Statistical Machine Learning for Complex Data Sets

by

Xiaowu Dai

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2019

Date of final oral examination: 05/01/2019

The dissertation is approved by the following members of the Final Oral Committee:

Grace Wahba, Advisor, IJ Schoenberg-Hildale Professor, Statistics

Peter Chien, Advisor, Professor, Statistics

Zhengjun Zhang, Professor, Statistics

Dimitris Papailiopoulos, Assistant Professor, Electrical and Computer Engineering

Anru Zhang, Assistant Professor, Statistics

To my wife Yuhua Zhu, and parents Qichun Dai and Liuhua Tu

Acknowledgments

I want to express my deepest gratitude to my advisors, Professor Grace Wahba and Professor Peter Chien. Grace's passion for pursuing science was greatly influential in my development as a researcher. Grace is my role model. It is really an honor and a privilege to work with Grace. Peter's enthusiasm for statistics and computing has substantially influenced my research direction and interests. It has always been a source of inspiration to discuss with Peter. This thesis would not have been possible without the patient mentorship and encouragement from Grace and Peter.

I am grateful to my thesis committee members: Professors Zhengjun Zhang, Dimitris Papailiopoulos, and Anru Zhang for their valuable inputs and suggestions. I want to thank Professor Zhengjun Zhang for supporting me and writing letters in behalf of me. Thank Professor Dimitris Papailiopoulos for introducing me into deep learning which is such an exciting area to work on. Thank Professor Anru Zhang for bringing the semi-supervised learning to my research plan. Collaborations with Jared Huling has also been a great motivation for the research presented in this thesis and beyond.

I am fortunate to join two dynamic and inspiring research group meetings: "Thursday group" and Peter's group. I want to thank all enlightening lectures and challenging questions in "Thursday group" meetings from Professors Ming Yuan, Anru Zhang, Garvesh Raskutti, Karl Rohe, Sijian Wang. I want to thank all former and current graduate students in "Thursday group" and Peter's group for their presentations and questions. I am grateful to Professor Kam-Wah Tsui and Dr. Derek Bean for coaching me for teaching skills. Thank Professors who have contributed in my Ph.D. study: Jun Zhu, Yazhen Wang, Chunming

Zhang, Vikas Singh, Po-Ling Loh, Richard Chappell, Sunduz Keles, Michael Newton.

I am grateful to Professor Shi Jin and Professor Yaguang Wang for their support in Madison and Shanghai. I would like to acknowledge my friends in Madison, and fellow graduate students in the Department of Statistics, too many to name, for making my life in Madison to be filled with the joy of talking, swimming, hiking, biking, camping, and running marathon.

I owe so much to my parents, parents-in-law, and sisters for their unconditional love and support. I dedicate this thesis to Yuhua for making my life much better in all ways.

Contents

List of Tables

List of Figures

Abstract

This thesis is focused on developing theory and computational methods for a set of problems involving complex data.

Chapter 2 studies multivariate nonparametric predictions with gradient information. Gradients can be easily estimated in stochastic simulations and computer experiments. We propose a unified framework to incorporate the noisy and correlated gradients into predictions. We show theoretically, through minimax optimal rates of convergence, that incorporating gradients tends to significantly improve predictions with deterministic or random designs.

Chapters 3 proposes high-dimensional smoothing splines with applications to Alzheimer's disease (AD) prediction. While traditional prediction based on structural MRI uses imaging acquired at a single time point, a longitudinal study is more sensitive in detecting early pathological changes of the AD. Our novel method can be applied to extract features from heterogeneous and longitudinal MRI for the AD prediction, outperforming existing methods.

Chapters 4 introduces a novel class of variable selection penalties called TWIN, which provides sensible data-adaptive penalization. Under a linear sparsity regime, we show that TWIN penalties have a high probability of selecting correct models and result in minimax optimal estimators. We demonstrate in challenging and realistic simulation settings with high correlations between active and inactive variables that TWIN has high power in variable selection while controlling the number of false discoveries, outperforming standard penalties.

Chapters 5 investigates generalizations of mini-batch SGD in deep neural networks. We theoretically justify a hypothesis that large-batch SGD tends to converge to sharp minimizers by providing new properties of SGD. In particular, we give an explicit escaping time of SGD from a local minimum in the finite-time regime and prove that SGD tends to converge to flatter minima in the asymptotic regime (although may take exponential time to converge) regardless of the batch size.

Chapter 6 provides another look at statistical calibration problems in computer models. This viewpoint is inspired by two overarching practical considerations: (i) Many computer models are inadequate for perfectly modeling physical systems; (ii) Only a finite number of data are available from physical experiments to calibrate related computer models. We provide a non-asymptotic theory and derive a novel prediction-oriented calibration method.

Chapter 1

Introduction

The main focus of this dissertation is to develop theory and computational methods for a set of problems involving complex data: (i) Statistical machine learning for "big data", where data are heterogeneous, high-dimensional, and high-volume, and (ii) Uncertainty quantification for model errors in the non-asymptotic regime.

1.1 Statistical Machine Learning for Complex Data Sets

First, the heterogeneous and complex nature of data is increasingly collectable in the era of big data. As an example, derivative observations are available in many applications. Economists estimate cost functions, where data on factor demands and costs are collected together, and the demand functions are partial derivatives of the cost function by Shephard's Lemma. In dynamic systems and traffic engineering, real-time motion sensors can record velocity, acceleration in addition to position. To date, the fundamental question of how much benefit can be gained by incorporating noisy derivative data into function estimation and prediction has not been answered satisfactorily.

Chapter 2 aims to propose new nonparametric methods to incorporate derivatives for estimation and show that incorporation of first-order partial derivatives can adequately improve minimax optimal rates. In particular, the general multivariate nonparametric

functional ANOVA models can be estimated as efficiently as *additive models* by incorporating first-order partial derivatives.

Second, data heterogeneity is common in many existing datasets. A motivating example for my work is a longitudinal study where brain magnetic resonance imaging (MRI) is used to detect early pathological changes of Alzheimer's disease (AD). Two main difficulties arise here: (i) the longitudinal scans are collected in a highly inconsistent manner across and within subjects; (ii) the regions of interest (ROIs) in brain MRI is of a large amount and atrophy at heterogeneous rates (e.g.,the atrophy rate of entorhinal cortex is significantly higher than that of hippocampus). Chapters 3 provides a statistical modeling solution to simultaneously consider these two sources of heterogeneity.

Chapter 3 utilizes varying-coeffficient models (Hastie and Tibshirani, 1993) to capture these nonlinear relations and to model the heterogeneous atrophies of ROIs, which is motivated by the fact that functional relations between atrophies of AD-related ROIs and changes in clinical cognition are nonlinear in time (Jