambda_j)^r \lambda}{\lambda_i \lambda_j + \lambda} G_{\rho_{ij}} \Phi_{ij}.$$

Thus

$$\|F_\lambda - F_\rho\|_\rho^2 = \left\| L_{\mathscr{K}}^{1/2} (F_\lambda - F_\rho) \right\|_{\mathscr{K}}^2 = \sum_{i,j=1}^{\infty} \left(\frac{(\lambda_i \lambda_j)^{1/2+r} \lambda}{\lambda_i \lambda_j + \lambda} G_{\rho_{ij}} \right)^2$$

$$= \lambda^{2r+1} \sum_{i,j=1}^{\infty} \left(\frac{\lambda_i \lambda_j}{\lambda_i \lambda_j + \lambda} \right)^{1+2r} \left(\frac{\lambda}{\lambda_i \lambda_j + \lambda} \right)^{1-2r} G_{\rho_{ij}}^2 \leq \lambda^{2r+1} \|G_\rho\|_{\mathscr{K}}^2.$$

□

The next proposition is the analysis of $F_\lambda^{\mathbf{z}} - F_\lambda$.

Proposition 2.8. *There holds*

$$\mathbb{E}\left[\|F_\lambda^{\mathbf{z}} - F_\lambda\|_\rho \right] \leq \tilde{C} \mathcal{A}_{N,\lambda} \left(\frac{\mathcal{A}_{N,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{N,\lambda}^2}{\lambda} + 1 \right), \tag{2.50}$$

where $\tilde{C}$ is a constant that will be specified in the proof.

Proof. By (2.21) and (2.38), we have

$$F_\lambda^{\mathbf{z}} - F_\lambda = (L_{\mathscr{K}}^{\mathbf{x}} + \lambda I)^{-1} \frac{1}{N^2} S_{\mathbf{x}}^T Y - (L_{\mathscr{K}} + \lambda I)^{-1} L_{\mathscr{K}} F_\rho$$

$$= \left((L_{\mathscr{K}}^{\mathbf{x}} + \lambda I)^{-1} \frac{1}{N^2} S_{\mathbf{x}}^T Y - (L_{\mathscr{K}}^{\mathbf{x}} + \lambda I)^{-1} L_{\mathscr{K}} F_\rho \right)$$

$$+ \left((L_{\mathscr{K}}^{\mathbf{x}} + \lambda I)^{-1} L_{\mathscr{K}} F_\rho - (L_{\mathscr{K}} + \lambda I)^{-1} L_{\mathscr{K}} F_\rho \right) =: \mathcal{T}_1 + \mathcal{T}_2. \tag{2.51}$$

For $\mathcal{T}_1$, since $\left\| L_{\mathscr{K}}^{1/2} (L_{\mathscr{K}} + \lambda I)^{-1/2} \right\|_{\text{op}} \leq 1$ and

$$\|F\|_\rho = \|L_{\mathscr{K}}^{1/2} F\|_{\mathscr{K}} = \left\| L_{\mathscr{K}}^{1/2} (L_{\mathscr{K}} + \lambda I)^{-1/2} (L_{\mathscr{K}} + \lambda I)^{1/2} F \right\|_{\mathscr{K}}$$

$$\leq \left\| (L_{\mathscr{K}} + \lambda I)^{1/2} F \right\|_{\mathscr{K}},$$

we obtain

$$\|\mathcal{T}_1\|_\rho \leq \left\| (L_{\mathscr{K}} + \lambda I)^{1/2} \mathcal{T}_1 \right\|_{\mathscr{K}}$$

$$\leq \left\| (L_{\mathscr{K}} + \lambda I)^{1/2} (L_{\mathscr{K}}^{\mathbf{x}} + \lambda I)^{-1} (L_{\mathscr{K}} + \lambda I)^{1/2} \right\|_{\text{op}} \left\| (L_{\mathscr{K}} + \lambda I)^{-1/2} \left(\frac{1}{N^2} S_{\mathbf{x}}^T Y - L_{\mathscr{K}} F_\rho \right) \right\|_{\mathscr{K}}$$

$$= \mathcal{Q}_{\mathbf{z},\lambda} \mathcal{S}_{\mathbf{z},\lambda}.$$

By (2.46), (2.48) and the Cauchy-Schwarz inequality, there holds

$$\mathbb{E}\mathcal{Q}_{\mathbf{z},\lambda}\mathcal{S}_{\mathbf{z},\lambda} \leq \left(\mathbb{E}\mathcal{Q}^2_{\mathbf{z},\lambda}\mathbb{E}\mathcal{S}^2_{\mathbf{z},\lambda}\right)^{1/2} \leq \sqrt{(C_{\mathcal{Q}}+2)C_{\mathcal{S}}}\mathcal{A}_{N,\lambda}\left(\frac{\mathcal{A}_{N,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}^2_{N,\lambda}}{\lambda} + 1\right),$$

where the last inequality comes from the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, for all $a, b \geq 0$. In conclusion,

$$\mathbb{E}\|\mathcal{T}_1\|_\rho \leq \sqrt{(C_{\mathcal{Q}}+2)C_{\mathcal{S}}}\mathcal{A}_{N,\lambda}\left(\frac{\mathcal{A}_{N,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}^2_{N,\lambda}}{\lambda} + 1\right). \tag{2.52}$$

For $\mathcal{T}_2$, by the first order decomposition (2.41), we have

$$\begin{aligned}
\|\mathcal{T}_2\|_\rho &\leq \left\|(L_{\mathscr{K}}+\lambda I)^{1/2}\,\mathcal{T}_2\right\|_{\mathscr{K}} = \left\|(L_{\mathscr{K}}+\lambda I)^{1/2}\left((L^{\mathbf{x}}_{\mathscr{K}}+\lambda I)^{-1} - (L_{\mathscr{K}}+\lambda I)^{-1}\right)L_{\mathscr{K}}F_\rho\right\|_{\mathscr{K}} \\
&\leq \left\|(L_{\mathscr{K}}+\lambda I)^{1/2}\,(L^{\mathbf{x}}_{\mathscr{K}}+\lambda I)^{-1}\,(L_{\mathscr{K}}-L^{\mathbf{x}}_{\mathscr{K}})\,(L_{\mathscr{K}}+\lambda I)^{-1}\,L_{\mathscr{K}}F_\rho\right\|_{\mathscr{K}} \\
&= \left\|(L_{\mathscr{K}}+\lambda I)^{1/2}\,(L^{\mathbf{x}}_{\mathscr{K}}+\lambda I)^{-1}\,(L_{\mathscr{K}}+\lambda I)^{1/2}\,(L_{\mathscr{K}}+\lambda I)^{-1/2}\,(L_{\mathscr{K}}-L^{\mathbf{x}}_{\mathscr{K}})\,(L_{\mathscr{K}}+\lambda I)^{-1}\,L_{\mathscr{K}}F_\rho\right\|_{\mathscr{K}} \\
&\leq \mathcal{Q}_{\mathbf{z},\lambda}\mathcal{P}_{\mathbf{z},\lambda}\left\|(L_{\mathscr{K}}+\lambda I)^{-1}\,L_{\mathscr{K}}F_\rho\right\|_{\mathscr{K}} = \mathcal{Q}_{\mathbf{z},\lambda}\mathcal{P}_{\mathbf{z},\lambda}\,\|F_\lambda\|_{\mathscr{K}}\,.
\end{aligned} \tag{2.53}$$

As a result of (2.46) and (2.47), we obtain

$$\mathbb{E}\mathcal{Q}_{\mathbf{z},\lambda}\mathcal{P}_{\mathbf{z},\lambda} \leq \sqrt{\mathbb{E}\mathcal{Q}^2_{\mathbf{z},\lambda}\mathbb{E}\mathcal{P}^2_{\mathbf{z},\lambda}} \leq \sqrt{(C_{\mathcal{Q}}+2)C_{\mathcal{P}}}\mathcal{A}_{N,\lambda}\left(\frac{\mathcal{A}_{N,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}^2_{N,\lambda}}{\lambda} + 1\right). \tag{2.54}$$

Moreover, it holds

$$\|F_\lambda\|_{\mathscr{K}} = \left\|(L_{\mathscr{K}}+\lambda I)^{-1}\,L_{\mathscr{K}}F_\rho\right\|_{\mathscr{K}} \leq \left\|(L_{\mathscr{K}}+\lambda I)^{-1}\,L_{\mathscr{K}}\right\|_{\mathrm{op}}\|F_\rho\|_{\mathscr{K}} \leq \|F_\rho\|_{\mathscr{K}}\,. \tag{2.55}$$

Due to (2.53), (2.54) and (2.55), we obtain

$$\mathbb{E}\,\|\mathcal{T}_2\|_\rho \leq \|F_\rho\|_{\mathscr{K}}\sqrt{(C_{\mathcal{Q}}+2)C_{\mathcal{P}}}\mathcal{A}_{N,\lambda}\left(\frac{\mathcal{A}_{N,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}^2_{N,\lambda}}{\lambda} + 1\right). \tag{2.56}$$

Combining (2.51), (2.52) and (2.56) together, we conclude

$$\mathbb{E}\,\|F^{\mathbf{z}}_\lambda - F_\lambda\|_\rho \leq \tilde{C}\mathcal{A}_{N,\lambda}\left(\frac{\mathcal{A}_{N,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}^2_{N,\lambda}}{\lambda} + 1\right),$$

where $\tilde{C} = \|F_\rho\|_{\mathscr{K}}\sqrt{(C_{\mathcal{Q}}+2)C_{\mathcal{P}}} + \sqrt{(C_{\mathcal{Q}}+2)C_{\mathcal{S}}}$. □

Proof of Theorem 2.2. Since

$$\|F_\lambda^{\mathbf{z}} - F_\rho\|_\rho \le \|F_\lambda^{\mathbf{z}} - F_\lambda\|_\rho + \|F_\lambda - F_\rho\|_\rho ,$$

the proof of (2.25) is a direct corollary of Proposition 2.7 and Proposition 2.8 with $C = \tilde{C} + \|G_\rho\|_{\mathscr{K}}$. □

Proposition 2.9. *There holds*

$$\mathbb{E}\left[\left\|F_\lambda^{\mathbf{z}'} - F_\lambda\right\|_\rho\right] \le \tilde{C}'\mathcal{A}_{n,\lambda}\left(\frac{\mathcal{A}_{n,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{n,\lambda}^2}{\lambda} + 1\right), \tag{2.57}$$

with some universal constant $\tilde{C}'$ to be specified in the proof.

Proof. This proof parallels the proof of Proposition 2.8. First,

$$\left\|F_\lambda^{\mathbf{z}'} - F_\rho\right\|_\rho \le \mathcal{Q}_{\mathbf{z}',\lambda}\mathcal{S}_{\mathbf{z}',\lambda} + \mathcal{Q}_{\mathbf{z}',\lambda}\mathcal{P}_{\mathbf{z}',\lambda}\|F_\lambda\|_{\mathscr{K}} . \tag{2.58}$$

By (2.43) and (2.45) with $\theta = 2$, we have

$$\mathbb{E}\mathcal{Q}_{\mathbf{z}',\lambda}\mathcal{S}_{\mathbf{z}',\lambda} \le \sqrt{\mathbb{E}\mathcal{Q}_{\mathbf{z}',\lambda}^2\mathbb{E}\mathcal{S}_{\mathbf{z}',\lambda}^2} \le \sqrt{C_{2,\mathcal{Q}'}C_{2,\mathcal{S}'}}\mathcal{A}_{n,\lambda}\left(\frac{\mathcal{A}_{n,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{n,\lambda}^2}{\lambda} + 1\right). \tag{2.59}$$

Furthermore, thanks to (2.44) and (2.45), there holds

$$\mathbb{E}\mathcal{Q}_{\mathbf{z}',\lambda}\mathcal{P}_{\mathbf{z}',\lambda} \le \sqrt{\mathbb{E}\mathcal{Q}_{\mathbf{z}',\lambda}^2\mathbb{E}\mathcal{P}_{\mathbf{z}',\lambda}^2} \le \sqrt{C_{2,\mathcal{Q}'}C_{2,\mathcal{P}'}}\mathcal{A}_{n,\lambda}\left(\frac{\mathcal{A}_{n,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{n,\lambda}^2}{\lambda} + 1\right). \tag{2.60}$$

By substituting (2.59), (2.60) and (2.55) into (2.58), we conclude (2.57) with $\tilde{C}' = \sqrt{C_{2,\mathcal{Q}'}C_{2,\mathcal{S}'}} + \sqrt{C_{2,\mathcal{Q}'}C_{2,\mathcal{P}'}}\,\|F_\rho\|_{\mathscr{K}}$. □

Proof of Theorem 2.1. By Proposition 2.7 and Proposition 2.9, we obtain

$$\left\|F_\lambda^{\mathbf{z}'} - F_\rho\right\|_\rho \le \left\|F_\lambda^{\mathbf{z}'} - F_\lambda\right\|_\rho + \|F_\lambda - F_\rho\|_\rho$$

$$\le (\tilde{C}' + \|G_\rho\|_{\mathscr{K}})\left[\mathcal{A}_{n,\lambda}\left(\frac{\mathcal{A}_{n,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{n,\lambda}^2}{\lambda} + 1\right) + \lambda^{r+1/2}\right].$$

□

Proof of Corollary 2.1. From the selection of λ, it's obvious that

$$\mathcal{A}_{N,\lambda} \leq \left(\sqrt{\frac{C_0}{2r+1+s}}+1\right) N^{-\frac{r+1/2}{2r+1+s}} \log^{\frac{r+1/2}{2r+1+s}} N,$$

$$\frac{\mathcal{A}_{N,\lambda}}{\sqrt{\lambda}} \leq \left(\sqrt{\frac{C_0}{2r+1+s}}+1\right).$$

Note that $\log N \geq 1$ and $\log^2 N \geq \log N$. From (2.25),

$$\mathbb{E}\left\|F_\lambda^{\mathbf{z}} - F_\rho\right\|_\rho \leq C_0^* N^{-\frac{r+1/2}{2r+1+s}} \log^{\frac{r+1/2}{2r+1+s}} N$$

with $C_* = C((\sqrt{C_0/(s+1)}+1)^3+1)$.

The proof of (2.30) parallels that of (2.29) by replacing C with C'. Here, $C^{*\prime} = C'((\sqrt{C_0/(s+1)}+1)^3+1)$.

□

The proof of Corollary 2.2 is similar to that of corollary 2.1 and is omitted.

2.4.3 Proofs of the Minimax Lower Bounds

Proof of Theorem 2.3. For $F = L_{\mathscr{K}}^r G$ with $G \in \mathcal{H}_{\mathscr{K}}$ and $\|G\|_{\mathscr{K}} \leq R$, we define the corresponding probability measure ρ_F by

$$d\rho_F(x,u,y') = \left[\frac{B+F(x,u)}{2B} d\delta_B(y') + \frac{B-F(x,u)}{2B} d\delta_{-B}(y')\right] d\mu(x) d\mu(u)$$

with $B = 4\kappa^{4r+2}R$ and $d\delta_B$ being the Dirac measure massing at B, where μ is a probability measure on $\mathcal{X}$ such that $D_1 i^{-1/s_1} \leq \lambda_i \leq D_2 i^{-1/s_2}$ for each i. Hence $\rho_F \in \mathcal{P}(s_1, s_2, r)$ with $|y'| \leq B$ and F being the target function.

Let s_γ be the cardinality of the set $\mathcal{S}_\gamma = \{(i,j) : \gamma \leq ij \leq 2\gamma\}$ for a given integer $\gamma > 0$. According to Varshamov-Gilbert's Lemma (c.f., Lemma 4.7, [51]), for any integer $s_\gamma \geq 8$, there is $\Theta = \{w_0, w_1, ..., w_m\} \subset \{0,1\}^{2\gamma \times 2\gamma}$ with $w_i^{(st)}$ being the (s,t)-th component of w_i for $s,t = 1,2,\cdots,2\gamma$, such that

1. $w_0 = (0)_{2\gamma\times 2\gamma}$;

2. $w_i^{(st)} = 0$ for $(s,t) \notin \mathcal{S}_\gamma$ and $i = 1, 2, \cdots m$;

2. For $i \neq j$, $\|w_i - w_j\|_F^2 = \sum_{\gamma \leq st \leq 2\gamma} \left(w_i^{(st)} - w_j^{(st)}\right)^2 \geq s_\gamma/4$;

3. $\log m \geq s_\gamma/8$.

We construct a function

$$G_i = \sum_{\gamma \leq st \leq 2\gamma} w_i^{(st)} R s_\gamma^{-1/2} \Phi_{st},$$

which is obviously in $\mathcal{H}_{\mathscr{K}}$ with

$$\|G_i\|_{\mathscr{K}}^2 = \sum_{\gamma \leq st \leq 2\gamma} \left(w_i^{(st)}\right)^2 s_\gamma^{-1} R^2 \leq R^2.$$

Define $F_i = L_{\mathscr{K}}^r G_i$. Then $\rho_{F_i} \in \mathcal{P}(s_1, s_2, r)$. By the reproducing property, there holds

$$\|F_i\|_\infty \leq \kappa^2 \left\|L_{\mathscr{K}}^r G_i\right\|_K \leq \kappa^{4r+2} R \leq \frac{B}{4}.$$

Let ξ_m be a random variable drawn from the uniform distribution on $\{1, 2, \cdots, m\}$. According to a standard argument through Fano's Lemma, see, for example, [73], it holds

$$\begin{aligned}\inf_{F_{\mathbf{z}}'} \sup_{\rho \in \mathcal{P}(s_1,s_2,r)} \mathbb{P}_{\mathbf{z}' \sim \rho^n} \left\{ \|F_{\mathbf{z}}' - F_\rho\|_\rho^2 \geq \frac{1}{4} \min_{i \neq j} \|F_i - F_j\|_\rho^2 \right\} \\ \geq 1 - \frac{\mathbb{E}_{X'}[\mathbb{I}_{X'}(Y'; \xi_m)] + \log 2}{\log m},\end{aligned}$$

where $\mathbb{I}_{X'}(Y'; \xi_m)$ is the mutual information between Y' and ξ_m conditioned on X'.

Let

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}\left(\rho_{F_i} \| \rho_{F_j}\right) = \int_{\mathcal{X}\times\mathcal{X}} \left\{ \frac{B + F_i(x,u)}{2B} \log\left(1 + \frac{F_i(x,u) - F_j(x,u)}{B + F_j(x,u)}\right) \right. \\ \left. + \frac{B - F_i(x,u)}{2B} \log\left(1 - \frac{F_i(x,u) - F_j(x,u)}{B - F_j(x,u)}\right) \right\} d\mu(x) d\mu(u)\end{aligned}$$

be the the KL-divergence between ρ_{F_i} and ρ_{F_j}. Since $\|F_i\|_\infty, \|F_j\|_\infty \le B/4$, there holds

$$\left|\frac{F_i(x,u) - F_j(x,u)}{B + F_j(x,u)}\right| \le 2/3 < 1$$

and

$$\left|\frac{F_i(x,u) - F_j(x,u)}{B - F_j(x,u)}\right| \le 2/3 < 1.$$

By the elementary inequality $\log(1+t) \le t$ for $t > -1$, we obtain

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}\left(\rho_{F_i} \| \rho_{F_j}\right) &\le \int_{\mathcal{X}\times\mathcal{X}} \frac{F_i(x,u) - F_j(x,u)}{2B} \left\{\frac{B + F_i(x,u)}{B + F_j(x,u)} - \frac{B - F_i(x,u)}{B - F_j(x,u)}\right\} d\mu(x) d\mu(u) \\ &\le \frac{16}{15B^2} \|F_i - F_j\|_\rho^2.\end{aligned}$$

Note that

$$\frac{16}{15B^2} \|F_i - F_j\|_\rho^2 = \frac{16R^2}{15B^2 s_\gamma} \sum_{\gamma \le st \le 2\gamma} \lambda_s^{2r+1} \lambda_t^{2r+1} \left(w_i^{(st)} - w_j^{(st)}\right)^2 \le \frac{D_2^{2r+2}}{15\kappa^{2(4r+2)}} \gamma^{-\frac{2r+1}{s_2}}.$$

It holds that

$$\begin{aligned}\mathbb{I}_{X'}(Y'; \xi_m) &= \frac{1}{m} \sum_{i=1}^m \mathcal{D}_{\mathrm{KL}}\left((\rho_{F_i})^n \| \frac{1}{m} \sum_{j=1}^m (\rho_{F_j})^n\right) \le \frac{1}{m^2} \sum_{i,j=1}^m \mathcal{D}_{\mathrm{KL}}\left((\rho_{F_i})^n \| (\rho_{F_j})^n\right) \\ &= \frac{n}{m^2} \sum_{i \ne j} \mathcal{D}_{\mathrm{KL}}\left(\rho_{F_i} \| \rho_{F_j}\right) \le \frac{D_2^{2(2r+2)}}{15\kappa^{2(4r+2)}} n \gamma^{-\frac{2r+1}{s_2}}.\end{aligned}$$

Note that

$$s_\gamma = \sum_{i=1}^{2\gamma} \lfloor 2\gamma/i \rfloor - \sum_{i=1}^{\gamma-1} \lfloor (\gamma-1)/i \rfloor \ge \gamma \log \frac{4\gamma}{e^2}.$$

We introduce constants $\alpha > 0$ and $\tau_1 > 0$ and take $\gamma \ge \tau_1 (n/\delta)^{\frac{s_2}{2r+1+s_2}} / \log^\alpha \frac{n}{\delta}$. Then

$$\log m \ge s_\gamma/8 \ge \frac{1}{8} \gamma \log \frac{4\gamma}{e^2}$$

and

$$\frac{\mathbb{I}_{X'}(Y';\xi_m)}{\log m} \leq \delta$$

for τ_1 and n large enough, $0 < \delta < 1$, and $\alpha = \frac{s_2}{2r+1+s_2}$.

By taking γ to be the smallest integer greater than or equal to $\tau_1 (n/\delta)^{\frac{s_2}{2r+1+s_2}} / \log^{\alpha} \frac{n}{\delta}$, we obtain

$$\begin{aligned}
\|F_i - F_j\|_\rho^2 &= \sum_{\gamma \leq st \leq 2\gamma} R^2 s_\gamma^{-1} \lambda_s^{2r+1} \lambda_t^{2r+1} \left(w_i^{(st)} - w_j^{(st)} \right)^2 \\
&\geq R^2 s_\gamma^{-1} (2D_1^2 \gamma)^{-\frac{2r+1}{s_1}} \|w_i - w_j\|_F^2 \\
&\geq \frac{R^2}{(2D_1)^{\frac{2(2r+1)}{s_1}} 8} \gamma^{-\frac{2r+1}{s_1}} \geq \tau (n/\delta)^{-\frac{s_2(2r+1)}{s_1(2r+1)+s_1 s_2}} \log^{\frac{2r+1}{s_1}\alpha} \frac{n}{\delta},
\end{aligned}$$

for some constants τ.

In conclustion,

$$\begin{aligned}
&\inf_{F_{\mathbf{z}'}} \sup_{\rho \in \mathcal{P}(s_1,s_2,r)} \mathbb{P}_{\mathbf{z}' \sim \rho^n} \left\{ \|F_{\mathbf{z}'} - F_\rho\|_\rho^2 \geq \frac{1}{4} \tau \delta^{\frac{s_2(2r+1)}{s_1(2r+1)+s_1 s_2}} n^{-\frac{s_2(2r+1)}{s_1(2r+1)+s_1 s_2}} \log^{\frac{s_2(2r+1)}{s_1(2r+1+s_2)}} \frac{n}{\delta} \right\} \\
&\qquad \geq 1 - \delta - \log 2 / \log m.
\end{aligned}$$

This completes the proof of (2.35) by noting that $m \to \infty$ as $n \to \infty$.

To prove (2.36), consider the set $\mathcal{R}_\gamma = \{(i,j) : \gamma_1 \leq i + j \leq \gamma_2\}$ with cardinality r_γ, where γ_1 and γ_2 are two integers to be specified later. Note that for $(i,j) \in \mathcal{R}_\gamma$, there holds $e^{-t_1 \gamma_2} \leq \lambda_i \lambda_j \leq e^{-t_2 \gamma_1}$. We can obtain (2.36) by the similar argument as a result of

$$\begin{aligned}
r_\gamma &= \left(\frac{1}{2}\gamma_2^2 - \frac{\gamma_2}{2} \right) - \left(\frac{1}{2}(\gamma_1 - 1)^2 - \frac{\gamma_1 - 1}{2} \right) \\
&\geq C_\delta \left(\log^2 \frac{n}{\delta} - \log^2 \frac{n}{2\delta} \right) = 2 C_\delta \log 2 \log \frac{n}{\sqrt{2}\delta}
\end{aligned}$$

by taking $\gamma_2 = \lceil \frac{1}{t_1(2r+1)} (\log \frac{n}{\delta} - 2 \log \log \frac{n}{\delta}) \rceil$ and $\gamma_1 = \lfloor \frac{1}{t_2(2r+1)} (\log \frac{n}{2\delta} - \log \left(c_\delta \log \frac{n}{\delta} \right)) \rfloor$, for some constants C_δ and c_δ. □

2.5 Proofs of Propositions in Section 2.2

In this section, we prove the properties of Kronecker product pairwise kernels stated in section 2.2.

Proof of Proposition 2.1. According to [5], $\Phi_{ij} \in \mathcal{H}_{\mathscr{K}}$ and

$$\langle \Phi_{ij}, \Phi_{st} \rangle_{\mathscr{K}} = \langle \phi_i, \phi_s \rangle_K \langle \phi_j, \phi_t \rangle_K = \delta_{((i,j),(s,t))},$$

where δ is the Kronecker delta function. This proves the orthonormality of $\{\Phi_{ij}\}_{i,j=1}^{\infty}$. Moreover, by Mercer's Theorem [54] and the definition of $\mathscr{K}$, we obtain

$$\mathscr{K}((x,u),(x',u')) = K(x,x')K(u,u') = \left(\sum_{i=1}^{\infty} \phi_i(x)\phi_i(x')\right)\left(\sum_{j=1}^{\infty} \phi_j(u)\phi_j(u')\right)$$

$$= \sum_{i,j=1}^{\infty} (\phi_i(x)\phi_j(u))\,(\phi_i(x')\phi_j(u')) = \sum_{i,j=1}^{\infty} \Phi_{ij}(x,u)\Phi_{ij}(x',u'),$$

where the third equality is a result of the absolute and uniform convergence of $\sum_{i=1}^{\infty} \phi_i(x)\phi_i(x')$ thanks to Mercer's Theorem. We have finished the proof of *(a)* and *(b)*.

Item *(c)* is obtained by noting that

$$\begin{aligned} L_{\mathscr{K}}\Phi_{ij}(x',u') &= \int_{\mathcal{X}^2} \phi_i(x)\phi_j(u)K(x,x')K(u,u')d\rho_{\mathcal{X}}(x)d\rho_{\mathcal{X}}(u) \\ &= \int_{\mathcal{X}} \phi_i(x)K(x,x')d\rho_{\mathcal{X}}(x) \int_{\mathcal{X}} \phi_j(u)K(u,u')d\rho_{\mathcal{X}}(u) \\ &= L_K\phi_i(x')L_K\phi_j(u') = \lambda_i\lambda_j\Phi_{ij}(x',u'). \end{aligned}$$

Now we prove item *(d), (e)* and *(f)*. Note that

$$\Phi_{ij}(x,u) = \phi_i(x)\phi_j(u) = \phi_j(u)\phi_i(x) = \Phi_{ji}(u,x).$$

We obtain that

$$\mathscr{K}^{\mathrm{S}}_{(x,u)} = \frac{1}{2}\left(\mathscr{K}_{(x,u)} + \mathscr{K}_{(u,x)}\right) = \frac{1}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\Phi_{ij}(x,u)\left(\Phi_{ij}+\Phi_{ji}\right)$$

$$= \sum_{i<j}\Phi_{ij}(x,u)\left(\Phi_{ij}+\Phi_{ji}\right) + \sum_{i=1}^{\infty}\Phi_{ii}(x,u)\Phi_{ii}$$

and that $\mathscr{K}^{\mathrm{S}}_{(x,u)}$ is symmetric. As a consequence, $\left\{\frac{1}{\sqrt{2}}(\Phi_{ij}+\Phi_{ji})\right\}_{i<j} \cup \{\Phi_{ii}\}_{i=1}^{\infty}$ is the orthonormal basis of $\mathcal{H}_{\mathscr{K}^{\mathrm{S}}}$ spanned by $\mathscr{K}^{\mathrm{S}}_{(x,u)}$ and any $F \in \mathcal{H}_{\mathscr{K}^{\mathrm{S}}}$ is symmetric. On the other hand, for any $F = \sum_{i,j=1}^{\infty} F_{ij}\Phi_{ij} \in \mathcal{H}_{\mathscr{K}}$ such that $F(x,u) = F(u,x), x, u \in \mathcal{X}$, it holds

$$2F(x,u) = (F(x,u)+F(u,x)) = \sum_{i,j=1}^{\infty} F_{ij}\left(\Phi_{ij}(x,u)+\Phi_{ji}(x,u)\right) \in \mathcal{H}_{\mathscr{K}^{\mathrm{S}}}.$$

Since $\Phi_{ij}(x,u) = \Phi_{ji}(u,x)$, item *(f)* is a result of

$$L_{\mathscr{K}^{\mathrm{S}}}\left(\Phi_{ij}-\Phi_{ji}\right) = \frac{1}{2}\int_{\mathcal{X}\times\mathcal{X}}\left(\Phi_{ij}(x,u)-\Phi_{ji}(x,u)\right)\mathscr{K}_{(x,u)}d\rho_{\mathcal{X}}(x)d\rho_{\mathcal{X}}(u)$$

$$- \frac{1}{2}\int_{\mathcal{X}\times\mathcal{X}}\left(\Phi_{ij}(x,u)-\Phi_{ji}(x,u)\right)\mathscr{K}_{(u,x)}d\rho_{\mathcal{X}}(x)d\rho_{\mathcal{X}}(u)$$

$$= \frac{1}{2}\left(L_{\mathscr{K}}(\Phi_{ij}-\Phi_{ji}) - L_{\mathscr{K}}(\Phi_{ji}-\Phi_{ij})\right) = \lambda_i\lambda_j(\Phi_{ij}-\Phi_{ji}).$$

The proof of the properties of $\mathscr{K}^{\mathrm{SS}}$ is done in the same way. □

Proof of Proposition 2.3. Recall that

$$F(x,u) = \sum_{i,j=1}^{\infty} F_{ij}\Phi_{ij}(x,u) = \sum_{i,j=1}^{\infty} F_{ij}\phi_i(x)\phi_j(u).$$

Thus

$$\begin{aligned}\|F\|_\rho^2 &= \int_{\mathcal{X}\times\mathcal{X}} \left(\sum_{i,j=1}^{\infty} F_{ij}\phi_i(x)\phi_j(u)\right)^2 d\rho_{\mathcal{X}}(x)d\rho_{\mathcal{X}}(u) \\ &= \sum_{i,j,s,t=1}^{\infty} F_{ij}F_{st}\int_{\mathcal{X}} \phi_i(x)\phi_s(x)d\rho_{\mathcal{X}}(x)\int_{\mathcal{X}} \phi_j(u)\phi_t(u)d\rho_{\mathcal{X}}(u) \\ &= \sum_{i,j,s,t=1}^{\infty} F_{ij}F_{st}\left\langle\phi_i,\phi_s\right\rangle_{L^2_{\rho_{\mathcal{X}}}}\left\langle\phi_j,\phi_t\right\rangle_{L^2_{\rho_{\mathcal{X}}}}.\end{aligned}$$

By (2.15) and the orthogonality of $\{\phi_i\}_{i=1}^{\infty}$ and $\{\Phi_{ij}\}_{i,j=1}^{\infty}$, there holds

$$\begin{aligned}&\sum_{i,j,s,t=1}^{\infty} F_{ij}F_{st}\left\langle\phi_i,\phi_s\right\rangle_{L^2_{\rho_{\mathcal{X}}}}\left\langle\phi_j,\phi_t\right\rangle_{L^2_{\rho_{\mathcal{X}}}} \\ =&\sum_{i,j=1}^{\infty}\lambda_i\lambda_j F_{ij}^2\|\phi_i\|_K^2\|\phi_j\|_K^2 = \left\|\sum_{i,j=1}^{\infty}\sqrt{\lambda_i\lambda_j}F_{ij}\Phi_{ij}\right\|_{\mathscr{K}}^2 \\ =&\left\|\sum_{i,j=1}^{\infty}F_{ij}\left(L_{\mathscr{K}}^{1/2}\Phi_{ij}\right)\right\|_{\mathscr{K}}^2 = \left\|L_{\mathscr{K}}^{1/2}F\right\|_{\mathscr{K}}^2.\end{aligned}$$

□

2.6 Proofs of the bounds on $\mathcal{N}_{\mathscr{K}}(\lambda)$

In this section, we derive the upper and lower bounds of $\mathcal{N}_{\mathscr{K}}(\lambda)$ given in section 2.3.2.

Proof of Proposition 2.4.

Proof of the upper bound. Without loss of generality, we assume $D_2 = 1$ and $\lambda_i \le i^{-1/s_2}$. Otherwise, it can be scaled into λ without affecting the degree of λ. Since $\frac{t}{t+\lambda}$ is increasing with respect to t on $[0,+\infty)$, by letting $p_2 = 1/s_2 > 1$, we get

$$\mathcal{N}_{\mathscr{K}}(\lambda) = \sum_{i,j=1}^{\infty}\frac{\lambda_i\lambda_j}{\lambda_i\lambda_j+\lambda} \le 2\sum_{i=1}^{\infty}\frac{\lambda_i}{\lambda+\lambda_i} + \sum_{i=2}^{\infty}\sum_{j=2}^{\infty}\frac{1}{1+\lambda i^{p_2}j^{p_2}} =: 2\mathcal{N}_K(\lambda)+\mathcal{T}_U. \tag{2.61}$$

We bound $\mathcal{T}_U$ by the decomposition

$$\sum_{i=2}^{\infty}\sum_{j=2}^{\infty}\frac{1}{1+\lambda i^{p_2}j^{p_2}} \le \sum_{i=2}^{\infty}\int_1^{\infty}\frac{1}{1+(\lambda^{1/p_2}ix)^{p_2}}dx$$

$$=\sum_{i=2}^{\infty}\lambda^{-1/p_2}i^{-1}\int_{\lambda^{1/p_2}i}^{\infty}\frac{1}{1+t^{p_2}}dt \le \lambda^{-1/p_2}\int_1^{\infty}\frac{1}{x}\int_{\lambda^{1/p_2}x}^{\infty}\frac{dt}{1+t^{p_2}}dx$$

$$=\lambda^{-1/p_2}\left(\int_1^{\lambda^{-1/p_2}}\frac{1}{x}\int_{\lambda^{1/p_2}x}^{\infty}\frac{dt}{1+t^{p_2}}dx + \int_{\lambda^{-1/p_2}}^{\infty}\frac{1}{x}\int_{\lambda^{1/p_2}x}^{\infty}\frac{dt}{1+t^{p_2}}dx\right).$$

Note that

$$\int_1^{\lambda^{-1/p_2}}\frac{1}{x}\int_{\lambda^{1/p_2}x}^{\infty}\frac{dt}{1+t^{p_2}}dx \le \left(\int_0^{\infty}\frac{1}{1+t^{p_2}}dt\right)\int_1^{\lambda^{-1/p_2}}\frac{dx}{x} = C_{U,1}\log(\lambda^{-1/p_2})$$

with $C_{U,1}=\int_0^{\infty}\frac{1}{1+x^{p_2}}dx<\infty$, and that

$$\int_{\lambda^{-1/p_2}}^{\infty}\frac{1}{x}\int_{\lambda^{1/p_2}x}^{\infty}\frac{dt}{1+t^{p_2}}dx \le \int_{\lambda^{-1/p_2}}^{\infty}\frac{1}{x}\int_{\lambda^{1/p_2}x}^{\infty}\frac{dt}{t^{p_2}}dx = \frac{\lambda^{-1+1/p_2}}{(p_2-1)^2}\left(\lambda^{-1/p_2}\right)^{-p_2+1}.$$

We have

$$\mathcal{T}_U \le \lambda^{-1/p_2}\left[C_{U,1}\log(\lambda^{-1/p_2}) + \frac{\lambda^{-1+1/p_2}}{(p_2-1)^2}\left(\lambda^{-1/p_2}\right)^{-p_2+1}\right]$$

$$=\lambda^{-1/p_2}\left[C_{U,1}\log(\lambda^{-1/p_2}) + \frac{1}{(p_2-1)^2}\right].$$

Since $\lambda \le e^{-1}$ implies $\log(1/\lambda)\ge 1$, we obtain

$$\mathcal{T}_U = \sum_{i=2}^{\infty}\sum_{j=2}^{\infty}\frac{1}{1+\lambda i^{p_2}j^{p_2}} \le C_U\lambda^{-1/p_2}\log(1/\lambda), \tag{2.62}$$

with $C_U = \max\left\{C_{U,1}/p_2, \frac{1}{(p_2-1)^2}\right\}$. The proof of (2.26) is finished by substituting (2.62) into (2.61).

Proof of the lower bound. Since $\lambda_i \ge D_1 i^{-p_1}$ for $p_1 = 1/s_1$ and $D_1>0$, we get

$$\mathcal{N}_{\mathscr{K}}(\lambda) = \sum_{i,j=1}^{\infty}\frac{\lambda_i\lambda_j}{\lambda_i\lambda_j+\lambda} \ge \sum_{i,j=1}^{\infty}\frac{1}{1+(\lambda/D_1^2)i^{p_1}j^{p_1}}.$$

Since

$$\sum_{i,j=1}^{\infty} \frac{1}{1+(\lambda/D_1^2)i^{p_1}j^{p_1}} \geq \sum_{i=1}^{\infty} \int_1^{\infty} \frac{1}{1+((\lambda/D_1^2)^{1/p_1}ix)^{p_1}} dx$$

$$= \sum_{i=1}^{\infty} \lambda^{-1/p_1} i^{-1} \int_{(\lambda/D_1^2)^{1/p_1} i}^{\infty} \frac{1}{1+t^{p_1}} dt$$

$$\geq \lambda^{-1/p_1} \int_1^{\infty} \frac{1}{x} \int_{(\lambda/D_1^2)^{1/p_1} x}^{\infty} \frac{1}{1+t^{p_1}} dt dx,$$

there holds

$$\mathcal{N}_{\mathscr{K}}(\lambda) \geq \lambda^{-1/p_1} \int_1^{\infty} \frac{1}{x} \int_{(\lambda/D_1^2)^{1/p_1} x}^{\infty} \frac{1}{1+t^{p_1}} dt dx$$

$$\geq \lambda^{-1/p_1} \int_1^{\lambda^{-1/p_1}} \frac{1}{x} \int_{(\lambda/D_1^2)^{1/p_1} x}^{\infty} \frac{1}{1+t^{p_1}} dt dx$$

$$\geq \lambda^{-1/p_1} \int_1^{\lambda^{-1/p_1}} \frac{1}{x} \int_{D_1^{2/p_1}}^{\infty} \frac{1}{1+t^{p_1}} dt dx$$

$$=: C_L \lambda^{-1/p_1} \log(\lambda^{-1/p_1}).$$

Thus we obtain (2.27) with $D_0 = C_L/p_1$. □

Proof of Proposition 2.5. Since

$$\mathcal{N}_K(\lambda) = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \lambda} \leq \tilde{C}_0 \lambda^{-s}, for \lambda \leq \lambda_1,$$

it holds

$$\sum_{j=1}^{i} \frac{\lambda_j}{\lambda_j + \lambda_i} \leq \tilde{C}_0 \lambda_i^{-s}$$

by taking $\lambda = \lambda_i$. Note that $\frac{\lambda_j}{\lambda_j+\lambda_i} \geq 1/2$ for any $j \leq i$. Thus

$$\frac{i}{2} \leq \sum_{j=1}^{i} \frac{\lambda_j}{\lambda_j + \lambda_i} \leq \tilde{C}_0 \lambda_i^{-s}.$$

□

Proof of Proposition 2.6.

Proof of the upper bound. Without loss of generality, we assume $\hat{D}_2 = 1$. Then $\lambda_i \leq \exp(-t_2 i)$ and we have

$$\mathcal{N}_{\mathscr{K}}(\lambda) \leq \sum_{i,j=1}^{\infty} \frac{1}{1+\lambda \exp(t_2(i+j))} \leq \int_0^{\infty} \int_0^{\infty} \frac{1}{1+(\lambda \exp(t_2(x+y))} dydx. \quad (2.63)$$

Since $\lambda \leq e^{-1}$ and $\log(1/\lambda) \geq 1 > 0$, we decompose the integral in (2.63) as

$$\int_0^{\infty} \int_0^{\infty} \frac{1}{1+(\lambda \exp(t_2(x+y))} dydx = \int_0^{\infty} \int_{\log\lambda + t_2 x}^{\infty} \frac{1}{1+e^s} dsdx$$

$$= \int_0^{\frac{\log(1/\lambda)}{t_2}} \int_{\log\lambda+t_2x}^{\infty} \frac{1}{1+e^s} dsdx + \int_{\frac{\log(1/\lambda)}{t_2}}^{\infty} \int_{\log\lambda+t_2x}^{\infty} \frac{1}{1+e^s} dsdx.$$

Note that

$$\int_0^{\frac{\log(1/\lambda)}{t_2}} \int_{\log\lambda+t_2x}^{\infty} \frac{dsdx}{1+e^s} = \int_0^{\frac{\log(1/\lambda)}{t_2}} \int_{\log\lambda+t_2x}^{0} \frac{dsdx}{1+e^s} + \int_0^{\frac{\log(1/\lambda)}{t_2}} \int_0^{\infty} \frac{dsdx}{1+e^s}$$

$$\leq \int_0^{\frac{\log(1/\lambda)}{t_2}} \int_{\log\lambda+t_2x}^{0} dsdx + \int_0^{\frac{\log(1/\lambda)}{t_2}} \int_0^{\infty} \frac{dsdx}{e^s}$$

$$= \frac{\log^2 \frac{1}{\lambda}}{2t_2} + \frac{\log(1/\lambda)}{t_2} \leq \frac{3\log^2(1/\lambda)}{2t_2},$$

and

$$\int_{\frac{\log(1/\lambda)}{t_2}}^{\infty} \int_{\log\lambda+t_2x}^{\infty} \frac{1}{1+e^s} dsdx \leq \int_{\frac{\log(1/\lambda)}{t_2}}^{\infty} \int_{\log\lambda+t_2x}^{\infty} \frac{dsdx}{e^s} = \frac{1}{t_2}.$$

We obtain $\mathcal{N}_{\mathscr{K}}(\lambda) \leq \frac{5\log^2(1/\lambda)}{2t_2}$.

Proof of the lower bound. For $\lambda_i \geq \hat{D}_1 \exp(-t_1 i)$ and $\tilde{\lambda} = \lambda/\hat{D}_1^2$, there holds

$$\mathcal{N}_{\mathscr{K}}(\lambda) \geq \frac{1}{1+(\lambda/\hat{D}_1^2)\exp(t_1(i+j))} \geq \int_1^{\infty} \int_1^{\infty} \frac{1}{1+\tilde{\lambda}e^{t_1(x+y)}}.$$

Since $\lambda \leq \hat{D}_1^2 \exp(-8t_1) \leq \hat{D}_1^2 \exp(-2t_1)$ and $\frac{1}{t_1}\log(1/\tilde{\lambda}) - 1 \geq 1$, we obtain

$$\begin{aligned}\int_1^\infty \int_1^\infty \frac{1}{1+\tilde{\lambda}e^{t_1(x+y)}} &\geq \int_1^{\frac{1}{t_1}\log(1/\tilde{\lambda})-1} \int_{\log\tilde{\lambda}+t_1(x+1)}^0 \frac{1}{1+e^s} ds dx \\ &\geq \int_1^{\frac{1}{t_1}\log(1/\tilde{\lambda})-1} \int_{\log\tilde{\lambda}+t_1(x+1)}^0 \frac{1}{2} ds dx \\ &= \frac{1}{2}\left(\frac{1}{2t_1}\log^2\left(\frac{1}{\tilde{\lambda}}\right) - 2\log\left(\frac{1}{\tilde{\lambda}}\right) + 2t_1\right).\end{aligned}$$

Note that $\frac{1}{4t_1}\log^2(1/\tilde{\lambda}) \geq 2\log(1/\tilde{\lambda})$ for $\lambda \leq \hat{D_1}^2 \exp(-8t_1)$. There holds

$$\mathcal{N}_{\mathscr{K}}(\lambda) \geq \frac{1}{8t_1}\log^2(1/\tilde{\lambda}).$$

Since $\frac{1}{2}\log(1/\lambda) + \log(\hat{D}_1^2) \geq 0$ for $\lambda \leq \hat{D}_1^4$, we obtain

$$\frac{1}{8t_1}\log^2(1/\tilde{\lambda}) \geq \frac{1}{32t_1}\log^2\frac{1}{\lambda}.$$

□

2.7 Proofs of Technical Lemmas

Lemma 2.2 and Lemma 2.3 are proved in this section.

2.7.1 Proof of Lemma 2.2

The proof of Lemma 2.2 is based on the following lemmas. The first one is Pinelis' concentration inequality [59].

Lemma 2.4. *For a random variable ξ on $(\mathcal{Z}, \rho)$ with values in a separable Hilbert space $(H, \|\cdot\|)$ satisfying $\|\xi\| \leq M < \infty$ almost surely, and a random sample $\{z_i\}_{i=1}^s$ independently drawn according to ρ, there holds with confidence $1-\delta$,*

$$\left\|\frac{1}{s}\sum_{i=1}^s [\xi(z_i) - \mathbb{E}\xi]\right\| \leq \frac{2M\log(2/\delta)}{s} + \sqrt{\frac{2\mathbb{E}(\|\xi\|^2)\log(2/\delta)}{s}}. \tag{2.64}$$

The next Lemma is about the application of $\mathcal{N}_{\mathscr{K}}(\lambda)$.

Lemma 2.5. *Let $\{z_i' = (x_i, u_i, y_i')\}_{i=1}^n$ be a sequence of i.i.d. copies of $z' = (x, u, y') \in (\mathcal{X}^2 \times \mathcal{Y}, \rho)$. Then, the following statements hold true.*

(a) $\mathbb{E}\left\|(L_{\mathscr{K}} + \lambda I)^{-1/2}\, \mathscr{K}_{(x,u)}\right\|_{\mathscr{K}}^2 = \mathcal{N}_{\mathscr{K}}(\lambda)$.

(b) For any bounded measurable real-valued function g and $\xi_g(z') = g(z')\mathscr{K}_{(x,u)}$, with confidence at least $1-\delta$, it holds

$$\left\|(L_{\mathscr{K}} + \lambda I)^{-1/2}\left(\frac{1}{n}\sum_{i=1}^n \xi_g(z_i') - \mathbb{E}[\xi_g]\right)\right\|_{\mathscr{K}} \leq 2\|g\|_\infty(\kappa^2+1)\mathcal{A}_{n,\lambda}\log(2/\delta). \tag{2.65}$$

Proof. Note that

$$\mathscr{K}_{(x,u)} = \sum_{i,j=1}^\infty \Phi_{ij}(x,u)\Phi_{ij},$$

and

$$(L_{\mathscr{K}} + \lambda I)^{-1/2}\, \mathscr{K}_{(x,u)} = \sum_{i,j=1}^\infty \frac{\Phi_{ij}(x,u)}{(\lambda_i\lambda_j + \lambda)^{1/2}}\Phi_{ij}.$$

Then

$$\mathbb{E}\left\|(L_{\mathscr{K}} + \lambda I)^{-1/2}\, \mathscr{K}_{(x,u)}\right\|_{\mathscr{K}}^2 = \sum_{i,j=1}^\infty \frac{\int_{\mathcal{X}\times\mathcal{X}} \Phi_{ij}(x,u)^2 d\rho_{\mathcal{X}}(x) d\rho_{\mathcal{X}}(u)}{\lambda_i\lambda_j + \lambda}$$

$$= \sum_{i,j=1}^\infty \frac{\|\Phi_{ij}\|_\rho^2}{\lambda_i\lambda_j + \lambda} = \sum_{i,j=1}^\infty \frac{\left\|L_{\mathscr{K}}^{1/2}\Phi_{ij}\right\|_{\mathscr{K}}^2}{\lambda_i\lambda_j + \lambda} = \sum_{i,j=1}^\infty \frac{\lambda_i\lambda_j}{\lambda_i\lambda_j + \lambda} = \mathcal{N}_{\mathscr{K}}(\lambda).$$

The proof of (2.65) is a direct application of Lemma 2.4 by noting that

$$\|\,(L_{\mathscr{K}} + \lambda I)^{-1/2}\,\xi_g\|_{\mathscr{K}} \leq \|g\|_\infty \left\|(L_{\mathscr{K}} + \lambda I)^{-1/2}\right\|_{\mathrm{op}} \|\mathscr{K}_{(x,u)}\|_{\mathscr{K}} \leq \frac{\|g\|_\infty\kappa^2}{\sqrt{\lambda}},$$

$$\mathbb{E}\|\,(L_{\mathscr{K}} + \lambda I)^{-1/2}\,\xi_g\|_{\mathscr{K}}^2 \leq \|g\|_\infty^2 \mathbb{E}\left\|(L_{\mathscr{K}} + \lambda I)^{-1/2}\, \mathscr{K}_{(x,u)}\right\|_{\mathscr{K}}^2 = \|g\|_\infty^2 \mathcal{N}_{\mathscr{K}}(\lambda).$$

□

Based on the lemmas above, we derive the probabilistic bounds for quantities defined above in Lemma 2.2.

Lemma 2.6. *Assume that $|y'| \leq M$. Then for $0 < \delta < 1$, each of the following bound holds with confidence $1 - \delta$.*

$$\mathcal{S}_{\mathbf{z}',\lambda} \leq C_{\mathcal{S}'} \mathcal{A}_{n,\lambda} \log(2/\delta), \tag{2.66}$$

$$\mathcal{P}_{\mathbf{z}',\lambda} \leq C_{\mathcal{P}'} \mathcal{A}_{n,\lambda} \log(2/\delta), \tag{2.67}$$

$$\mathcal{Q}_{\mathbf{z}',\lambda} \leq C_{\mathcal{Q}'} \left(\frac{\mathcal{A}_{n,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{n,\lambda}^2}{\lambda} \right) \log^2(2/\delta) + 1, \tag{2.68}$$

with $C_{\mathcal{S}'} = 2M(\kappa^2 + 1), C_{\mathcal{P}'} = 2(\kappa^4 + \kappa^2)$ and $C_{\mathcal{Q}'} = C_{\mathcal{P}'} + C_{\mathcal{P}'}^2$.

Proof. We first bound $\mathcal{S}_{\mathbf{z}',\lambda}$. Note that

$$\mathbb{E} y' \mathscr{K}_{(x,u)} = \int_{\mathcal{X}^2} \int_{\mathcal{Y}} y' d\rho(y'|x,u) d\rho_{\mathcal{X}}(x) d\rho_{\mathcal{X}}(u) = \int_{\mathcal{X}^2} F_\rho(x,u) d\rho_{\mathcal{X}}(x) d\rho_{\mathcal{X}}(u) = L_{\mathscr{K}} F_\rho$$

and

$$(L_{\mathscr{K}} + \lambda I)^{-1/2} \left(\frac{1}{n} S_{\mathbf{x}'}^T Y' - L_{\mathscr{K}} F_\rho \right) = (L_{\mathscr{K}} + \lambda I)^{-1/2} \frac{1}{n} \sum_{i=1}^n (y_i' \mathscr{K}_{(x_i,u_i)} - L_{\mathscr{K}}).$$

Apply (2.65) to $\mathcal{S}_{\mathbf{z}',\lambda}$ with $g(z') = y'$, and we obtain (2.66) as a result of $\|g\|_\infty = \sup_{y' \in \mathcal{Y}} |y'| \leq M$.

Now we bound $\mathcal{P}_{\mathbf{z}',\lambda}$. Apply Lemma 2.5 to $\xi_F(z') = F(x,u)\mathscr{K}_{(x,u)}$ for any $F \in \mathcal{H}_{\mathscr{K}}$, and there holds

$$\left\| (L_{\mathscr{K}} + \lambda I)^{-1/2} \left(L_{\mathscr{K}}^{\mathbf{x}'} F - L_{\mathscr{K}} F \right) \right\|_{\mathscr{K}} = \left\| (L_{\mathscr{K}} + \lambda I)^{-1/2} \left(\frac{1}{n} \sum_{i=1}^n \xi_F(z_i') - \mathbb{E}\xi_F(z') \right) \right\|_{\mathscr{K}}$$

$$\leq 2 \|F\|_\infty (\kappa^2 + 1) \mathcal{A}_{n,\lambda} \log(2/\delta) \leq 2\kappa^2 \|F\|_{\mathscr{K}} (\kappa^2 + 1) \mathcal{A}_{n,\lambda} \log(2/\delta),$$

where the last inequality follows $\|F\|_\infty \leq \kappa^2 \|F\|_{\mathscr{K}}$. This implies (2.67) according to the definition of the operator norm.

To derive the bound of $\mathcal{Q}_{\mathbf{z}',\lambda}$, we use the following decomposition

$$\mathcal{Q}_{\mathbf{z}',\lambda} \leq \left\| (L_{\mathscr{K}} + \lambda I)^{1/2} \left[\left(L_{\mathscr{K}}^{\mathbf{x}'} + \lambda I \right)^{-1} - (L_{\mathscr{K}} + \lambda I)^{-1} \right] (L_{\mathscr{K}} + \lambda I)^{1/2} \right\|_{\mathrm{op}}$$

$$+ \left\| (L_{\mathscr{K}} + \lambda I)^{1/2} (L_{\mathscr{K}} + \lambda I)^{-1} (L_{\mathscr{K}} + \lambda I)^{1/2} \right\|_{\mathrm{op}} =: \mathcal{Q}_{\mathbf{z}',0} + 1. \tag{2.69}$$

Note that

$$\left\| \left(L_{\mathscr{K}}^{\mathbf{x}'} + \lambda I \right)^{-1} \left(L_{\mathscr{K}} - L_{\mathscr{K}}^{\mathbf{x}'} \right) \right\|_{\mathrm{op}} = \left\| \left(L_{\mathscr{K}} - L_{\mathscr{K}}^{\mathbf{x}'} \right) \left(L_{\mathscr{K}}^{\mathbf{x}'} + \lambda I \right)^{-1} \right\|_{\mathrm{op}}$$

as a result of the symmetry of $\left(L_{\mathscr{K}}^{\mathbf{x}'} + \lambda I \right)^{-1}$ and $L_{\mathscr{K}} - L_{\mathscr{K}}^{\mathbf{x}'}$. By applying the second order decomposition (2.42) to $\left(L_{\mathscr{K}}^{\mathbf{x}'} + \lambda I \right)^{-1} - (L_{\mathscr{K}} + \lambda I)^{-1}$, we have

$$\mathcal{Q}_{\mathbf{z}',0} \leq \left\| (L_{\mathscr{K}} + \lambda I)^{-1/2} \left(L_{\mathscr{K}} - L_{\mathscr{K}}^{\mathbf{x}'} \right) (L_{\mathscr{K}} + \lambda I)^{-1/2} \right\|_{\mathrm{op}}$$

$$+ \left\| (L_{\mathscr{K}} + \lambda I)^{-1/2} \left(L_{\mathscr{K}} - L_{\mathscr{K}}^{\mathbf{x}'} \right) \left(L_{\mathscr{K}}^{\mathbf{x}'} + \lambda I \right)^{-1} \left(L_{\mathscr{K}} - L_{\mathscr{K}}^{\mathbf{x}'} \right) (L_{\mathscr{K}} + \lambda I)^{-1/2} \right\|_{\mathrm{op}}$$

$$\leq \frac{\mathcal{P}_{\mathbf{z},\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{P}_{\mathbf{z},\lambda}^2}{\lambda}. \tag{2.70}$$

We obtain (2.68) by substituting (2.67) and (2.70) into (2.69). □

We use the following Lemma to derive expected error bounds from the probabilistic error bounds.

Lemma 2.7. *Let ξ be a positive random variable. If there are constants $a > 0, b > 0, \tau > 0$ such that for any $0 < \delta \leq 1$, with confidence at least $1 - \delta$, there holds $\xi \leq a(\log \frac{b}{\delta})^{\tau}$, then for any $\theta > 0$ we have $\mathbb{E}[\xi^{\theta}] \leq a^{\theta} b \Gamma(\tau\theta + 1)$.*

Lemma 2.7 is a standard result, of which the proof can be found, e.g., in [39].

Proof of Lemma 2.2. The proofs of (2.43) and (2.44) are direct applications of Lemma 2.7 to (2.66) and (2.67), respectively, with $C_{\theta,\mathcal{S}'} = 2C_{\mathcal{S}'}^{\theta}\Gamma(\theta+1), C_{\theta,\mathcal{P}'} = 2C_{\mathcal{P}'}^{\theta}\Gamma(\theta+1)$.

By noting that $\log(2/\delta) > 1$, (2.68) becomes

$$\mathcal{Q}_{\mathbf{z}',\lambda} \leq C_{\mathcal{Q}'}\left(\frac{\mathcal{A}_{n,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{n,\lambda}^2}{\lambda}\right)\log^2(2/\delta) + 1 \leq \max\{C_{\mathcal{Q}'}, 1\}\left(\frac{\mathcal{A}_{n,\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{A}_{n,\lambda}^2}{\lambda} + 1\right)\log^2\frac{2}{\delta}.$$

(2.45) follows Lemma 2.7 with $C_{\theta,\mathcal{Q}'} = 2\left(\max\{C'_{\mathcal{Q}}, 1\}\right)^{\theta}\Gamma(2\theta+1)$. □

2.7.2 Proof of Lemma 2.3

Average of sums of i.i.d. blocks. Consider a sequence of real numbers $\{a_{ij}\}_{i,j=1}^{N}$. We have

$$(N-2)!\,\lfloor N/2\rfloor \sum_{i\neq j} a_{ij} = \sum_{\pi\in\Pi_N}\sum_{i=1}^{\lfloor N/2\rfloor} a_{\pi(i),\pi(\lfloor N/2\rfloor+i)}, \tag{2.71}$$

where Π_N is the set of all the permutations of $\{1, 2, ..., N\}$. In fact, it is easy to see that each a_{ij} with $i \neq j$, has been added $(N-2)!\lfloor N/2\rfloor$ times on the right-hand side of (2.71). Equation (2.71) is widely used in U-statistics and is known as the *average of sums of i.i.d. blocks* technique [23] since it can simplify the analysis of dependent random variables to the independent case as shown in the following lemma.

Lemma 2.8 is a variant of (Lemma A.1, [23]) for random variables taking values in a Hilbert space.

Lemma 2.8. *Consider a sequence of random variables $\{\xi_{ij}\}_{i,j=1}^{N}$ taking values in a Hilbert space $(H, \|\cdot\|)$, such that for any distinct positive integers i, j, s, t, ξ_{ij} and ξ_{st} are independent and identically distributed. Then for any convex nondecreasing function $\psi : \mathbb{R} \to \mathbb{R}$, there holds*

$$\mathbb{E}\psi\left(\left\|\frac{1}{N(N-1)}\sum_{i\neq j}\xi_{ij}\right\|\right) \leq \mathbb{E}\psi\left(\left\|\frac{1}{\lfloor N/2\rfloor}\sum_{i=1}^{\lfloor N/2\rfloor}\xi_{i,\lfloor N/2\rfloor+i}\right\|\right). \tag{2.72}$$

Proof. It is straightforward to generalize (2.71) to sequences in a Hilbert space by doing inner product with each element in that space due to the linearity of inner

products. Thus, we have

$$\frac{1}{N(N-1)}\sum_{i\neq j}\xi_{ij}=\frac{1}{N!}\sum_{\pi\in\Pi_N}\frac{1}{\lfloor N/2\rfloor}\sum_{i=1}^{\lfloor N/2\rfloor}\xi_{\pi(i),\pi(\lfloor N/2\rfloor+i)}.$$

Then it holds

$$\begin{aligned}\left\|\frac{1}{N(N-1)}\sum_{i\neq j}\xi_{ij}\right\| &= \left\|\frac{1}{N!}\sum_{\pi\in\Pi_N}\frac{1}{\lfloor N/2\rfloor}\sum_{i=1}^{\lfloor N/2\rfloor}\xi_{\pi(i),\pi(\lfloor N/2\rfloor+i)}\right\| \\ &\leq \frac{1}{N!}\sum_{\pi\in\Pi_N}\left\|\frac{1}{\lfloor N/2\rfloor}\sum_{i=1}^{\lfloor N/2\rfloor}\xi_{\pi(i),\pi(\lfloor N/2\rfloor+i)}\right\|.\end{aligned}$$

Since ψ is convex and nondecreasing, (2.72) is proved by Jensen's inequality and the i.i.d. assumption on ξ_{ij},ξ_{st} for distinct positive integers i,j,s,t . □

Proof of Lemma 2.3. We first bound $\mathcal{S}_{\mathbf{z},\lambda}$. Note that

$$\frac{1}{N^2}S_{\mathbf{x}}^T Y=\frac{1}{N^2}\sum_{i,j=1}^{N}y_{ij}\mathscr{K}_{(x_i,x_j)}=\frac{1}{N^2}\sum_{i=1}^{N}y_{ii}\mathscr{K}_{(x_i,x_i)}+\frac{1}{N^2}\sum_{i\neq j}y_{ij}\mathscr{K}_{(x_i,x_j)}.$$

Thus

$$\begin{aligned}\mathcal{S}_{\mathbf{z},\lambda} &\leq \left\|(L_{\mathscr{K}}+\lambda I)^{-1/2}\frac{1}{N^2}\sum_{i=1}^{N}\left(y_{ii}\mathscr{K}_{(x_i,x_i)}-L_{\mathscr{K}}F_\rho\right)\right\|_{\mathscr{K}} \\ &\quad+\left\|(L_{\mathscr{K}}+\lambda I)^{-1/2}\frac{1}{N(N-1)}\sum_{i\neq j}\left(y_{ij}\mathscr{K}_{(x_i,x_j)}-L_{\mathscr{K}}F_\rho\right)\right\|_{\mathscr{K}} \\ &=: \mathcal{S}_{\lambda,1}+\mathcal{S}_{\lambda,2},\end{aligned}\tag{2.73}$$

as a result of $1/N^2\leq 1/(N(N-1))$. Since $|y_{ii}|\leq M$, $\left\|\mathscr{K}_{(x,u)}\right\|_{\mathscr{K}}\leq\kappa^2$ and $\|L_{\mathscr{K}}\|_{\mathrm{op}}\leq\kappa^4$, we obtain

$$\mathcal{S}_{\lambda,1}\leq\frac{M\kappa^2+\kappa^4\|F_\rho\|_{\mathscr{K}}}{N\sqrt{\lambda}}\leq\left(M\kappa^2+\kappa^4\|F_\rho\|_{\mathscr{K}}\right)\mathcal{A}_{N,\lambda}.\tag{2.74}$$

For $\mathcal{S}_{\lambda,2}$, by Lemma 2.8 and (2.43), there holds

$$\mathbb{E}\mathcal{S}_{\lambda,2}^2 \leq \mathbb{E}\left\|(L_{\mathscr{K}}+\lambda I)^{-1/2}\frac{1}{n}\sum_{i=1}^{n}\left(y_i'\mathscr{K}_{(x_i,u_i)}-L_{\mathscr{K}}F_\rho\right)\right\|_{\mathscr{K}}^2 = \mathbb{E}\mathcal{S}_{\mathbf{z}',\lambda}^2 \leq C_{2,\mathcal{S}'}\mathcal{A}_{n,\lambda}^2, \tag{2.75}$$

thanks to the monotonicity and convexity of x^2 for $x>0$, where $n=\lfloor N/2\rfloor$, $y_i'=y_{i,n+i}$ and $u_i=x_{n+i}$. Note that for $n=\lfloor N/2\rfloor$ and $N\geq 4$,

$$\mathcal{A}_{n,\lambda}\leq 4\mathcal{A}_{N,\lambda}.$$

Thus (2.75) becomes

$$\mathbb{E}\mathcal{S}_{\lambda,2}^2 \leq 16C_{2,\mathcal{S}'}\mathcal{A}_{N,\lambda}^2. \tag{2.76}$$

By choosing $C_{\mathcal{S}}=2(16C_{2,\mathcal{S}'}+(M\kappa^2+\kappa^4\|F_\rho\|_{\mathscr{K}})^2)$, according to (2.74), (2.76) and (2.73), we obtain

$$\mathbb{E}\mathcal{S}_{\mathbf{z},\lambda}^2 \leq 2\left(\mathbb{E}\mathcal{S}_{\lambda,1}^2+\mathbb{E}\mathcal{S}_{\lambda,1}^2\right)\leq C_{\mathcal{S}}\mathcal{A}_{N,\lambda}^2.$$

Now we bound $\mathcal{P}_{\mathbf{z},\lambda}$. Consider the decomposition

$$\begin{aligned}\mathcal{P}_{\mathbf{z},\lambda} \leq & \left\|(L_{\mathscr{K}}+\lambda I)^{-1/2}\frac{1}{N^2}\sum_{i=1}^{N}\left(\left\langle\cdot,\mathscr{K}_{(x_i,x_i)}\right\rangle_{\mathscr{K}}\mathscr{K}_{(x_i,x_i)}-L_{\mathscr{K}}\right)\right\|_{\mathrm{op}} \\ & + \left\|(L_{\mathscr{K}}+\lambda I)^{-1/2}\frac{1}{N(N-1)}\sum_{i\neq j}\left(\left\langle\cdot,\mathscr{K}_{(x_i,x_j)}\right\rangle_{\mathscr{K}}\mathscr{K}_{(x_i,x_j)}-L_{\mathscr{K}}\right)\right\|_{\mathrm{op}} \\ & =: \mathcal{P}_{\lambda,1}+\mathcal{P}_{\lambda,2.}\end{aligned} \tag{2.77}$$

For $\mathcal{P}_{\lambda,1}$, similarly to (2.13), we have

$$\left\|\left\langle\cdot,\mathscr{K}_{(x_i,x_i)}\right\rangle_{\mathscr{K}}\mathscr{K}_{(x_i,x_i)}\right\|_{\mathrm{op}}\leq\kappa^4,$$

and

$$\mathcal{P}_{\lambda,1}\leq\frac{2\kappa^4}{N\sqrt{\lambda}}. \tag{2.78}$$

To bound $\mathcal{P}_{\lambda,2}$, by applying (2.72) again, there holds

$$\mathbb{E}\mathcal{P}_{\lambda,2}^2 \le \mathbb{E}\mathcal{P}_{\mathbf{z}',\lambda}^2 \le C_{2,\mathcal{P}'}\mathcal{A}_{n,\lambda}^2 \le 16C_{2,\mathcal{P}'}\mathcal{A}_{N,\lambda}^2. \tag{2.79}$$

By (2.77), (2.78) and (2.79), we get

$$\mathbb{E}\mathcal{P}_{\mathbf{z},\lambda}^2 \le 2(\mathbb{E}\mathcal{P}_{\lambda,1}^2 + \mathcal{P}_{\lambda,2}^2) \le C_{\mathcal{P}}\mathcal{A}_{N,\lambda}^2 \tag{2.80}$$

with $C_{\mathcal{P}} = 32C_{2,\mathcal{P}'} + 8\kappa^8$.

To bound $\mathcal{Q}_{\mathbf{z},\lambda}$, we do the decomposition

$$\begin{aligned}\mathcal{Q}_{\mathbf{z},\lambda} \le& \left\|(L_{\mathscr{K}} + \lambda I)^{1/2}\left[(L_{\mathscr{K}}^{\mathbf{x}} + \lambda I)^{-1} - (L_{\mathscr{K}} + \lambda I)^{-1}\right](L_{\mathscr{K}} + \lambda I)^{1/2}\right\|_{\mathrm{op}} \\ &+ \left\|(L_{\mathscr{K}} + \lambda I)^{1/2}(L_{\mathscr{K}} + \lambda I)^{-1}(L_{\mathscr{K}} + \lambda I)^{1/2}\right\|_{\mathrm{op}} =: \mathcal{Q}_{\mathbf{z},0} + 1. \end{aligned}\tag{2.81}$$

Use the second order decomposition (2.42) and we obtain

$$\begin{aligned}\mathcal{Q}_{\mathbf{z},0} \le& \left\|(L_{\mathscr{K}} + \lambda I)^{-1/2}(L_{\mathscr{K}} - L_{\mathscr{K}}^{\mathbf{x}})(L_{\mathscr{K}} + \lambda I)^{-1/2}\right\|_{\mathrm{op}} \\ &+ \left\|(L_{\mathscr{K}} + \lambda I)^{-1/2}(L_{\mathscr{K}} - L_{\mathscr{K}}^{\mathbf{x}})(L_{\mathscr{K}}^{\mathbf{x}} + \lambda I)^{-1}(L_{\mathscr{K}} - L_{\mathscr{K}}^{\mathbf{x}})(L_{\mathscr{K}} + \lambda I)^{-1/2}\right\|_{\mathrm{op}} \\ \le& \frac{\mathcal{P}_{\mathbf{z},\lambda}}{\sqrt{\lambda}} + \frac{\mathcal{P}_{\mathbf{z},\lambda}^2}{\lambda}. \end{aligned}\tag{2.82}$$

By (2.81) and (2.82), we have

$$\mathbb{E}\mathcal{Q}_{\mathbf{z},\lambda}^2 \le 2\mathbb{E}\mathcal{Q}_{\mathbf{z},0}^2 + 2 \le 4\left(\frac{\mathbb{E}\mathcal{P}_{\mathbf{z},\lambda}^2}{\lambda} + \frac{\mathbb{E}\mathcal{P}_{\mathbf{z},\lambda}^4}{\lambda^2}\right) + 2. \tag{2.83}$$

By Lemma 2.8, (2.77) and (2.78),

$$\begin{aligned}\mathbb{E}\mathcal{P}_{\mathbf{z},\lambda}^4 \le& 8\left(\mathbb{E}\mathcal{P}_{\lambda,1}^4 + \mathbb{E}\mathcal{P}_{\lambda,2}^4\right) \le \frac{128\kappa^{16}}{N^4\lambda^2} + 8\mathbb{E}\mathcal{P}_{\mathbf{z}',\lambda}^4 \\ \le& \left(128\kappa^{16} + 4^4 \times 8C_{4,\mathcal{P}'}\right)\mathcal{A}_{N,\lambda}^4 =: C_{4,\mathcal{P}}\mathcal{A}_{N,\lambda}^4. \end{aligned}\tag{2.84}$$

Substitute (2.47) and (2.84) into (2.83), and we obtain (2.46) with

$$C_{\mathcal{Q}} = 4\max\{C_{\mathcal{P}}, C_{4,\mathcal{P}}\}.$$

. □

Chapter 3

Sparse Semi-supervised Learning with Summary Statistics

Many kernel-based machine learning algorithms need the availability of input data during the training process while the data may be unavailable in many circumstances due to privacy issues. However, there are usually unlabeled data published without privacy issues from the same distribution as the input of the private data [90]. The learning process with both labeled (a.k.a. supervised) and ublabeled (a.k.a. unsupervised) data, is often called semi-supervised learning [11, 18, 36]. Based on the unlabeled data and the summary statistics (a statistic generated with the labeled data with the hope to reduce the leak of sensitive data), a novel algorithm for linear models has been proposed [50, 90] and been extended to the kernel-based learning scheme with empirical features [61]. In this chapter, we propose a semi-supervised learning algorithm that achieves both sparsity and approximation accuracy based on the summary statistics and empirical features.

3.1 Summary Statistics

Let $\mathcal{X}$ be an input space and $\mathcal{Y} = \mathbb{R}$ be the output space. Here $\mathcal{X} \times \mathcal{Y}$ is equipped with a Borel probability measure ρ that can be decomposed as a marginal measure $\rho_{\mathcal{X}}$ on $\mathcal{X}$ and a conditional measure $\rho(\cdot|x)$ at $x \in \mathcal{X}$ on $\mathcal{Y}$. Recall the target function

f_ρ to be recovered in least-squares regression

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x), x \in \mathcal{X}.$$

Given a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, let $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$ be the corresponding reproducing kernel Hilbert space (RKHS). Assume that $\kappa^2 := \sup_{x \in \mathcal{X}} K(x, x) < +\infty$ and $|y| \leq M < +\infty$. Recall the integral operator defined by (1.2) and the notation $K_x(u) = K(x, u), x, u \in \mathcal{X}$. Let $\{(\lambda_i, \phi_i)\}_{i=1}^{\infty}$ be the eigensystem of L_K orthonormalized in $\mathcal{H}_K$. Based on the input observations $\mathbf{x} = \{x_i\}_{i=1}^m$, we can define the empirical integral operator with respect to $\mathbf{x}$ as

$$L_K^{\mathbf{x}} : \mathcal{H}_K \to \mathcal{H}_K,$$

$$f \mapsto \frac{1}{m} \sum_{i=1}^{m} f(x_i) K_{x_i}.$$

Since $L_K^{\mathbf{x}}$ is positive semi-definite with rank at most m [38], we can write its eigensystem orthonormalized in $\mathcal{H}_K$ as $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}_{i=1}^{\infty}$ with $\lambda_1^{\mathbf{x}} \geq \lambda_2^{\mathbf{x}} \geq \cdots \geq \lambda_m^{\mathbf{x}} \geq 0 = \lambda_{m+1}^{\mathbf{x}} = \cdots$. Note that $L_K^{\mathbf{x}}$ depends only on $\mathbf{x}$. With unlabeled data $\mathbf{u} = \{u_i\}_{i=1}^n \subset \mathcal{X}$, we can define similarly $L_K^{\mathbf{u}}$ the empirical integral operator with respect to $\mathbf{u}$ and write the corresponding eigensystem $\{(\lambda_i^{\mathbf{u}}, \phi_i^{\mathbf{u}})\}_{i=1}^{\infty}$ with $\lambda_1^{\mathbf{u}} \geq \lambda_2^{\mathbf{u}} \geq \cdots \geq \lambda_n^{\mathbf{u}} \geq 0 = \lambda_{n+1}^{\mathbf{u}} = \cdots$.

Recently, a novel estimator was introduced in linear regression for privacy consideration. Consider a linear model $Y = \mathbf{X}\beta + \epsilon$ and its least squares estimator $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T Y$. Since the access to the coefficient matrix $\mathbf{X}$ may be impossible due to privacy issues, [50, 90] define a new estimator $\hat{\beta}' = \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1} \mathbf{X}^T Y$. Here $\tilde{X}$ is from openly accessible and unlabeled data. To obtain $\hat{\beta}'$, one need only the summary statistics $\mathbf{X}^T Y$ and the covariance matrix $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$. Thus, the direct exposure of private information in $\mathbf{X}$ is avoided. [61] generalized the summary statistics to the non-parametric case. Based on the empirical features, the summary statistic is

defined by $\mathbf{d}^{\mathbf{u},\mathbf{z}} = (d_1^{\mathbf{u},\mathbf{z}}, \cdots, d_n^{\mathbf{u},\mathbf{z}})$ with

$$d_i^{\mathbf{u},\mathbf{z}} = \left\langle \phi_i^{\mathbf{u}}, \frac{1}{m}\sum_{j=1}^{m} y_j K_{x_j} \right\rangle_K, 1 \leq i \leq n. \tag{3.1}$$

3.2 Algorithm

In [38], an empirical feature-based learning algorithm based on ℓ_1 regularization was proposed that produce output function with both sparsity and approximation accuracy. In particular, the output function is

$$f_\gamma^{\mathbf{z}} = \sum_{i=1}^{m} c_{\gamma,i}^{\mathbf{z}} \phi_i^{\mathbf{x}}$$

with $\mathbf{c}_\gamma^{\mathbf{z}} = \left(c_{\gamma,1}^{\mathbf{z}}, \cdots, c_{\gamma,m}^{\mathbf{z}}\right)^T \in \mathbb{R}^m$ defined by

$$\mathbf{c}_\gamma^{\mathbf{z}} = \arg\min_{\mathbf{c}\in\mathbb{R}^m} \left\{ \frac{1}{m}\sum_{i=1}^{m} \left(\left(\sum_{j=1}^{m} c_j \phi_j^{\mathbf{x}} \right)(x_i) - y_i \right)^2 + \gamma \|\mathbf{c}\|_1 \right\}.$$

Here $\|\cdot\|_1$ is the ℓ_1 norm of a vector, i.e., $\|\mathbf{c}\|_1 = \sum_{i=1}^m |c_i|$ for any $\mathbf{c} = (c_1, \cdots, c_m)^T \in \mathbb{R}^m$. [38] reveals that $c_\gamma^{\mathbf{z}}$ has a closed form representation as

$$c_{\gamma,i}^{\mathbf{z}} = \begin{cases} 0, & \text{if } 2|d_i^{\mathbf{z}}| \leq \gamma \text{ or } \lambda_i^{\mathbf{x}} = 0, \\ \frac{1}{\lambda_i^{\mathbf{x}}}(d_i^{\mathbf{z}} - \gamma/2) & \text{if } d_i^{\mathbf{z}} > \gamma/2 \text{ and } \lambda_i^{\mathbf{x}} > 0, \\ \frac{1}{\lambda_i^{\mathbf{x}}}(d_i^{\mathbf{z}} + \gamma/2) & \text{if } d_i^{\mathbf{z}} < -\gamma/2 \text{ and } \lambda_i^{\mathbf{x}} > 0, \end{cases}$$

with

$$d_i^{\mathbf{z}} = \frac{1}{m}\sum_{j=1}^{m} y_j \phi_i^{\mathbf{x}}(x_j) = \left\langle \phi_i^{\mathbf{x}}, \frac{1}{m}\sum_{j=1}^{m} y_j K_{x_j} \right\rangle_K, 1 \leq i \leq m.$$

Note that $d_i^{\mathbf{u},\mathbf{z}}$ defined in (3.1) can be obtained by replacing $\phi_i^{\mathbf{x}}$ with $\phi_i^{\mathbf{u}}$ in the definition of $d_i^{\mathbf{x}}$. According to the idea of summary statistics and ℓ_1 regularized empirical feature-based learning, we propose a new learning algorithm as

$$f_\gamma^{\mathbf{u},\mathbf{z}} = \sum_{i=1}^{n} c_{\gamma,i}^{\mathbf{u},\mathbf{z}} \phi_i^{\mathbf{u}}$$

with $\mathbf{c}_{\gamma}^{\mathbf{u},\mathbf{z}} = \left(\mathbf{c}_{\gamma,1}^{\mathbf{u},\mathbf{z}}, \cdots, \mathbf{c}_{\gamma,m}^{\mathbf{u},\mathbf{z}}\right)$ defined by

$$c_{\gamma,i}^{\mathbf{u},\mathbf{z}} = \begin{cases} 0, & \text{if } 2\left|d_i^{\mathbf{u},\mathbf{z}}\right| \leq \gamma \text{ or } \lambda_i^{\mathbf{u}} = 0, \\ \frac{1}{\lambda_i^{\mathbf{u}}}\left(d_i^{\mathbf{u},\mathbf{z}} - \gamma/2\right), & \text{if } d_i^{\mathbf{u},\mathbf{z}} > \gamma/2 \text{ and } \lambda_i^{\mathbf{u}} > 0, \\ \frac{1}{\lambda_i^{\mathbf{u}}}\left(d_i^{\mathbf{u},\mathbf{z}} + \gamma/2\right), & \text{if } d_i^{\mathbf{u},\mathbf{z}} < -\gamma/2 \text{ and } \lambda_i^{\mathbf{u}} > 0. \end{cases}$$

In the following of this chapter, we simply drop the superscripts $\mathbf{u}$ and $\mathbf{z}$ and rewrite $\mathbf{d} = \mathbf{d}^{\mathbf{u},\mathbf{z}}$, $\mathbf{c}_{\gamma}^{\mathbf{u},\mathbf{z}} = \mathbf{c}_{\gamma}$, $d_i = d_i^{\mathbf{u},\mathbf{z}}$, and $c_{\gamma,i} = c_{\gamma,i}^{\mathbf{u},\mathbf{z}}$.

3.3 Main Results

Theorem 3.1. *Let $p \in \{1, 2, \cdots, n\}$ and assume that $f_\rho = L_K^r(g_\rho)$ with some $r > 0$ and $g_\rho \in \mathcal{H}_K$. For any $0 < \delta < 1/3$, if we choose*

$$\gamma \geq 2^{1+2r} \left\|g_\rho\right\|_K \lambda_p^{1+r} + C_{K,\rho}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)\log^{1+r}\frac{2}{\delta},$$

where $C_{K,\rho}$ is a constant given in [38], then there is a universal constant $C < \infty$, which will be specified in the proof, such that

$$\left\|f_\gamma^{\mathbf{u},\mathbf{z}} - f_\rho\right\|_K \leq C\left(\lambda_{p+1}^r + \lambda_p^{\min\{r-1,0\}}\left(\sum_{i=p+1}^{\infty}\lambda_i^{\max\{2r,2\}}\right)^{1/2} + \frac{1}{\lambda_p}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)\log\frac{2}{\delta} + \frac{\sqrt{2p}\gamma}{\lambda_p}\right)$$

with confidence $1 - 3\delta$.

Theorem 3.2. *Assume that $f_\rho = L_K^r(g_\rho)$ for some $r > 0$ and $g_\rho \in \mathcal{H}_K$ and assume that the eigenvalues satisfy*

$$D_1 i^{-\alpha_1} \leq \lambda_i \leq D_2 i^{-\alpha_2}, \text{ for any } i \in \mathbb{N},$$

with $0 < \alpha_2 \leq \alpha_1 < \infty$. We have

(i) If $r \geq 1$ with $\frac{1}{2r} < \alpha_2 \leq \alpha_1 < \alpha_2(1+r) - \frac{1}{2}$, then, by taking

$$\gamma = 2^{1+2r}\left\|g_\rho\right\|_K D_2^{r+1} n^{-\frac{\alpha_2(r+1)}{2(\alpha_1+r\alpha_2)}} + C_{K,\rho}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)\log^{1+r}\frac{2}{\delta},$$

there holds

$$\left\|f_\gamma^{\mathbf{u},\mathbf{z}} - f_\rho\right\|_K \leq C_1 \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} + n^{-\frac{\alpha_2(1+r)}{2(\alpha_1+r\alpha_2)}}\right) n^{\frac{2\alpha_1+1}{4(\alpha_1+r\alpha_2)}} \log^{1+r}\frac{2}{\delta},$$

where C_1 is a universal constant to be specified in the proof.

(ii) If $0 < r < 1$ and $\frac{1}{2r} < \alpha_2 \leq \alpha_1 < \alpha_2(1-r) - \frac{1}{2}$, then, by taking

$$\gamma = 2^{1+2r}\left\|g_\rho\right\|_K D_2^{r+1} n^{-\frac{\alpha_2(r+1)}{2(\alpha_2+r\alpha_1)}} + C_{K,\rho}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)\log^{1+r}\frac{2}{\delta},$$

there holds

$$\left\|f_\gamma^{\mathbf{u},\mathbf{z}} - f_\rho\right\|_K \leq C_2 \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} + n^{-\frac{\alpha_2(1+r)}{2(\alpha_2+r\alpha_1)}}\right) n^{\frac{2\alpha_1+1}{4(\alpha_2+r\alpha_1)}} \log^{1+r}\frac{2}{\delta},$$

where C_2 is a universal constant to be specified in the proof.

When $\alpha_1 = \alpha_2 = \alpha$, one takes $n = \lceil m^s \rceil$ with $s > 0$ to see that the best choice of s is $s = 1$, i.e., $m = n$. In this case, the convergence rate matches the rate of [38] and is slower than the most up-to-date work on empirical features [35]. Bridging the gap between our semi-supervised learning case and [35] is not trivial since the analysis of [35] depends heavily on the orthogonality of $\phi_i^{\mathbf{x}}$ restricted on $\mathbf{x}$, which is not satisfied when $\phi_i^{\mathbf{u}}$ is restricted on $\mathbf{x}$.

3.4 Proof

3.4.1 Technical Lemmas

Lemma 3.1. *For $0 < \delta < 1/3$, there holds*

$$\sum_{i=1}^{n}\left(d_i - \lambda_i^{\mathbf{u}}\left\langle f_\rho, \phi_i^{\mathbf{u}}\right\rangle_K\right)^2 \leq C_3^2\left(\frac{1}{n} + \frac{1}{m}\right)\log^2\frac{2}{\delta}$$

with confidence $1 - 3\delta$. Here $C_3^2 = 4\left(8M\kappa + 4\kappa^2\left\|f_\rho\right\|_K\right)^2$ is a constant independent of $\delta, n, m,$ or γ.

Proof. Note that

$$\lambda_i^{\mathbf{u}} \langle f_\rho, \phi_i^{\mathbf{u}} \rangle_K = \langle L_K^{\mathbf{u}} f_\rho, \phi_i^{\mathbf{u}} \rangle_K .$$

We obtain

$$d_i - \lambda_i^{\mathbf{u}} \langle f_\rho, \phi_i^{\mathbf{u}} \rangle_K = \left\langle \frac{1}{m} \sum_{j=1}^m y_j K_{x_j} - L_K^{\mathbf{u}} f_\rho, \phi_i^{\mathbf{u}} \right\rangle_K$$

and

$$\sum_{i=1}^n \left(d_i - \lambda_i^{\mathbf{u}} \langle f_\rho, \phi_i^{\mathbf{u}} \rangle_K \right)^2 = \sum_{i=1}^n \left\langle \frac{1}{m} \sum_{j=1}^m y_j K_{x_j} - L_K^{\mathbf{u}} f_\rho, \phi_i^{\mathbf{u}} \right\rangle_K^2 \leq \left\| \frac{1}{m} \sum_{j=1}^m y_j K_{x_j} - L_K^{\mathbf{u}} f_\rho \right\|_K^2$$

$$\leq 4 \left(\left\| \frac{1}{m} \sum_{j=1}^m y_j K_{x_j} - L_K^{\mathbf{x}} f_\rho \right\|_K^2 + \| L_K^{\mathbf{x}} f_\rho - L_K f_\rho \|_K^2 + \| L_K^{\mathbf{u}} f_\rho - L_K f_\rho \|_K^2 \right).$$

According to Lemma 2 and Lemma 3 of [38], for each $0 < \delta < 1$, each of the following inequality holds with confidence $1 - \delta$.

$$\left\| \frac{1}{m} \sum_{j=1}^m y_j K_{x_j} - L_K^{\mathbf{x}} f_\rho \right\|_K \leq \frac{8M\kappa \log \frac{2}{\delta}}{\sqrt{m}},$$

$$\| L_K^{\mathbf{x}} f_\rho - L_K f_\rho \|_K \leq \| L_K - L_K^{\mathbf{x}} \|_{\mathrm{op}} \| f_\rho \|_K \leq \| L_K - L_K^{\mathbf{x}} \|_{\mathrm{HS}} \| f_\rho \|_K \leq \frac{4\kappa^2 \| f_\rho \|_K \log \frac{2}{\delta}}{\sqrt{m}},$$

$$\| L_K^{\mathbf{u}} f_\rho - L_K f_\rho \|_K \leq \frac{4\kappa^2 \| f_\rho \|_K \log \frac{2}{\delta}}{\sqrt{n}}.$$

□

Lemma 3.2. *Assume that $f_\rho = L_K^r(g_\rho)$ for some $r > 0$ and $g_\rho \in \mathcal{H}_K$. For any $p \in \{1, 2, \cdots, n\}$ and $0 < \delta < \frac{1}{3}$, if we choose*

$$\gamma \geq 2^{1+2r} \| g_\rho \|_K \lambda_p^{1+r} + C_{K,\rho} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right) \log^{1+r} \frac{2}{\delta},$$

then

$$c_{\gamma,i} = 0, \text{ for any } i = p+1, \cdots, n,$$

with confidence $1 - 3\delta$.

Proof. According to Lemma 3.1, we see that

$$2|d_i| \leq 2\lambda_i^{\mathbf{u}} \left|\langle f_\rho, \phi_i^{\mathbf{u}}\rangle_K\right| + 2\left|d_i - \lambda_i^{\mathbf{u}} \langle f_\rho, \phi_i^{\mathbf{u}}\rangle_K\right| \leq 2\lambda_i^{\mathbf{u}} \left|\langle f_\rho, \phi_i^{\mathbf{u}}\rangle_K\right| + 2C_3 \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right) \log\frac{2}{\delta}.$$

Thanks to Lemma 5 of [38], there holds

$$\begin{aligned} 2\lambda_i^{\mathbf{u}} \left|\langle f_\rho, \phi_i^{\mathbf{u}}\rangle_K\right| &\leq 2\left(\lambda_1^r \|g_\rho\|_K \|L_K - L_K^{\mathbf{u}}\|_{\mathrm{HS}} + 2^r \|g_\rho\|_K (\lambda_i^{\mathbf{u}})^{1+r}\right) \\ &\leq 2\lambda_1^r \|g_\rho\|_K \frac{4\kappa^2 \log\frac{2}{\delta}}{\sqrt{n}} + 2^{1+r} \|g_\rho\|_K (\lambda_i^{\mathbf{u}})^{1+r}. \end{aligned} \tag{3.2}$$

By the proof of Theorem 5 of [38],

$$(\lambda_i^{\mathbf{u}})^{1+r} \leq 2^r \left(\lambda_i^{1+r} + \|L_K - L_K^{\mathbf{u}}\|_{\mathrm{HS}}^{1+r}\right) \leq 2^r \left(\lambda_i^{1+r} + \frac{4\kappa^2 \log^{1+r}\frac{2}{\delta}}{\sqrt{n}}\right).$$

In conclusion,

$$2|d_i| \leq 2^{1+2r} \|g_\rho\|_K \lambda_p^{1+r} + C_{K,\rho} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right) \log^{1+r}\frac{2}{\delta} \leq \gamma$$

with $C_{K,\rho} = 4\kappa^2 (2\lambda_1^r + 2^{1+2r}) \|g_\rho\|_K + 2C_3$. □

3.4.2 Proof of Main Results

Proof of Theorem 3.1. Let $\mathcal{S} = \{1 \leq i \leq p : \lambda_i^{\mathbf{u}} > \lambda_p/2\}$ for $p \in \{1, 2, \cdots, n\}$. Due to equation (3.2), we have

$$\begin{aligned} 2\lambda_i^{\mathbf{u}} \left|\langle f_\rho, \phi_i^{\mathbf{u}}\rangle_K\right| &\leq 2\lambda_1^r \|g_\rho\|_K \|L_K - L_K^{\mathbf{u}}\|_{\mathrm{HS}} + 2^{1+r} \|g_\rho\|_K (\lambda_i^{\mathbf{u}})^{1+r} \\ &\leq 2\lambda_1^r \|g_\rho\|_K \frac{4\kappa^2 \log\frac{2}{\delta}}{\sqrt{n}} + \|g_\rho\|_K \lambda_p^{1+r}, \end{aligned}$$

and $2|d_i| \leq \gamma$, for any $i \in \mathbb{N}\backslash\mathcal{S}$. Thus $c_{\gamma,i} = 0$ for any $i \in \mathbb{N}\backslash\mathcal{S}$. Decompose

$$\left\|f_\gamma^{\mathbf{u},\mathbf{z}} - f_\rho\right\|_K^2 = \sum_{i\in\mathbb{N}\backslash\mathcal{S}} \langle f_\rho, \phi_i^{\mathbf{u}}\rangle_K^2 + \sum_{i\in\mathcal{S}} \left(\langle f_\rho, \phi_i^{\mathbf{u}}\rangle_K - c_{\gamma,i}\right)^2 =: \Delta_1 + \Delta_2.$$

Similarly as the proof of Theorem 4 of [38], we get

$$\sqrt{\Delta_1} \le C_{r,\rho}\left(\lambda_{p+1}^r + \lambda_p^{\min\{r-1,0\}}\left(\sum_{i=p+1}^{\infty}\lambda_i^{\max\{2r,2\}}\right)^{1/2} + \|L_K - L_K^{\mathbf{u}}\|_{\mathrm{HS}}\right)$$

$$\le C_{r,\rho}\left(4\kappa^2+1\right)\left(\lambda_{p+1}^r + \lambda_p^{\min\{r-1,0\}}\left(\sum_{i=p+1}^{\infty}\lambda_i^{\max\{2r,2\}}\right)^{1/2} + \frac{\log\frac{2}{\delta}}{\sqrt{n}}\right).$$

with some constants $c_{r,\rho}$ that is specified in [38].

For Δ_2, we have

$$\Delta_2 = \sum_{i\in\mathcal{S}}\left(\langle f_\rho, \phi_i^{\mathbf{u}}\rangle_K - c_{\gamma,i}\right)^2 \le \sum_{i\in\mathcal{S}}\frac{4}{\lambda_p^2}\left(\lambda_i^{\mathbf{u}}\left(\langle f_\rho, \phi_i^{\mathbf{u}}\rangle_K - c_{\gamma,i}\right)\right)^2$$

$$\le \frac{8}{\lambda_p^2}\left[\sum_{i=1}^{p}\left(d_i - \lambda_i^{\mathbf{u}}\langle f_\rho, \phi_i^{\mathbf{u}}\rangle\right)^2 + \sum_{1\le i\le p, 2|d_i|\le\gamma} d_i^2\right] \le 8\left[C_3^2\left(\frac{1}{n}+\frac{1}{m}\right)\frac{\log^2\frac{2}{\delta}}{\lambda_p^2} + \frac{p\gamma^2}{4\lambda_p^2}\right].$$

We complete the proof by taking $C = 2\sqrt{2}\left(C_3+1\right) + C_{r,\rho}(4\kappa^2+1)$. □

Proof of Theorem 3.2. Let $p = \lceil n^\beta \rceil$ for some $0 < \beta < 1$ to be decided later. Thus $n^\beta \le p \le 2n^\beta$ and

$$\lambda_p^r + \left(\frac{1}{\lambda_p\sqrt{n}} + \frac{1}{\lambda_p\sqrt{m}}\right)\log\frac{2}{\delta} \le D_2^r n^{-\alpha_2\beta r} + \frac{2^{\alpha_1}}{D_1}\left(n^{-\frac{1}{2}+\alpha_1\beta} + m^{-1/2}n^{\alpha_1\beta}\right)\log\frac{2}{\delta}.$$

For $r \ge 1$, there holds

$$\lambda_p^{\min\{r-1,0\}}\left(\sum_{i=p+1}^{\infty}\lambda_i^{\max\{2r,2\}}\right)^{1/2} \le \frac{D_2^r}{\sqrt{2r\alpha_2-1}}n^{-\frac{\beta(2r\alpha_2-1)}{2}}.$$

By the selection of γ, we have

$$\frac{\sqrt{2p}\gamma}{\lambda_p} \le \sqrt{2p}\left(2^{1+2r}\|g_\rho\|_K\lambda_p^r + C_{K,\rho}\frac{1}{\lambda_p}\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right)\log^{1+r}\frac{2}{\delta}\right)$$

$$\le 2\left(2^{1+2r}\|g_\rho\|_K + C_{K,\rho}\right)n^{\frac{\beta}{2}}\left(D_2^r n^{-\alpha_2\beta r} + \frac{2^{\alpha_1}}{D_1}\left(n^{-\frac{1}{2}+\alpha_1\beta} + m^{-\frac{1}{2}}n^{\alpha_1\beta}\right)\right)$$

We obtain the upper bound by taking $\beta = \frac{1}{2(\alpha_1+\alpha_2 r)}$ with

$$C_1 = C\left[\left(D_2^r + \frac{2^{\alpha_1}}{D_1}\right)\left(2\left(2^{1+2r}\left\|g_\rho\right\|_K + C_{K,\rho}\right)+1\right) + \frac{D_2^r}{\sqrt{2\alpha_2 r - 1}}\right].$$

For $0 < r < 1$, it holds

$$\begin{aligned}\lambda_p^{\min\{r-1,0\}}\left(\sum_{i=p+1}^{\infty}\lambda_i^{\max\{2r,2\}}\right)^{1/2} &\leq \frac{D_1^{r-1}p^{-\alpha_1(r-1)}D_2 p^{\frac{1-2\alpha_2}{2}}}{\sqrt{2\alpha_2-1}} \\ &\leq \frac{D_1^{r-1}D_2}{\sqrt{2\alpha_2-1}}n^{\frac{\beta(-2\alpha_2+2\alpha_1(1-r)+1)}{2}}. \qquad (3.3)\end{aligned}$$

We get the upper bound by taking $p = \left\lceil n^{\frac{1}{2(\alpha_2+r\alpha_1)}}\right\rceil$ and

$$C_2 = C\left[\left(D_2^r + \frac{2^{\alpha_1}}{D_1}\right)\left(2\left(2^{1+2r}\left\|g_\rho\right\|_K + C_{K,\rho}\right)+1\right) + \frac{D_1^{r-1}D_2}{\sqrt{2\alpha_2 r - 1}}\right].$$

□

Chapter 4

Modified Poisson Estimators for Grouped and Right-censored Counts

Grouped and right-censored (GRC) count data are widely adopted to study some sensitive topics or to collect information from less cognitive respondents in many research fields, such as psychology, sociology, and criminology. However, theoretical analysis of GRC counts is involved due to the co-existence of grouping schemes and right-censoring schemes. Recently, a modified Poisson regression model has been proposed to analyze GRC count data under the framework of maximum likelihood estimation. In this chapter, we study the asymptotic properties of the maximum likelihood estimators of GRC counts that can cover the modified Poisson estimator. Existing results on modified Poisson estimators for GRC counts are only applicable to stochastic regressors with strictly positive definite Fisher information matrices. Results in this chapter are derived under a milder condition that the information matrix of observations is divergent, which can cover the results for the stochastic case in the almost sure sense. Real data simulations are provided to investigate drug use in America.

4.1 Grouped and Right-censored Count Data

Grouped and right-censored (GRC) counts, as the name suggests, combine both grouped counts, where observations are groups (for example, "1-2 times") instead of separate categories ("once" and "twice"), and right-censored counts, where the upper-end category is from a constant (or a bounded random variable, [57]) to infinity, such as "40 or more times". In many research fields, such as psychology [1] and sociology [4, 7]), GRC counts are widely used to study some sensitive topics (e.g., marijuana use among adolescents [8], Monitoring the Future study among U.S. high school seniors [44]), or to collect information from less cognitive populations [12].

In statistics, studies on right-censored counts have long been established [14,15,65] and implemented [62]. Most of these articles focus on Poisson and zero-inflated Poisson (ZIP) regression models [40, 45]. Sometimes random effects are considered in ZIP regression [55,56]. In parametric regression, the maximum likelihood estimator (MLE) is one of the most efficient estimators. The MLE for generalized linear models (GLM) and its asymptotic theory have been established in statistics for a long time [31,47,74]. The maximum likelihood estimation of right-censored data has also been investigated recently [57].

Methodologically, analyzing GRC count data is more complicated than the right-censored data due to the existence of grouping schemes in right-censored data. Even though grouped and right-censored counts are adopted in survey research for a long time, statistical methods to analyze GRC count data just started recently from [33], where they proposed a Poisson-multinomial mixture approach. A three-step M algorithm was introduced to find the optimal grouping scheme that maximizes the objective function of the Fisher information [32]. Based on MLE, the modified Poisson estimator for GRC counts has been derived recently under a general framework that can cover the Poisson and ZIP models for GRC count data [34].

In this chapter, the asymptotic properties, counting asymptotic existence, (weak and strong) consistency, and asymptotic normality, of MLE of GRC counts are studied under assumptions on the information matrix of the first n observations $\mathbf{F}_n$ that will be defined later. Precisely, we require that $\sigma_{\min}\mathbf{F}_n \to +\infty$ as n tends to infinity. Here $\sigma_{\min}$ is the minimal eigenvalue of a matrix. In some recent work on ZIP regression for right-censored data without grouping schemes [57], they proved the weak consistency of MLE of right-censored data under the assumption that $\{n/\sigma_{\min}\mathbf{F}_n\}_{n=1}^{\infty}$ is a bounded sequence, which is equivalent to that the limit matrix of $\mathbf{F}_n/n$ as $n \to +\infty$ is strictly positive definite. Comparing with [57], our results are applicable to the more involved GRC data under the condition that $\sigma_{\min}\mathbf{F}_n \to +\infty$ without further assumptions on the divergence rate of $\sigma_{\min}\mathbf{F}_n$. For stochastic regressors, the asymptotic properties of MLE of GRC count data were proved under some conditions such that the Fisher information exists and is strictly positive definite [34]. In this case, by the strong law of large numbers, the Fisher information matrix is the (almost sure) limit of $\mathbf{F}_n/n$ as $n \to +\infty$. The results in the strong sense of this chapter (Theorem 4.2) is applicable to both fixed and stochastic regressors. Thus our assumption is weaker than [34] and the asymptotic results for stochastic regressors are given in Corollary 4.3 in this chapter.

The rest of this chapter is organized as follows. In Section 4.2, the maximum likelihood estimators for GRC counts are introduced. Section 4.3 investigates the main results on the asymptotic theory of MLE for GRC counts. Proofs of asymptotic results are provided in Section 4.4. Numerically, the large sample performance of modified Poisson estimators with real data from the MTF (Monitoring the Future) project [44] is studied in Section 4.5.

4.2 Maximum Likelihood Estimators of Grouped and Right-censored Counts

We start from a general framework. Let Y be a discrete random variable from a distribution with the probability mass function $\text{Prob}(Y = k) = \theta(k, \boldsymbol{\xi}), k = 1, 2, \cdots$, parameterized by $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_r)^T \in \mathbb{R}^r, r \in \mathbb{N}$. Each component $\xi_s, 1 \leq s \leq r$, is related to a linear combination of $\mathbf{x}_s = (x_{s,0}, \cdots, x_{s,d_s})^T \in \mathbb{R}^{d_s+1}$ with $d_s \in \mathbb{N}$ through a homeomorphic link function $g_s : \mathbb{R} \to \mathbb{R}$, i.e., $\xi_s = g_s^{-1}\left(\boldsymbol{\beta}_s^T \mathbf{x}_s\right)$. Here each $x_{s,k}$ is a covariate and $\boldsymbol{\beta}_s = (\beta_{s,0}, \cdots, \beta_{s,d_s})^T \in \mathbb{R}^{d_s+1}$ is a vector of parameters to be estimated. When $x_{s,0} = 1$, $\beta_{s,0}$ is known as an intercept term. For simplicity, let $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_r^T\right)^T \in \mathfrak{B} \subset \mathbb{R}^d$ and $\mathbf{x} = \left(\mathbf{x}_1^T, \cdots, \mathbf{x}_r^T\right)^T \in \mathcal{X} \subset \mathbb{R}^d$ with $d = d_1 + \cdots + d_r + r$. Assume tacitly that the parameter space $\mathfrak{B}$ of $\boldsymbol{\beta}$ is convex with non-empty interior. Denote $\Xi \subset \mathbb{R}^r$ the parameter space of $\boldsymbol{\xi} \in \mathbb{R}^r$, that is, $\Xi = \left\{\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\beta}, \mathbf{x}) \in \mathbb{R}^r : \xi_s = g_s^{-1}(\boldsymbol{\beta}_s^T \mathbf{x}_s), 1 \leq s \leq r, \boldsymbol{\beta} \in \mathfrak{B}, \mathbf{x} \in \mathcal{X}\right\}$.

The general framework above can cover two specific models that are ubiquitous in statistics: the Poisson model ($r = 1$) and the zero-inflated Poisson (ZIP) model ($r = 2$). For the Poisson regression model with a Poisson parameter $\mu > 0$, where the probability mass function is

$$\theta_{\text{P}}(k, \mu) = e^{-\mu} \frac{\mu^k}{k!}, k = 0, 1, 2, \cdots, \tag{4.1}$$

the natural link function for μ is the log link $g_{\log}(\mu) = \log \mu$ with $g_{\log}^{-1}(t) = e^t$. The Poisson distribution possesses the equi-dispersion property, that is, the mean and the variance of the Poisson distribution are equal. However, it has been observed that sometimes the equi-dispersion assumption is violated and one proposed the ZIP model [40, 45]. For the zero-inflated Poisson model with a Bernoulli parameter $0 < p < 1$ and a Poisson parameter $\mu > 0$, the probability mass function is

$$\theta_{\text{ZIP}}\left(k, (\mu, p)^T\right) = \begin{cases} p + (1-p)e^{-\mu}, & k = 0, \\ (1-p)e^{-\mu}\frac{\mu^k}{k!}, & k = 1, 2, \cdots. \end{cases} \tag{4.2}$$

For the Poisson parameter μ of ZIP models, one can still use $g_{\log}$ as a link function. For the Bernoulli parameter p, one may select, for example, the logit link $g_{\text{logit}}(p) = \log\left(\frac{p}{1-p}\right)$ with $g^{-1}_{\text{logit}}(t) = (1+e^{-t})^{-1}$.

For GRC counts, a grouping scheme $\mathcal{G}$ with $M \in \mathbb{N}$ groups is defined through partitioning $\mathbb{N}$ by fixed integers $0 = l_1 < l_2 < \cdots < l_{M+1} = +\infty$ with the k'th group given by $\text{Group}_k = \{m \in \mathbb{N}, l_k \leq m < l_{k+1}\}, 1 \leq k \leq M$. In this chapter, we consider the case $M < +\infty$, i.e., the number of groups is finite, which is general in practice. Consider a random variable $Y_{\mathcal{G}}$, that is the group in which Y lies, taking values in $\{1, 2, \cdots, M\}$. $Y_{\mathcal{G}}$ is obviously from a multinomial distribution with the probability mass function

$$\text{Prob}\left(Y_{\mathcal{G}} = k\right) = \theta^{\mathcal{G}}(k, \boldsymbol{\xi}) = \sum_{j=l_k}^{l_{k+1}-1} \theta(j, \boldsymbol{\xi}), 1 \leq k \leq M.$$

Let $\{(\mathbf{x}^i, Y^i_{\mathcal{G}})\}_{i=1}^n \subset \mathcal{X} \times \{1, 2, \cdots, M\}, n \in \mathbb{N}$, be a sample drawn from a distribution with respect to the parameter $\boldsymbol{\beta}^*$, that is, $\text{Prob}\left(Y^i_{\mathcal{G}} = k\right) = \theta^{\mathcal{G}}(k, \boldsymbol{\xi}^i_* = \boldsymbol{\xi}(\boldsymbol{\beta}^*, \mathbf{x}^i))$. Here $\left\{Y^i_{\mathcal{G}}\right\}_{i=1}^n$ is a sequence of independent random variables. In most cases of this chapter, let the regressors $\mathbf{x}^i, i = 1, 2, \cdots, n$, be fixed and expectations are taken over $Y^i_{\mathcal{G}} \in \{1, 2, \cdots, M\}$. As shown in Corollary 4.3, our results in the strong sense can be extended to stochastic regressors by considering the conditional expectation as conditioned on $\{\mathbf{x}^i\}_{i=1}^n$ and the law of total probability.

In parametric regression, the aim is to estimate the true parameter $\boldsymbol{\beta}^*$ by generating an estimator $\hat{\boldsymbol{\beta}}_n$ from the sample $\{(\mathbf{x}^i, Y^i_{\mathcal{G}})\}_{i=1}^n$. In this chapter, we consider the maximum likelihood estimator of GRC counts, i.e., $\hat{\boldsymbol{\beta}}_n$ is the maximizer of the log-likelihood function

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^n \log \theta^{\mathcal{G}}(Y^i_{\mathcal{G}}, \boldsymbol{\xi}^i = \boldsymbol{\xi}(\boldsymbol{\beta}, \mathbf{x}^i)). \tag{4.3}$$

Assume that the true parameter $\boldsymbol{\beta}^*$ is contained in the interior of $\mathfrak{B}$. For simplicity,

in the following of this chapter, we just write $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ without specifying the corresponding parameter spaces $\mathfrak{B}$ and Ξ.

Let $\|\cdot\|$ be the Frobenius norm of a matrix and the Euclidean norm of a vector, respectively. The operator norm of a matrix is written as $\|\cdot\|_{\text{op}}$. $\mathbb{E}_{\boldsymbol{\beta}}$ ($\mathbb{E}_{\boldsymbol{\xi}}$) denotes the expectation of a random variable with respect to the parameter $\boldsymbol{\beta}$ ($\boldsymbol{\xi}$). $\text{Var}_{\boldsymbol{\beta}}$ is the variance of a random variable with respect to $\boldsymbol{\beta}$. $\sigma_{\min}$ ($\sigma_{\max}$) denotes the minimal (maximal) eigenvalue of a matrix. $\mathbf{0}$ and $\mathbf{I}$ are the zero matrix and the identity matrix whose dimension can be verified from contexts, correspondingly. $\to_d$ and $\to_p$ mean convergence in distribution and in probability, respectively.

4.3 Asymptotic Theory

Recall the log-likelihood function $\ell_n(\boldsymbol{\beta})$ of GRC counts defined by (4.3). Let $\mathbf{s}_n(\boldsymbol{\beta}) \in \mathbb{R}^d$ and $\mathbf{H}_n(\boldsymbol{\beta}) \in \mathbb{R}^{d\times d}$ be the gradient and the Hessian matrix of $\ell_n(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, respectively. Since our error analysis is based on Taylor's expansion

$$\ell_n(\boldsymbol{\beta}) - \ell_n(\boldsymbol{\beta}^*) = \Delta\boldsymbol{\beta}^T \mathbf{s}_n(\boldsymbol{\beta}^*) + \frac{\Delta\boldsymbol{\beta}^T \mathbf{H}_n(\tilde{\boldsymbol{\beta}})\Delta\boldsymbol{\beta}}{2}, \tag{4.4}$$

where $\Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and $\tilde{\boldsymbol{\beta}}$ is a point between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, we need some assumptions on $\mathbf{s}_n(\boldsymbol{\beta})$ and on the Fisher information of the first n observations $\mathbf{F}_n(\boldsymbol{\beta}) := -\mathbb{E}_{\boldsymbol{\beta}}[\mathbf{H}_n(\boldsymbol{\beta})]$ for $\boldsymbol{\beta}$ in a neighborhood around $\boldsymbol{\beta}^*$.

The structure of $\mathbf{F}_n(\boldsymbol{\beta})$ for GRC counts is complicated. Let $\mathbf{X}^i = \text{Diag}\{\mathbf{x}_1^i, \cdots, \mathbf{x}_r^i\}$, which is a $d \times r$ block diagonal matrix, and

$$\mathbf{U}(\boldsymbol{\beta}, \mathbf{x}) := \text{Diag}\left\{U_1(\boldsymbol{\beta}, \mathbf{x}), U_2(\boldsymbol{\beta}, \mathbf{x}), \cdots, U_r(\boldsymbol{\beta}, \mathbf{x})\right\} \in \mathbb{R}^{r\times r}$$

with $U_s(\boldsymbol{\beta}, \mathbf{x}) = \left(g_s^{-1}\right)'(\boldsymbol{\beta}_s^T \mathbf{x}_s), s = 1, 2, \cdots, r$. Simply rewrite $\mathbf{U}^i(\boldsymbol{\beta}) = \mathbf{U}(\boldsymbol{\beta}, \mathbf{x}^i)$. Straightforward but tedious calculations show that $\mathbf{s}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{X}^i \mathbf{U}^i(\boldsymbol{\beta}) \mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}^i)$ with

$$\mathbf{s}_{Y_{\mathcal{G}}}(\boldsymbol{\xi}) = \left(\frac{\partial}{\partial\xi_1}\log\theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi}), \frac{\partial}{\partial\xi_2}\log\theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi}), \cdots, \frac{\partial}{\partial\xi_r}\log\theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi})\right)^T.$$

Moreover, we have

$$\mathbf{H}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[-\mathbf{X}^i \mathbf{U}^i(\boldsymbol{\beta}) \mathbb{I}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}^i) \left(\mathbf{X}^i \mathbf{U}^i(\boldsymbol{\beta}) \right)^T + \mathbf{R}^i(\boldsymbol{\beta}) \right]. \tag{4.5}$$

Here, for a given random variable $Y_{\mathcal{G}} \in \{1, 2, \cdots, M\}$, $\mathbb{I}_{Y_{\mathcal{G}}} \in \mathbb{R}^{r \times r}$ is defined by

$$\left(\mathbb{I}_{Y_{\mathcal{G}}}(\boldsymbol{\xi}) \right)_{st} = -\frac{\partial^2}{\partial \xi_s \partial \xi_t} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi}), s, t = 1, 2, \cdots, r,$$

and $\mathbf{R}^i(\boldsymbol{\beta}) = \mathbf{X}^i \mathbf{W}^i(\boldsymbol{\beta}) \mathbf{S}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}^i) \left(\mathbf{X}^i \right)^T$ with

$$\mathbf{W}^i(\boldsymbol{\beta}) = \mathrm{Diag} \left\{ \left(g_1^{-1} \right)'' (\boldsymbol{\beta}_1^T \mathbf{x}_1^i), \cdots, \left(g_r^{-1} \right)'' (\boldsymbol{\beta}_r^T \mathbf{x}_r^i) \right\} \in \mathbb{R}^{r \times r}$$

and

$$\mathbf{S}_{Y_{\mathcal{G}}}(\boldsymbol{\xi}) = \mathrm{Diag} \left(\frac{\partial}{\partial \xi_1} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi}), \cdots, \frac{\partial}{\partial \xi_r} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi}) \right) \in \mathbb{R}^{r \times r}.$$

Note that $\mathbb{E}_{\boldsymbol{\xi}} \left[\frac{\partial}{\partial \xi_s} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi}) \right] = \frac{\partial}{\partial \xi_s} \sum_{j=1}^{M} \theta^{\mathcal{G}}(j, \boldsymbol{\xi}) = \frac{\partial}{\partial \xi_s} 1 = 0$ and $\mathbb{E}_{\boldsymbol{\beta}}[\mathbf{s}_n(\boldsymbol{\beta})] = 0$. We obtain $\mathbb{E}_{\boldsymbol{\beta}}[\mathbf{R}^i(\boldsymbol{\beta})] = 0$ and

$$\mathbf{F}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}^i \mathbf{U}^i(\boldsymbol{\beta}) \mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}^i) \left(\mathbf{X}^i \mathbf{U}^i(\boldsymbol{\beta}) \right)^T = \mathbb{E}_{\boldsymbol{\beta}} \left[\mathbf{s}_n(\boldsymbol{\beta}) \mathbf{s}_n(\boldsymbol{\beta})^T \right], \tag{4.6}$$

where $\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{\xi}} \left[\mathbb{I}_{Y_{\mathcal{G}}}(\boldsymbol{\xi}) \right] = \left(I_{s,t}^{\mathcal{G}}(\boldsymbol{\xi}) \right)_{r \times r}$ is the Fisher information of $Y_{\mathcal{G}}$ with respect to $\boldsymbol{\xi}$ and the last equality is because

$$I_{s,t}^{\mathcal{G}}(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{\xi}} \left[-\frac{\partial^2}{\partial \xi_s \partial \xi_t} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi}) \right] = \mathbb{E}_{\boldsymbol{\xi}} \left[\frac{\partial}{\partial \xi_s} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi}) \frac{\partial}{\partial \xi_t} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi}) \right].$$

When $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, we simply drop the parameter $\boldsymbol{\beta}^*$ and rewrite $\mathbf{s}_n(\boldsymbol{\beta}^*), \mathbf{F}_n(\boldsymbol{\beta}^*), \mathbf{H}_n(\boldsymbol{\beta}^*)$, $\mathbf{U}^i(\boldsymbol{\beta}^*), \mathbb{E}_{\boldsymbol{\beta}^*}, \mathrm{Var}_{\boldsymbol{\beta}^*}$ as $\mathbf{s}_n, \mathbf{F}_n, \mathbf{H}_n, \mathbf{U}^i, \mathbb{E}, \mathrm{Var}$, accordingly.

Theorem 4.1. *Let $\mathcal{X}$ be a compact set. Assume that, for any $s = 1, 2, \cdots, r$,*

(i) g_s^{-1} is C^2 with $(g_s^{-1})' > 0$.

(ii) $\theta^{\mathcal{G}}\left(Y_{\mathcal{G}}, \boldsymbol{\xi}\right)$ *is* C^2 *with respect to* $\boldsymbol{\xi}$ *such that* $\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi})$ *is strictly positive definite everywhere.*

(iii) $\sigma_{\min}\mathbf{F}_n \to +\infty$, *as* $n \to +\infty$.

Then, there is a sequence $\{\hat{\boldsymbol{\beta}}_n\}_{n=1}^{+\infty}$ *of random variables such that, as* $n \to +\infty$,

(i) $\mathbb{P}\left[\mathbf{s}_n(\hat{\boldsymbol{\beta}}_n) = 0\right] \to 1$ *(asymptotic existence),*

(ii) $\hat{\boldsymbol{\beta}}_n \to_p \boldsymbol{\beta}^*$ *(weak consistency),*

(iii) $\mathbf{F}_n^{T/2}\left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right] \to_d \mathcal{N}(0, \mathbf{I})$ *(asymptotic normality).*

The assumption *(i)* of Theorem 4.1, which implies that the inverse function of the link function is strictly increasing, is fulfilled by most link functions used in practice. Specifically, for Poisson and ZIP models, it is obvious that $\left(g_{\log}^{-1}\right)'(t) = e^t > 0$ and $\left(g_{\text{logit}}^{-1}\right)'(t) = \frac{e^{-t}}{(1+e^{-t})^2} > 0$. The assumption *(ii)* of Theorem 4.1 is satisfied when $M \geq 2$ for the Poisson model and $M \geq 3$ for the ZIP case, according to [32]. The assumption that $\mathbf{F}_n$ is strictly positive definite for $n \in \mathbb{N}$ large enough with

$$\sigma_{\min}\mathbf{F}_n \to +\infty, n \to +\infty, \tag{4.7}$$

is common in literature [31].

We now give another insight on the assumption (4.7). For any $\mathbf{u} = \left(\mathbf{u}_1^T, \cdots, \mathbf{u}_r^T\right)^T \in \mathbb{R}^d$, denote $\tilde{\mathbf{u}}^i = \left(\mathbf{u}_1^T\mathbf{x}_1^i, \cdots, \mathbf{u}_r^T\mathbf{x}_r^i\right)^T \in \mathbb{R}^r$. Then

$$\begin{aligned}
\mathbf{u}^T\mathbf{F}_n\mathbf{u} &= \sum_{i=1}^{n}\left(\tilde{\mathbf{u}}^i\right)^T\mathbf{U}^i\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}_*^i)\mathbf{U}^i\tilde{\mathbf{u}}^i \\
&\geq \sum_{i=1}^{n}\sigma_{\min}\left[\mathbf{U}^i\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}_*^i)\mathbf{U}^i\right]\sum_{s=1}^{r}\left(\mathbf{u}_s^T\mathbf{x}_s^i\right)^2 \\
&\geq \left(\min_{\mathbf{x}\in\mathcal{X}}\left\{\sigma_{\min}\left[\mathbf{U}(\boldsymbol{\beta}^*, \mathbf{x})\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}(\boldsymbol{\beta}^*, \mathbf{x}))\mathbf{U}(\boldsymbol{\beta}^*, \mathbf{x})\right]\right\}\right)\sum_{i=1}^{n}\sum_{s=1}^{r}\left(\mathbf{u}_s^T\mathbf{x}_s^i\right)^2.
\end{aligned}$$

By the eigenvalue perturbation theory (Corollary 6.3.8, [42]), $\sigma_{\min}$ and $\sigma_{\max}$ are (Lipschitz) continuous on the Hermitian matrix space. Thus, under assumptions *(i)* and *(ii)* of Theorem 4.1, $\left(\min_{\mathbf{x}\in\mathcal{X}}\left\{\sigma_{\min}\left[\mathbf{U}(\boldsymbol{\beta}^*,\mathbf{x})\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}(\boldsymbol{\beta}^*,\mathbf{x}))\mathbf{U}(\boldsymbol{\beta}^*,\mathbf{x})\right]\right\}\right)$ is bounded away from 0, when $\mathcal{X}$ is compact. And (4.7) is implied by

$$\sigma_{\min}\left(\sum_{i=1}^{n}\mathbf{x}_s^i\left(\mathbf{x}_s^i\right)^T\right)\to+\infty, n\to+\infty, \text{ for any } s=1,2,\cdots,r. \tag{4.8}$$

Corollary 4.1. *Let $\mathcal{X}$ be compact and assume (4.8). If assumptions (i) and (ii) of Theorem 4.1 hold, then there is a sequence $\{\hat{\boldsymbol{\beta}}_n\}_{n=1}^{+\infty}$ of random variables such that there hold all conclusions of Theorem 4.1.*

To state the strong consistency, for any given $\epsilon>0$, we define a ball $B_\epsilon(\boldsymbol{\beta}^*)=\{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|\le\epsilon\}$.

Theorem 4.2. *Let $\mathcal{X}$ be a compact set. Assume that*

(i) For any $s=1,2,\cdots,r$, g_s^{-1} is C^2 with $\left(g_s^{-1}\right)'>0$.

(ii) $\theta^{\mathcal{G}}\left(Y_{\mathcal{G}},\boldsymbol{\xi}\right)$ is C^2 with respect to $\boldsymbol{\xi}$ such that $\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi})$ is strictly positive definite everywhere.

(iii) $\sigma_{\min}\mathbf{F}_n\to+\infty$, as $n\to+\infty$.

(iv) There exist $\epsilon>0$ and a fixed number $n_0\in\mathbb{N}$ such that for any $n\ge n_0$ and $\boldsymbol{\beta}\in B_\epsilon(\boldsymbol{\beta}^*)$, $\sigma_{\min}\mathbf{F}_n(\boldsymbol{\beta})\ge c\sigma_{\max}\mathbf{F}_n$ with a universal constant $c>0$.

Then, there exist a sequence $\{\hat{\boldsymbol{\beta}}_n\}_{n=1}^{\infty}$ of random variables and a random number $\tilde{n}_0\in\mathbb{N}$ such that

(i) $\mathbb{P}[\mathbf{s}_n(\hat{\boldsymbol{\beta}}_n)=0,$ for all $n\ge\tilde{n}_0]=1$ (asymptotic existence),

(ii) $\hat{\boldsymbol{\beta}}_n\to\boldsymbol{\beta}^*$, a.s., as $n\to+\infty$ (strong consistency),

(iii) $\mathbf{F}_n^{T/2}\left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right] \to_d \mathcal{N}(0, \mathbf{I})$, *as* $n \to +\infty$ *(asymptotic normality).*

Corollary 4.2. *Let* $\mathcal{X}$ *be a compact set. Assume that for any* $s \in \{1, 2, \cdots, r\}$, *there hold (4.8) and*

$$\sigma_{\min}\left(\sum_{i=1}^{n} \mathbf{X}^i \left(\mathbf{X}^i\right)^T\right) \geq \tilde{c}\sigma_{\max}\left(\sum_{i=1}^{n} \mathbf{X}^i \left(\mathbf{X}^i\right)^T\right), \text{ for all } n \geq n_1, \tag{4.9}$$

with a fixed number $n_1 \in \mathbb{N}$ *and a universal constant* $\tilde{c} > 0$. *If, in addition, assumptions (i) and (ii) of Theorem 4.2 are satisfied, then there exist a sequence* $\{\hat{\boldsymbol{\beta}}_n\}_{n=1}^{\infty}$ *of random variables and a random number* $\tilde{n}_0 \in \mathbb{N}$ *such that all conclusions of Theorem 4.2 hold.*

In Corollary 4.2, equation (4.9) says that the sequence $\{\kappa_n\}_{n=1}^{\infty}$ of condition numbers of $\sum_{i=1}^{n} \mathbf{X}^i \left(\mathbf{X}^i\right)^T$, i.e.,

$$\kappa_n := \frac{\sigma_{\max}\left[\sum_{i=1}^{n} \mathbf{X}^i \left(\mathbf{X}^i\right)^T\right]}{\sigma_{\min}\left[\sum_{i=1}^{n} \mathbf{X}^i \left(\mathbf{X}^i\right)^T\right]},$$

is bounded uniformly for $n \in \mathbb{N}$. By the definition of $\mathbf{X}^i$, (4.9) is equivalent to

$$\min_{s\in\{1,2,\cdots,r\}}\left\{\sigma_{\min}\left(\sum_{i=1}^{n} \mathbf{x}_s^i \left(\mathbf{x}_s^i\right)^T\right)\right\} \geq \tilde{c} \max_{s\in\{1,2,\cdots,r\}}\left\{\sigma_{\max}\left(\sum_{i=1}^{n} \mathbf{x}_s^i \left(\mathbf{x}_s^i\right)^T\right)\right\}. \tag{4.10}$$

For stochastic regressors, consider $\mathbf{F}_n(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}}\left[-\mathbf{H}_n(\boldsymbol{\beta}) \left| \{\mathbf{x}^i\}_{i=1}^n \right.\right]$ and $\mathbf{F}_n = \mathbf{F}_n(\boldsymbol{\beta}^*)$, which is the same as the fixed design case. Thanks to the strong law of large numbers, (4.8) and (4.10) are implied by that $\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}_s\mathbf{x}_s^T\right]$ exists and is strictly positive definite for each $1 \leq s \leq r$, which is a condition required by [34]. Here the expectation is taken with respect to the marginal distribution of $\mathbf{x}_s, s = 1, 2, \cdots, r$. Moreover, for stochastic regressors with a strictly positive definite Fisher information matrix $\mathbb{F} := \mathbb{E}\mathbf{F}_1$ with the expectation taken over $\mathbf{x}^1 \in \mathcal{X}$, from the proof of Corollary 3 of [31], one can obtain the following corollary easily according to Theorem 4.2.

Corollary 4.3. *Let $\mathcal{X}$ be a compact set. If assumptions (i) and (ii) of Theorem 4.2 are fulfilled and the Fisher information $\mathbb{F}$ exists and is strictly positive definite, then there are a sequence $\{\hat{\boldsymbol{\beta}}_n\}_{n=1}^{\infty}$ of random variables and a random number $\tilde{n}_0 \in \mathbb{N}$ such that conclusions (i) and (ii) of Theorem 4.2 hold with $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right) \to_d \mathcal{N}(0, \mathbb{F}^{-1})$, as $n \to +\infty$.*

In Corollary 4.3, the strict positive definiteness of $\mathbb{F}$ is implied by the strict positive definiteness of $\mathbb{E}\left[\mathbf{x}_s\mathbf{x}_s^T\right]$ and $\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi})$, as pointed out by [34].

4.4 Proofs of Asymptotic Properties

For GRC count data, the Hessian matrix $\mathbf{H}_n(\boldsymbol{\beta})$ and the information matrix $\mathbf{F}_n(\boldsymbol{\beta})$, as shown in (4.5) and (4.6), become much more involved, comparing with classical analysis on the generalized linear models. The technical difficulties in analyzing GRC counts are caused by the discrepancy between $\mathbf{F}_n(\boldsymbol{\beta})$ and $-\mathbf{H}_n(\boldsymbol{\beta})$. Most of the auxiliary results in sections 4.4.1 and 4.4.2 are motivated by [31], but the technical details are different.

4.4.1 Some Properties of the Information Matrix

We first derive some properties of $\mathbf{F}_n$, which play important roles in the sequel proofs. Since $\mathbf{F}_n$ is positive semi-definite, one can decompose it as $\mathbf{F}_n = \mathbf{F}_n^{1/2}\mathbf{F}_n^{T/2}$ (for example, through the Cholesky decomposition or through the eigendecomposition and taking the square root of each eigenvalue), where $\mathbf{F}_n^{T/2} = \left(\mathbf{F}_n^{1/2}\right)^T$. To prove the next proposition, we introduce the Loewner partial order "$\succcurlyeq$" and "$\preccurlyeq$" between Hermitian matrices, that is, for two Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succcurlyeq \mathbf{B}$ ($\mathbf{A} \preccurlyeq \mathbf{B}$) if $\mathbf{A} - \mathbf{B}$ is positive (negative) semi-definite. For properties of Loewner's partial order, one may refer to Chapter 7.7 of [42]. Let $\mathrm{Tr}(\mathbf{A})$ be the trace of a matrix $\mathbf{A}$. We summarize some properties of $\mathbf{F}_n$ in the following proposition.

Proposition 4.1. *Let $\mathcal{X}$ be compact. Under assumptions (i) and (ii) of Theorem 4.1, there is a constant $C < +\infty$ such that*

$$\sum_{i=1}^{n} \mathrm{Tr}\left[\left(\mathbf{X}^i\right)^T \mathbf{F}_n^{-1} \mathbf{X}^i\right] \leq C, \text{ for } n \text{ large enough.} \tag{4.11}$$

If we further assume (iii) of Theorem 4.1, then

$$\mathrm{Tr}\left[\left(\mathbf{X}^n\right)^T \mathbf{F}_n^{-1} \mathbf{X}^n\right] \to 0 \tag{4.12}$$

and

$$\max_{1 \leq i \leq n} \mathrm{Tr}\left[\left(\mathbf{X}^i\right)^T \mathbf{F}_n^{-1} \mathbf{X}^i\right] \to 0, \tag{4.13}$$

as $n \to +\infty$.

Proof. Denote $\mathbf{X}_n = \sum_{i=1}^{n} \mathbf{X}^i \left(\mathbf{X}^i\right)^T \succcurlyeq \mathbf{0} \in \mathbb{R}^{d \times d}$. Then we have $\sum_{i=1}^{n} \mathrm{Tr}\left[\left(\mathbf{X}^i\right)^T \mathbf{F}_n^{-1} \mathbf{X}^i\right] = \mathrm{Tr}\left[\mathbf{F}_n^{-1} \mathbf{X}_n\right] = \mathrm{Tr}\left[\mathbf{F}_n^{-1/2} \mathbf{X}_n \mathbf{F}_n^{-T/2}\right]$. Rewrite

$$\mathbf{F}_n = \sum_{i=1}^{n} \mathbf{X}^i \mathbf{U}^i \mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}_*^i) \left(\mathbf{X}^i \mathbf{U}^i\right)^T = \sum_{i=1}^{n} \mathbf{X}^i \mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}^i) \left(\mathbf{X}^i\right)^T$$

with $\mathbf{C}(\boldsymbol{\beta}, \mathbf{x}) = \mathbf{U}(\boldsymbol{\beta}, \mathbf{x}) \mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}(\boldsymbol{\beta}, \mathbf{x})) \mathbf{U}(\boldsymbol{\beta}, \mathbf{x})$. Since $(g_s^{-1})'(\boldsymbol{\beta}_s^{*T} \mathbf{x}_s) > 0$, for each $s = 1, 2, \cdots, r$, which implies that $\mathbf{U}(\boldsymbol{\beta}^*, \mathbf{x})$ is of full rank, and $\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi})$ is strictly positive definite everywhere, we have $\mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x})$ is strictly positive definite, for any $\mathbf{x} \in \mathcal{X}$. Since $\sigma_{\min}$ is continuous on the Hermitian matrix space, we obtain that there is a constant $C < +\infty$ such that

$$\min_{\mathbf{x} \in \mathcal{X}} \left\{\sigma_{\min} \mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x})\right\} \geq d/C > 0 \text{ and } \mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}) \succcurlyeq \frac{d}{C} \mathbf{I}, \text{ for any } \mathbf{x} \in \mathcal{X}.$$

Thus $\mathbf{F}_n \succcurlyeq \frac{d}{C} \mathbf{X}_n$ and

$$\mathbf{I} = \mathbf{F}_n^{-1/2} \mathbf{F}_n \mathbf{F}_n^{-T/2} \succcurlyeq \frac{d}{C} \mathbf{F}_n^{-1/2} \mathbf{X}_n \mathbf{F}_n^{-T/2}. \tag{4.14}$$

We get (4.11) by taking trace to both sides of (4.14). Since $\mathbf{F}_n \succcurlyeq (\sigma_{\min} \mathbf{F}_n) \mathbf{I}$ and $\mathbf{F}_n^{-1} \preccurlyeq (\sigma_{\min} \mathbf{F}_n)^{-1} \mathbf{I}$.

(4.12) and (4.13) are proved by the compactness of $\mathcal{X}$ and equation (4.7). □

4.4.2 Some Lemmas

We list some lemmas that will be used in the proofs of asymptotic results. For simplicity, denote $\mathbf{V}_n(\boldsymbol{\beta}) = -\mathbf{F}_n^{-1/2}\mathbf{H}_n(\boldsymbol{\beta})\mathbf{F}_n^{-T/2}$ and introduce a neighborhood around $\boldsymbol{\beta}^*$ by $\mathcal{N}_n(\delta) = \left\{\boldsymbol{\beta} : \left\|\mathbf{F}_n^{T/2}\left[\boldsymbol{\beta} - \boldsymbol{\beta}^*\right]\right\| \leq \delta\right\}$, for any $n = 1, 2, \cdots$, and $\delta > 0$.

Lemma 4.1. *Assume (4.7) and assume that, for any $\delta > 0$,*

$$\mathbb{P}\left[\sigma_{\min}\mathbf{V}_n(\boldsymbol{\beta}) \geq c_1, \text{ for all } \boldsymbol{\beta} \in \mathcal{N}_n(\delta)\right] \to 1, \text{ as } n \to +\infty, \tag{4.15}$$

with some constants $c_1 > 0$ independent of δ, then there is a sequence $\{\hat{\boldsymbol{\beta}}_n\}$ of estimators such that conclusions (i) and (ii) of Theorem 4.1 hold.

The proof of Lemma 4.1 follows mainly the proof of Theorem 1 of [31] by noting that (4.15) is equivalent to

$$\mathbb{P}\left[-\mathbf{H}_n(\boldsymbol{\beta}) - c_1\mathbf{F}_n \text{ is positive semi-definite for all } \boldsymbol{\beta} \in \mathcal{N}_n(\delta)\right] \to 1,$$

as $n \to +\infty$, which is equivalent to the condition (C*) in Section 4 of [31].

The next lemma is the asymptotic normality of $\mathbf{F}_n^{-1/2}\mathbf{s}_n$.

Lemma 4.2. *Let $\mathcal{X}$ be compact. If assumptions (i), (ii), and (iii) of Theorem 4.1 hold, then $\mathbf{F}_n^{-1/2}\mathbf{s}_n \to_d \mathcal{N}(0, \mathbf{I})$.*

Proof. Note that $\mathbb{E}\mathbf{s}_n = 0$ and $\mathbb{E}\left[\mathbf{F}_n^{-1/2}\mathbf{s}_n\mathbf{s}_n^T\mathbf{F}_n^{-T/2}\right] = \mathbf{F}_n^{-1/2}\mathbf{F}_n\mathbf{F}_n^{-T/2} = \mathbf{I}$. It suffices to check the Lindeberg-Feller condition (cf. Proposition 2.27, [74]) for $\mathbf{v}_{ni} := \mathbf{F}_n^{-1/2}\mathbf{X}^i\mathbf{U}^i\mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}_*^i)$, i.e., to prove $g_n(\delta) = \sum_{i=1}^n \mathbb{E}\left[\|\mathbf{v}_{ni}\|^2 \mathbf{1}_{[\|\mathbf{v}_{ni}\|>\delta]}\right] \to 0$, as $n \to +\infty$, for any $\delta > 0$. If we denote $\mathbf{Z}_{ni} = \mathbf{F}_n^{-1/2}\mathbf{X}^i\mathbf{U}^i$, then

$$\begin{aligned} g_n(\delta) &\leq \sum_{i=1}^n \|\mathbf{Z}_{ni}\|_{\text{op}}^2 \, \mathbb{E}\left[\left\|\mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}_*^i)\right\|^2 \mathbf{1}_{\left[\left\|\mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}_*^i)\right\|^2 > \delta^2/\|\mathbf{Z}_{ni}\|_{\text{op}}^2\right]}\right] \\ &\leq \max_{1\leq i\leq n}\left\{\mathbb{E}\left[\left\|\mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}_*^i)\right\|^2 \mathbf{1}_{\left[\left\|\mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}_*^i)\right\|^2 > \delta^2/\|\mathbf{Z}_{ni}\|_{\text{op}}^2\right]}\right]\right\}\left(\sum_{i=1}^n \|\mathbf{Z}_{ni}\|_{\text{op}}^2\right). \end{aligned} \tag{4.16}$$

According to (4.11) and the boundedness of $\mathbf{U}(\boldsymbol{\beta}^*, \mathbf{x})$ with respect to $\mathbf{x} \in \mathcal{X}$, there is a constant $\tilde{C} < +\infty$ such that

$$\sum_{i=1}^{n} \|\mathbf{Z}_{ni}\|_{\text{op}}^2 \leq \left(\max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{U}(\boldsymbol{\beta}^*, \mathbf{x})\|_{\text{op}}^2 \right) \left(\sum_{i=1}^{n} \text{Tr} \left[\left(\mathbf{X}^i\right)^T \mathbf{F}_n^{-1} \mathbf{X}^i \right] \right) \leq \tilde{C},$$

for all $n \in \mathbb{N}$. According to (4.13),

$$\max_{1 \leq i \leq n} \|\mathbf{Z}_{ni}\|_{\text{op}}^2 \leq \left[\max_{1 \leq i \leq n} \left\| \mathbf{F}_n^{-1/2} \mathbf{X}^i \right\|_{\text{op}}^2 \right] \left[\max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{U}(\boldsymbol{\beta}^*, \mathbf{x})\|_{\text{op}}^2 \right] \to 0$$

and $C_n := \frac{1}{\max_{1 \leq i \leq n} \|\mathbf{Z}_{ni}\|_{\text{op}}^2} \to +\infty$, as $n \to +\infty$. Since $\mathcal{X}$ is compact and g_s is homeomorphic, $\boldsymbol{\xi}(\boldsymbol{\beta}^*, \mathbf{x})$ lies in a compact subset of $\mathbb{R}^r$ for $\mathbf{x}$ running over $\mathcal{X}$ and $\mathbf{s}_{Y_{\mathcal{G}}}(\boldsymbol{\xi}(\boldsymbol{\beta}^*, \mathbf{x}))$ is bounded with respect to $1 \leq Y_{\mathcal{G}} \leq M$ and $\mathbf{x} \in \mathcal{X}$. Thus, for any $\delta > 0$, there is a fixed number n_2 such that for any $n \geq n_2$,

$$\max_{1 \leq i \leq n} \mathbb{P} \left[\left\| \mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}_*^i) \right\|^2 > \delta^2 C_n \right] \leq \mathbb{P} \left[\max_{Y_{\mathcal{G}} \in \{1,2,\cdots,M\}, \mathbf{x} \in \mathcal{X}} \left\| \mathbf{s}_{Y_{\mathcal{G}}}(\boldsymbol{\xi}(\boldsymbol{\beta}^*, \mathbf{x})) \right\|^2 > \delta^2 C_n \right] = 0.$$

As a result,

$$\begin{aligned} &\max_{1 \leq i \leq n} \left\{ \mathbb{E} \left[\left\| \mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}_*^i) \right\|^2 \mathbf{1}_{\left[\left\| \mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}_*^i) \right\|^2 > \delta^2 / \|\mathbf{Z}_{ni}\|_{\text{op}}^2 \right]} \right] \right\} \\ &\leq \max_{1 \leq i \leq n} \left\{ \mathbb{E} \left[\left\| \mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}_*^i) \right\|^2 \mathbf{1}_{\left[\left\| \mathbf{s}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}_*^i) \right\|^2 > \delta^2 C_n \right]} \right] \right\} \to 0, \end{aligned}$$

as $n \to +\infty$.

In conclusion, the right-hand-side of (4.16) tends to 0 as $n \to +\infty$. □

Lemma 4.3. *Assume that, for any $\delta > 0$,*

$$\max_{\boldsymbol{\beta} \in \mathcal{N}_n(\delta)} \|\mathbf{V}_n(\boldsymbol{\beta}) - \mathbf{I}\| \to_p 0, \text{ as } n \to +\infty. \tag{4.17}$$

Then, under assumptions (i), (ii), and (iii) of Theorem 4.1, there hold all conclusions of Theorem 4.1.

Proof. It's easy to verify that (4.15) is implied by (4.17). Thus Lemma 4.1 holds. The proof of the asymptotic normality follows mainly Theorem 3 of [31], according to Lemma 4.2. □

The next lemma will be used in the proof of Theorem 4.2.

Lemma 4.4. *Assume (4.7) and assume that there exist $\epsilon > 0$ and a random number $n_1 \in \mathbb{N}$ such that for any $n \geq n_1$,*

$$\sigma_{\min}\left[-\mathbf{H}_n(\boldsymbol{\beta})\right] \geq c_2 \sigma_{\max}\mathbf{F}_n, \boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*), \text{ almost surely,} \tag{4.18}$$

with a universal constant $c_2 > 0$. Then there exist a random number $\tilde{n}_0$ and a sequence $\{\hat{\boldsymbol{\beta}}_n\}_{n=1}^{\infty}$ of estimators such that conclusions (i) and (ii) of Theorem 4.2 hold.

The proof of Lemma 4.4 follows mainly the proof of Theorem 2 in [31] since (4.18) is equivalent to the condition $\left(\mathrm{S}^*_{1/2}\right)$ in Section 4 of [31].

4.4.3 Proofs of Theorems and Corollaries

Proof of Theorem 4.1. It suffices to check (4.17) in Lemma 4.3. We decompose $\mathbf{V}_n(\boldsymbol{\beta})$ as

$$\begin{aligned}-\mathbf{F}_n^{-1/2}\mathbf{H}_n(\boldsymbol{\beta})\mathbf{F}_n^{-T/2} = \sum_{i=1}^{n} \mathbf{F}_n^{-1/2} \Big\{ \mathbf{X}^i \left[\mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}) - \mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*) + \mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*)\right](\mathbf{X}^i)^T \\ - \left[\mathbf{R}^i(\boldsymbol{\beta}) - \mathbf{R}^i(\boldsymbol{\beta}^*) + \mathbf{R}^i(\boldsymbol{\beta}^*)\right]\Big\} \mathbf{F}_n^{-T/2}\end{aligned}$$

with $\mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta}, \mathbf{x}) = \mathbf{U}(\boldsymbol{\beta}, \mathbf{x})\mathbb{I}_{Y_{\mathcal{G}}}(\boldsymbol{\xi}(\boldsymbol{\beta}, \mathbf{x}))\mathbf{U}(\boldsymbol{\beta}, \mathbf{x})$ and $\mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}) = \mathbf{C}_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}, \mathbf{x}^i)$. Then we have

$$\mathbf{V}_n(\boldsymbol{\beta}) - \mathbf{I} = \mathbf{F}_n^{-1/2}\left[-\mathbf{H}_n(\boldsymbol{\beta}) - \mathbf{F}_n\right]\mathbf{F}_n^{-T/2} = \mathcal{A}_n(\boldsymbol{\beta}) + \mathcal{B}_n + \mathcal{C}_n(\boldsymbol{\beta}) + \mathcal{D}_n,$$

where

$$\mathcal{A}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{A}^{ni} \left[\mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}) - \mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*) \right] (\mathbf{A}^{ni})^T,$$

$$\mathcal{B}_n = \sum_{i=1}^{n} \mathbf{A}^{ni} \left[\mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*) - \mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}^i) \right] (\mathbf{A}^{ni})^T,$$

$$\begin{aligned} \mathcal{C}_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{F}_n^{-1/2} \left(\mathbf{R}^i(\boldsymbol{\beta}^*) - \mathbf{R}^i(\boldsymbol{\beta}) \right) \mathbf{F}_n^{-T/2} \\ &= \sum_{i=1}^{n} \mathbf{A}^{ni} \left(\mathbf{W}^i(\boldsymbol{\beta}^*) \mathbf{S}_{Y^i_{\mathcal{G}}}(\boldsymbol{\xi}^i_*) - \mathbf{W}^i(\boldsymbol{\beta}) \mathbf{S}_{Y^i_{\mathcal{G}}}(\boldsymbol{\xi}^i) \right) \left(\mathbf{A}^{ni} \right)^T, \end{aligned}$$

$$\mathcal{D}_n = -\sum_{i=1}^{n} \mathbf{F}_n^{-1/2} \mathbf{R}^i(\boldsymbol{\beta}^*) \mathbf{F}_n^{-T/2} = -\sum_{i=1}^{n} \mathbf{A}^{ni} \mathbf{W}^i(\boldsymbol{\beta}^*) \mathbf{S}_{Y^i_{\mathcal{G}}}(\boldsymbol{\xi}^i_*) \left(\mathbf{A}^{ni} \right)^T,$$

with $\mathbf{A}^{ni} = \mathbf{F}_n^{-1/2} \mathbf{X}^i$ and $\mathbf{C}(\boldsymbol{\beta}, \mathbf{x}) = \mathbf{U}(\boldsymbol{\beta}, \mathbf{x}) \mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}(\boldsymbol{\beta}, \mathbf{x})) \mathbf{U}(\boldsymbol{\beta}, \mathbf{x}) = \mathbb{E}_{\boldsymbol{\beta}} \left[\mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta}, \mathbf{x}) \right]$. By (4.11),

$$\begin{aligned} \|\mathcal{A}_n(\boldsymbol{\beta})\| &\leq \left[\max_{1 \leq i \leq n} \left\| \mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}) - \mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*) \right\| \right] \sum_{i=1}^{n} \left\| \mathbf{A}^{ni} \right\|^2 \\ &\leq C \left[\max_{1 \leq i \leq n} \left\| \mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}) - \mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*) \right\| \right]. \end{aligned}$$

For any $\delta > 0$ and $\boldsymbol{\beta} \in \mathcal{N}_n(\delta)$,

$$\|\Delta\boldsymbol{\beta}\|^2 \leq \left\| \mathbf{F}_n^{T/2} \Delta\boldsymbol{\beta} \right\|^2 / (\sigma_{\min} \mathbf{F}_n) \leq \delta^2 / (\sigma_{\min} \mathbf{F}_n) \to 0, n \to +\infty,$$

with $\Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$. Note that

$$\max_{1 \leq i \leq n} \left\| \mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}) - \mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*) \right\| \leq \max_{Y_{\mathcal{G}} \in \{1, \cdots, M\}, \mathbf{x} \in \mathcal{X}} \left\| \mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta}, \mathbf{x}) - \mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta}^*, \mathbf{x}) \right\| =: \mathbf{C}^{\max}(\boldsymbol{\beta}). \tag{4.19}$$

By the continuity of $\mathbb{I}_{Y_{\mathcal{G}}}$, g_s^{-1}, and $(g_s^{-1})'$, and the compactness of $\mathcal{X}$, $\mathbf{C}^{\max}(\boldsymbol{\beta})$ is continuous with respect to $\boldsymbol{\beta}$. Let $\boldsymbol{\beta}^{\mathbf{C}}_{n,\delta} = \arg\max_{\boldsymbol{\beta} \in \mathcal{N}_n(\delta)} \mathbf{C}^{\max}(\boldsymbol{\beta}) \in \mathcal{N}_n(\delta)$. We have $\boldsymbol{\beta}^{\mathbf{C}}_{n,\delta} \to \boldsymbol{\beta}^*$ for any $\delta > 0$ as $n \to +\infty$ and

$$\max_{\boldsymbol{\beta} \in \mathcal{N}_n(\delta)} \|\mathcal{A}_n(\boldsymbol{\beta})\| \leq \mathbf{C}^{\max}(\boldsymbol{\beta}^{\mathbf{C}}_{n,\delta}) \to \mathbf{C}^{\max}(\boldsymbol{\beta}^*) = 0, \text{ as } n \to +\infty, \text{ for any } \delta > 0. \tag{4.20}$$

To bound $\mathcal{B}_n$, since $\mathbb{E}[\mathcal{B}_n] = \mathbf{0}$, it suffices to show that $\mathbb{E}\left[(\mathcal{B}_n)_{st}^2\right] \to 0$, $n \to +\infty$, for any $1 \leq s, t \leq d$, which implies $\mathcal{B}_n \to_p 0$ as $n \to +\infty$, thanks to Chebyshev's inequality. Since

$$(\mathcal{B}_n)_{st} = \sum_{i=1}^{n} \left(\sum_{l=1}^{r} \sum_{k=1}^{r} \mathbf{A}_{sl}^{ni} \left[\mathbf{C}_{Y_{\mathcal{G}}^i}^i(\boldsymbol{\beta}^*) - \mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}^i) \right]_{lk} \mathbf{A}_{tk}^{ni} \right)$$

and $\mathbb{E}\left[\mathbf{C}_{Y_{\mathcal{G}}^i}^i(\boldsymbol{\beta}^*) - \mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}^i) \right] = 0$, We obtain

$$\begin{aligned}
\mathbb{E}\left[(\mathcal{B}_n)_{st}^2\right] &= \sum_{i=1}^{n} \mathbb{E}\left[\left(\sum_{l=1}^{r} \sum_{k=1}^{r} \mathbf{A}_{sl}^{ni} \left[\mathbf{C}_{Y_{\mathcal{G}}^i}^i(\boldsymbol{\beta}^*) - \mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}^i) \right]_{lk} \mathbf{A}_{tk}^{ni} \right)^2 \right] \\
&\leq \sum_{i=1}^{n} \left\| \mathbf{A}^{ni} \right\|^4 \mathbb{E}\left[\left(\sum_{l=1}^{r} \sum_{k=1}^{r} \left| \left[\mathbf{C}_{Y_{\mathcal{G}}^i}^i(\boldsymbol{\beta}^*) - \mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}^i) \right]_{lk} \right| \right)^2 \right] \\
&\leq K \left(\max_{1 \leq i \leq n} \left\{ \left\| \mathbf{A}^{ni} \right\|^2 \right\} \right) \left(\sum_{i=1}^{n} \left\| \mathbf{A}^{ni} \right\|^2 \right), \qquad (4.21)
\end{aligned}$$

where $K = \max_{\mathbf{x} \in \mathcal{X}, Y_{\mathcal{G}} \in \{1,2,\cdots,M\}} \left\{ \left(\sum_{l=1}^{r} \sum_{k=1}^{r} \left| \left[\mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta}^*, \mathbf{x}) - \mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}) \right] \right|_{lk} \right)^2 \right\} < +\infty$, due to the continuity of $\mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta}, \mathbf{x})$ and $\mathbf{C}(\boldsymbol{\beta}, \mathbf{x})$ with respect to $\mathbf{x}$ and the compactness of $\mathcal{X}$. According to (4.11) and (4.13), the right-hand-side of (4.21) tends to 0 and $\mathcal{B}_n \to_p 0$, as $n \to +\infty$.

Thanks to the continuity of $(g_s^{-1})''$ and $\frac{\partial \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi})}{\partial \xi_s}$, $s = 1, 2, \cdots, r$, the proof of $\max_{\boldsymbol{\beta} \in \mathcal{N}_n(\delta)} \|\mathcal{C}_n(\boldsymbol{\beta})\| \to 0$ is similar to the proof of (4.20) and the proof of $\mathcal{D}_n \to_p 0$ is similar to the proof of $\mathcal{B}_n \to_p 0$. □

Proof of Theorem 4.2. The proof of the asymptotic normality follows mainly [31]. Now we are going to check (4.18) in Lemma 4.4. Rewrite $-\mathbf{H}_n(\boldsymbol{\beta}) = \mathbf{F}_n(\boldsymbol{\beta}) + \mathbf{E}_n(\boldsymbol{\beta})$, where

$$\mathbf{E}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \mathbf{X}^i \mathbf{U}^i(\boldsymbol{\beta}) \left[\mathbb{I}_{Y_{\mathcal{G}}^i}(\boldsymbol{\xi}^i) - \mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}^i) \right] \left(\mathbf{X}^i \mathbf{U}^i(\boldsymbol{\beta}) \right)^T - \mathbf{R}^i(\boldsymbol{\beta}) \right\}.$$

If we denote $\sigma_n^{\max} = \sigma_{\max}\mathbf{F}_n$, then we have $\sigma_{\min}\left[-\mathbf{H}_n(\boldsymbol{\beta})\right]/\sigma_n^{\max} \geq \sigma_{\min}\mathbf{F}_n(\boldsymbol{\beta})/\sigma_n^{\max} - \left\|\mathbf{E}_n(\boldsymbol{\beta})\right\|_{\text{op}}/\sigma_n^{\max}$. According to the assumption *(iv)*, $\sigma_{\min}\mathbf{F}_n(\boldsymbol{\beta})/\sigma_n^{\max}$ is bounded below by c for $n \geq n_0$. It's enough to show that $\left\|\mathbf{E}_n(\boldsymbol{\beta})\right\|_{\text{op}}/\sigma_n^{\max}$ can be arbitrarily small for n large enough and $\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*)$ with sufficiently small $\epsilon > 0$. Recall the matrices $\mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta})$ and $\mathbf{C}(\boldsymbol{\beta}, \mathbf{x})$ in the proof of Theorem 4.1. Decompose $\mathbf{E}_n(\boldsymbol{\beta}) = \mathcal{A}'_n(\boldsymbol{\beta}) + \mathcal{B}'_n + \mathcal{C}'_n(\boldsymbol{\beta}) + \mathcal{D}'_n(\boldsymbol{\beta}) + \mathcal{E}'_n$, where

$$\mathcal{A}'_n(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{X}^i \left[\mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}) - \mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*)\right]\left(\mathbf{X}^i\right)^T,$$

$$\mathcal{B}'_n = \sum_{i=1}^n \mathbf{X}^i \left[\mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*) - \mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}^i)\right]\left(\mathbf{X}^i\right)^T,$$

$$\mathcal{C}'_n(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{X}^i \left[\mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}^i) - \mathbf{C}(\boldsymbol{\beta}, \mathbf{x}^i)\right]\left(\mathbf{X}^i\right)^T,$$

$$\mathcal{D}'_n(\boldsymbol{\beta}) = \sum_{i=1}^n \left[\mathbf{R}^i(\boldsymbol{\beta}^*) - \mathbf{R}^i(\boldsymbol{\beta})\right],$$

$$\mathcal{E}'_n = -\sum_{i=1}^n \mathbf{R}^i(\boldsymbol{\beta}^*).$$

For any $\boldsymbol{\lambda} \in \mathbb{R}^d$,

$$\begin{aligned}\boldsymbol{\lambda}^T\mathbf{F}_n\boldsymbol{\lambda} &= \sum_{i=1}^n \boldsymbol{\lambda}^T\mathbf{X}^i\mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x}^i)\left(\mathbf{X}^i\right)^T\boldsymbol{\lambda} \\ &\geq \left(\min_{\mathbf{x}\in\mathcal{X}}\left\{\sigma_{\min}\mathbf{C}(\boldsymbol{\beta}^*, \mathbf{x})\right\}\right)\boldsymbol{\lambda}^T\mathbf{X}_n\boldsymbol{\lambda} =: c'\boldsymbol{\lambda}^T\mathbf{X}_n\boldsymbol{\lambda},\end{aligned}$$

where $\mathbf{X}_n = \sum_{i=1}^n \mathbf{X}^i\left(\mathbf{X}^i\right)^T$ and $c' > 0$ is a constant. Thus

$$\sigma_n^{\max} = \sigma_{\max}\mathbf{F}_n = \max_{\|\boldsymbol{\lambda}\|=1}\boldsymbol{\lambda}^T\mathbf{F}_n\boldsymbol{\lambda} \geq c'\sigma_{\max}\mathbf{X}_n. \tag{4.22}$$

For any $\delta' > 0$, there is $\epsilon > 0$ small enough such that for any $\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*)$ and $\boldsymbol{\lambda} \in \mathbb{R}^d$ with $\|\boldsymbol{\lambda}\| = 1$,

$$\frac{\left|\boldsymbol{\lambda}^T\mathcal{A}'_n(\boldsymbol{\beta})\boldsymbol{\lambda}\right|}{\sigma_n^{\max}} \leq \left(\max_{1\leq Y_{\mathcal{G}}\leq M, \mathbf{x}\in\mathcal{X}}\left\{\left\|\mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta}, \mathbf{x}) - \mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta}^*, \mathbf{x})\right\|_{\text{op}}\right\}\right)\left(\frac{\sigma_{\max}\left[\mathbf{X}_n\right]}{\sigma_n^{\max}}\right) \leq \delta',$$

where the last equality comes from the boundedness of $\sigma_{\max}\mathbf{X}_n/\sigma_n^{\max}$ due to (4.22) and the continuity of $\mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta},\mathbf{x})$ with respect to $\boldsymbol{\beta}$. Thus

$$\max_{\boldsymbol{\beta}\in B_\epsilon(\boldsymbol{\beta}^*)} \left\|\mathcal{A}_n'(\boldsymbol{\beta})\right\|_{\text{op}}/\sigma_n^{\max} = \max_{\boldsymbol{\beta}\in B_\epsilon(\boldsymbol{\beta}^*),\boldsymbol{\lambda}^T\boldsymbol{\lambda}=1} \left|\boldsymbol{\lambda}^T\mathcal{A}_n'(\boldsymbol{\beta})\boldsymbol{\lambda}\right|/\sigma_n^{\max} \leq \delta'.$$

Now we are going to bound $\mathcal{B}_n'$ by the strong law of large numbers. For any $\boldsymbol{\lambda}\in\mathbb{R}^d$ with $\|\boldsymbol{\lambda}\|=1$, there is a constant $C'<+\infty$ such that

$$\begin{aligned}
\text{Var}\left[\boldsymbol{\lambda}^T\mathcal{B}_n'\boldsymbol{\lambda}\right] &= \sum_{i=1}^n \mathbb{E}\left[\left(\left(\tilde{\mathbf{a}}^i\right)^T\left[\mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*)-\mathbf{C}(\boldsymbol{\beta}^*,\mathbf{x}^i)\right]\tilde{\mathbf{a}}^i\right)^2\right] \\
&\leq \sum_{i=1}^n \left\|\tilde{\mathbf{a}}^i\right\|^4 \mathbb{E}\left[\left\|\mathbf{C}^i_{Y^i_{\mathcal{G}}}(\boldsymbol{\beta}^*)-\mathbf{C}(\boldsymbol{\beta}^*,\mathbf{x}^i)\right\|^2_{\text{op}}\right] \\
&\leq \left(\max_{Y_{\mathcal{G}}\in\{1,2,\cdots,M\},\mathbf{x}\in\mathcal{X}}\left\{\left\|\mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta}^*,\mathbf{x})-\mathbf{C}(\boldsymbol{\beta}^*,\mathbf{x})\right\|^2_{\text{op}}\right\}\right) \\
&\times \left(\max_{1\leq i\leq n}\left\|\tilde{\mathbf{a}}^i\right\|^2\right)\left(\sum_{i=1}^n\left\|\tilde{\mathbf{a}}^i\right\|^2\right) \\
&\leq C'\sum_{i=1}^n\left\|\tilde{\mathbf{a}}^i\right\|^2
\end{aligned}$$

with $\tilde{\mathbf{a}}^i=\left(\mathbf{X}^i\right)^T\boldsymbol{\lambda}$, where the last inequality comes from the fact that $\mathbf{C}_{Y_{\mathcal{G}}}(\boldsymbol{\beta}^*,\mathbf{x})$, $\mathbf{C}(\boldsymbol{\beta}^*,\mathbf{x})$, and $\tilde{\mathbf{a}}^i$ are all bounded above since $Y_{\mathcal{G}}$ is finitely supported and $\mathcal{X}$ is compact. Recall that $\mathbf{X}_n=\sum_{i=1}^n\mathbf{X}^i\left(\mathbf{X}^i\right)^T$. Thanks to (4.22), we have

$$\sum_{i=1}^n\left\|\tilde{\mathbf{a}}^i\right\|^2/\sigma_n^{\max} = \left(\boldsymbol{\lambda}^T\mathbf{X}_n\boldsymbol{\lambda}\right)/\sigma_n^{\max} \leq \frac{\sigma_{\max}\left[\mathbf{X}_n\right]}{\sigma_n^{\max}} \leq 1/c'.$$

By the strong law of large numbers (cf., Lemma 2, [81]), $\boldsymbol{\lambda}^T\mathcal{B}_n'\boldsymbol{\lambda}/\sigma_n^{\max}\to 0$ almost surely, as $n\to+\infty$. Since $\mathcal{B}_n'$ is symmetric, we get that each entry of $\mathcal{B}_n'/\sigma_n^{\max}$ converges to 0 as $n\to+\infty$ through suitable choices of $\boldsymbol{\lambda}$. Thus, as $n\to+\infty$, $\left\|\mathcal{B}_n'\right\|_{\text{op}}/\sigma_n^{\max}\to 0$ almost surely.

Similarly, one can obtain that for any $\delta'>0$, there is $\epsilon>0$ small enough such that $\max_{\boldsymbol{\beta}\in B_\epsilon(\boldsymbol{\beta}^*)}\left\|\mathcal{C}_n'(\boldsymbol{\beta})\right\|_{\text{op}}/\sigma_n^{\max}\leq\delta'$, $\max_{\boldsymbol{\beta}\in B_\epsilon(\boldsymbol{\beta}^*)}\left\|\mathcal{D}_n'(\boldsymbol{\beta})\right\|_{\text{op}}/\sigma_n^{\max}\leq\delta'$, and

$\|\mathcal{E}_n'\|_{\text{op}}/\sigma_n^{\max} \to 0$, a.s. , $n \to +\infty$. By taking $\delta' > 0$ small enough, we obtain that there exist a constant $c_2 > 0$ and $\epsilon > 0$ such that $\sigma_{\min}\left[-\mathbf{H}_n(\boldsymbol{\beta})\right]/\sigma_n^{\max} \geq c_2$, for any $\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*)$ and sufficiently large n. □

Proof of Corollary 4.2. It's enough to prove that the assumption *(iv)* of Theorem 4.2 is satisfied under (4.9). In fact, since $(g_s^{-1})' > 0$ and $\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi})$ is strictly positive definite everywhere, by the continuity of g_s^{-1} and $\mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi})$, there exists $\epsilon > 0$ such that

$$c_{\min} = \min_{\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*), \mathbf{x} \in \mathcal{X}} \sigma_{\min}\left[\mathbf{U}(\boldsymbol{\beta}, \mathbf{x}) \mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}(\boldsymbol{\beta}, \mathbf{x})) \mathbf{U}(\boldsymbol{\beta}, \mathbf{x})\right] > 0,$$

and

$$c_{\max} = \max_{\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*), \mathbf{x} \in \mathcal{X}} \sigma_{\max}\left[\mathbf{U}(\boldsymbol{\beta}, \mathbf{x}) \mathbb{I}^{\mathcal{G}}(\boldsymbol{\xi}(\boldsymbol{\beta}, \mathbf{x})) \mathbf{U}(\boldsymbol{\beta}, \mathbf{x})\right] < \infty.$$

Thus, for any $\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*)$ and $n \geq n_1$,

$$\begin{aligned}
&\sigma_{\min}\mathbf{F}_n(\boldsymbol{\beta}) = \min_{\tilde{\boldsymbol{\lambda}}^T\tilde{\boldsymbol{\lambda}}=1} \left\{\tilde{\boldsymbol{\lambda}}^T \mathbf{F}_n(\boldsymbol{\beta}) \tilde{\boldsymbol{\lambda}}\right\} \\
\geq & c_{\min} \min_{\tilde{\boldsymbol{\lambda}}^T\tilde{\boldsymbol{\lambda}}=1} \left\{\sum_{i=1}^n \tilde{\boldsymbol{\lambda}}^T \mathbf{X}^i \left(\mathbf{X}^i\right)^T \tilde{\boldsymbol{\lambda}}\right\} = c_{\min}\sigma_{\min}\left(\sum_{i=1}^n \mathbf{X}^i \left(\mathbf{X}^i\right)^T\right) \\
\geq & c_{\min}\tilde{c}\sigma_{\max}\left(\sum_{i=1}^n \mathbf{X}^i \left(\mathbf{X}^i\right)^T\right) = \frac{c_{\min}\tilde{c}}{c_{\max}} c_{\max} \max_{\boldsymbol{\lambda}^T\boldsymbol{\lambda}=1} \left\{\boldsymbol{\lambda}^T \left(\sum_{i=1}^n \mathbf{X}^i \left(\mathbf{X}^i\right)^T\right) \boldsymbol{\lambda}\right\} \\
\geq & \frac{c_{\min}\tilde{c}}{c_{\max}} \max_{\boldsymbol{\lambda}^T\boldsymbol{\lambda}=1} \left\{\boldsymbol{\lambda}^T \mathbf{F}_n \boldsymbol{\lambda}\right\} = \frac{c_{\min}\tilde{c}}{c_{\max}} \sigma_{\max}\mathbf{F}_n.
\end{aligned}$$

We finish the proof by noting that $\frac{c_{\min}\tilde{c}}{c_{\max}} > 0$. □

4.5 Real Data Simulations

We experiment with the survey data concerning the marijuana use in America with sample size $n = 8478$ from the project MTF (Monitoring the Future) [44]. The code bases on an R package "GRCRegression" from [34]. Each response is the monthly frequency of marijuana use of a respondent. The number of covariates considered here

Table 4.1: Poisson Regression Estimates

	Estimate	95% Confidence Interval
μ		
Intercept	-0.581***	$(-0.668, -0.494)$
Grade10	1.387***	$(1.318, 1.455)$
Grade12	1.986***	$(1.916, 2.056)$
Male	0.433***	$(0.391, 0.475)$
Black	0.037	$(-0.021, 0.095)$
Intact Family	-0.780***	$(-0.824, -0.737)$
Parental Education	-0.395***	$(-0.440, -0.350)$
Metropolitan Areas	0.134***	$(0.082, 0.186)$
McFadden's Adj R^2: 0.121		
AIC: 52000		
BIC: 52060		

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

is 8, including: **Intercept**, grade (**Grade 10** and **Grade 12**), **male** (versus female), **black** (versus non-African American), **Intact family** (versus single- or no-parent family), **parental education** (one of parents has completed college education or not), **metropolitan areas** (the location of the school of a respondent is metropolitan or not). We adopt the grouping scheme [never, 1-2 times, 3-5 times, 6-9 times, 10-19 times, 20-39 times, 40+ times] according to the optimal design theory of grouping schemes [32].

For the Poisson regression model, we use the log link function $g_{\log}$ to estimate the Poisson parameter $\mu > 0$. The estimates are given in Table 4.1. For the ZIP model, we use $g_{\log}$ for the Poisson parameter $\mu > 0$ and the logit link function g_{logit} for the Bernoulli parameter $0 < p < 1$. The results are provided in Table 4.2.

From both Poisson and ZIP models, we can draw conclusions that the monthly marijuana use frequencies of students, from a junior grade, or from intact family, or with college-educated parents, are lower than the frequencies of their opposite parts, significantly, which are consistent with [34], where they studied the lifetime frequencies. The results also show that females use marijuana less frequently than

Table 4.2: Zero-inflated Poisson Regression Estimates

	Estimate	95% Confidence Interval
μ		
Intercept	2.108***	$(2.019, 2.198)$
Grade10	0.456***	$(0.388, 0.524)$
Grade12	0.732***	$(0.661, 0.802)$
Male	0.187***	$(0.142, 0.231)$
Black	-0.055·	$(-0.117, 0.006)$
Intact Family	-0.325***	$(-0.372, -0.279)$
Parental Education	-0.111***	$(-0.159, -0.064)$
Metropolitan Areas	-0.080**	$(-0.136, -0.024)$
p		
Intercept	2.411***	$(2.161, 2.661)$
Grade10	-1.042***	$(-1.215, -0.870)$
Grade12	-1.450***	$(-1.641, -1.260)$
Male	-0.296***	$(-0.426, -0.165)$
Black	-0.121	$(-0.308, 0.065)$
Intact Family	0.580***	$(0.437, 0.723)$
Parental Education	0.428***	$(0.283, 0.574)$
Metropolitan Areas	-0.260**	$(-0.429, -0.091)$
McFadden's Adj R^2: 0.061		
AIC: 19950		
BIC: 20070		

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$.

males. According to the Poisson model, students from metropolitan areas are more likely to use marijuana which conflicts with the corresponding conclusion from the ZIP model. Another conclusion contrasting to the study of lifetime frequencies is that estimates of "black", from both Poisson and ZIP models, are insignificant (versus the null hypothesis), while [34] shows that black students are less likely to use marijuana significantly. As shown by the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), the ZIP model fits better than the Poisson model, which is also in line with [34].

Chapter 5

Conclusions

The thesis investigates several topics in regression learning, including topics in both non-parametric estimation (such as kernel-based learning) and parametric estimation (such as maximum likelihood estimation).

In Chapter 2, the properties of the Kronecker product kernels are investigated. The analysis on the capacity of the reproducing kernel Hilbert spaces corresponding to the Kronecker product kernels is sharp. Based on the Kronecker product kernels, we study a pairwise learning algorithm called Kronecker kernel ridge regression. Both the upper bound and the minimax lower bound of the error of this pairwise learning algorithm are given. The convergence rate of the pairwise learning algorithm is minimax optimal.

In Chapter 3, we propose a sparse empirical feature-based semi-supervised learning algorithm. Sensitive information from the private data is avoided thanks to the summary statistics generated by the raw data the empirical features generated by published unlabeled data. This semi-supervised learning algorithm is a generalization of the linear model with summary statistics to the non-parametric case. This sparse learning algorithm achieves a fast convergence rate.

In Chapter 4, we established the asymptotic theory of the maximum likelihood estimators for grouped and right-censored count data. Grouped and right-censored

count data have been widely used in survey research. Yet the statistical analysis of grouped and right-censored counts is rare in literature. Recently, a novel modified Poisson estimator based on the maximum likelihood estimation has been proposed and proved to have a methodological advantage comparing with classical models on GRC regression. We derive the asymptotic properties of the maximum likelihood estimators of grouped and right-censored counts with divergent information matrices of the first n observations, which is a weaker condition than existing results. The empirical performance of these estimators is investigated with data on marijuana use in America. As further topics on GRC count data, one may study models with random effects.

Bibliography

[1] Diann M Ackard, Jillian K Croll, and Ann Kearney-Cooke. Dieting frequency among college females: Association with disordered eating, body image, and related psychological problems. *Journal of Psychosomatic Research*, 52(3):129 – 136, 2002.

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

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

[4] Ronald L. Akers, Anthony J. La Greca, John Cochran, and Christine Sellers. Social learning theory and alcohol behavior among the elderly. *The Sociological Quarterly*, 30(4):625–638, 1989.

[5] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[6] Francis Bach. Breaking the curse of dimensionality with convex neutral networks. *Journal of Machine Learning Research (JMLR)*, 18:Paper No. 19, 53, 2017.

[7] Jerald G Bachman, Lloyd D Johnston, and Patrick M O'Malley. Explaining the recent decline in cocaine use among young adults: Further evidence that perceived risks and disapproval lead to reduced drug use. *Journal of Health and Social Behavior*, 31(2):173–184, 1990.

[8] Susan L Bailey, Robert L Flewelling, and J Valley Rachal. Predicting continued use of marijuana among adolescents: The relative influence of drug-specific and social context factors. *Journal of Health and Social Behavior*, 33(1):51–65, 1992.

[9] Luca Baldassarre, Lorenzo Rosasco, Annalisa Barla, and Alessandro Verri. Multi-output learning via spectral filtering. *Machine Learning*, 87(3):259–301, 2012.

[10] 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.

[11] Andrea L. Bertozzi, Xiyang Luo, Andrew M. Stuart, and Konstantinos C. Zygalakis. Uncertainty quantification in graph-based classification of high dimensional data. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):568–595, 2018.

[12] Edward Blair and Scot Burton. Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research*, 14(2):280–288, 1987.

[13] 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.

[14] Kurt Brännäs. Limited dependent Poisson regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(4):413–423, 1992.

[15] A. Colin Cameron and Pravin K. Trivedi. *Regression analysis of count data*, volume 53 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, second edition, 2013.

[16] Qiong Cao, Zheng-Chu Guo, and Yiming Ying. Generalization bounds for metric and similarity learning. *Machine Learning*, 102(1):115–132, 2016.

[17] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics. The Journal of the Society for the Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[18] Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *Journal of Machine Learning Research (JMLR)*, 18:Paper No. 46, 22, 2017.

[19] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research (JMLR)*, 11:1109–1135, 2010.

[20] Di-Rong Chen, Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research (JMLR)*, 5:1143–1175, 2003/04.

[21] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[22] Andreas Christmann and Ding-Xuan Zhou. On the robustness of regularized pairwise learning methods based on kernels. *Journal of Complexity*, 37:1–33, 2016.

[23] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

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

[25] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *American Mathematical Society. Bulletin. New Series*, 39(1):1–49, 2002.

[26] Felipe Cucker and Ding-Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007.

[27] Ben Dai, Xiaotong Shen, Junhui Wang, and Annie Qu. Scalable collaborative ranking for personalized prediction. *Journal of the American Statistical Association*, 0(0):1–9, 2020.

[28] Lee H. Dicker, Dean P. Foster, and Daniel Hsu. Kernel ridge vs. principal component regression: minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, 11(1):1022–1047, 2017.

[29] John Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275, 2018.

[30] John C. Duchi, Lester Mackey, and Michael I. Jordan. The asymptotics of ranking algorithms. *The Annals of Statistics*, 41(5):2292–2323, 2013.

[31] Ludwig Fahrmeir and Heinz Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

[32] Qiang Fu, Xin Guo, and Kenneth C Land. Optimizing count responses in surveys: A machine-learning approach. *Sociological Methods & Research*, 2017. `https://doi.org/10.1177/0049124117747302`.

[33] Qiang Fu, Xin Guo, and Kenneth C. Land. A Poisson-multinomial mixture approach to grouped and right-censored counts. *Communications in Statistics - Theory and Methods*, 47(2):427–447, 2018.

[34] Qiang Fu, Tianyi Zhou, and Xin Guo. Modified Poisson regression analysis of grouped and right-censored counts. *Journal of the Royal Statistical Society: Series A*, forthcoming.

[35] Xin Guo, Jun Fan, and Ding-Xuan Zhou. Sparsity and error analysis of empirical feature-based regularization schemes. *Journal of Machine Learning Research (JMLR)*, 17:Paper No. 89, 34, 2016.

[36] Xin Guo, Ting Hu, and Qiang Wu. Distributed minimum error entropy algorithms. *Journal of Machine Learning Research (JMLR)*, 21:Paper No. 126, 31, 2020.

[37] Xin Guo, Ting Hu, and Qiang Wu. Centered reproducing kernel for variable and interaction selection. *Manuscript in Preparation*, 2021.

[38] Xin Guo and Ding-Xuan Zhou. An empirical feature-based learning algorithm producing sparse approximations. *Applied and Computational Harmonic Analysis. Time-Frequency and Time-Scale Analysis, Wavelets, Numerical Algorithms, and Applications*, 32(3):389–400, 2012.

[39] Zheng-Chu Guo, Lei Shi, and Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *Journal of Machine Learning Research (JMLR)*, 18:Paper No. 118, 25, 2017.

[40] Daniel B. Hall. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4):1030–1039, 2000.

[41] Roger A. Horn and Charles R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1994.

[42] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[43] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[44] Lloyd D. Johnston, Patrick M. O'Malley, Richard A. Miech, Jerald G. Bachman, and John E. Schulenberg. Monitoring the future national survey results on drug use, 1975-2016: Overview, key findings on adolescent drug use. `https://files.eric.ed.gov/fulltext/ED578534.pdf`, 2017. accessed July 17, 2019.

[45] Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

[46] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[47] E. L. Lehmann and George Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998.

[48] Guanghui Li, Jiawei Luo, Qiu Xiao, Cheng Liang, and Pingjian Ding. Prediction of microrna-disease associations with a kronecker kernel matrix dimension reduction model. *RSC Adv.*, 8:4377–4385, 2018.

[49] Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research (JMLR)*, 18:Paper No. 92, 31, 2017.

[50] J. Liu, C. Yang, Jiao Y., and Jian Huang. sslasso: A summary-statistic-based regression using lasso. *preprint*, 2017.

[51] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[52] Andreas Maurer. Learning similarity with operator-valued large-margin classifiers. *Journal of Machine Learning Research (JMLR)*, 9:1049–1082, 2008.

[53] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall, 1989.

[54] James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philos. Trans. Roy. Soc. London Ser. A*, 209(441-458):415–446, 1909.

[55] Yongyi Min and Alan Agresti. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5(1):1–19, 2005.

[56] Anthea Monod. Random effects modeling and the zero-inflated Poisson distribution. *Communications in Statistics - Theory and Methods*, 43(4):664–680, 2014.

[57] Van Trinh Nguyen and Jean-François Dupuy. Asymptotic results in censored zero-inflated Poisson regression. *Communications in Statistics - Theory and Methods*, 2019. `https://doi.org/10.1080/03610926.2019.1676442`.

[58] Tapio Pahikkala, Antti Airola, Michiel Stock, Bernard De Baets, and Willem Waegeman. Efficient regularized least-squares algorithms for conditional ranking on relational data. *Machine Learning*, 93(2-3):321–356, 2013.

[59] Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

[60] Tomaso Poggio and Steve Smale. The mathematics of learning: dealing with data. *Notices of the American Mathematical Society*, 50(5):537–544, 2003.

[61] Huihui Qin and Xin Guo. Semi-supervised learning with summary statistics. *Analysis and Applications*, 17(5):837–851, 2019.

[62] Rafal Raciborski. Right-censored Poisson regression model. *The Stata Journal*, 11(1):95–105, 2011.

[63] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[64] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[65] Debajyoti Sinha, Martin A. Tanner, and W. J. Hall. Maximization of the marginal likelihood of grouped survival data. *Biometrika*, 81(1):53–60, 1994.

[66] Steve Smale and Ding-Xuan Zhou. Shannon sampling. II. Connections to learning theory. *Applied and Computational Harmonic Analysis. Time-Frequency and Time-Scale Analysis, Wavelets, Numerical Algorithms, and Applications*, 19(3):285–302, 2005.

[67] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation. An International Journal for Approximations and Expansions*, 26(2):153–172, 2007.

[68] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research (JMLR)*, 2(1):67–93, 2002.

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

[70] Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

[71] Michiel Stock, Tapio Pahikkala, Antti Airola, Bernard De Baets, and Willem Waegeman. A comparative study of pairwise learning methods based on kernel ridge regression. *Neural Computation*, 30(8):2245–2283, 2018.

[72] Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research (JMLR)*, 18:Paper No. 233, 29, 2017.

[73] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[74] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[75] Aad W. Van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge university press, 2000.

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

[77] Jean-Philippe Vert, Jian Qiu, and William S Noble. A new pairwise kernel for biological network inference with support vector machines. In *BMC bioinformatics*, volume 8, page S8. BioMed Central, 2007.

[78] Willem Waegeman, Tapio Pahikkala, Antti Airola, Tapio Salakoski, Michiel Stock, and Bernard De Baets. A kernel-based framework for learning graded relations from data. *IEEE Transactions on Fuzzy Systems*, 20(6):1090–1101, 2012.

[79] Grace Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[80] Chendi Wang, Xin Guo, and Qiang Wu. Learning with centered reproducing kernels. *submitted*, 2021.

[81] Chien-Fu Wu. Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, 9(3):501–513, 1981.

[82] Yun Yang, Mert Pilanci, and Martin J. Wainwright. Randomized sketches for kernels: fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017.

[83] Yiming Ying and Peng Li. Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research*, 13(1):1–26, 2012.

[84] Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

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

[86] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

[87] Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

[88] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research (JMLR)*, 16:3299–3340, 2015.

[89] Peilin Zhao, Steven CH Hoi, Rong Jin, and Tianbo YANG. Online auc maximization. 2011.

[90] Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The Annals of Applied Statistics*, 11(3):1561–1592, 2017.

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

[92] Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2006.

[93] Laurent Zwald, Gilles Blanchard, Pascal Massart, and Régis Vert. Kernel projection machine: a new tool for pattern recognition. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005.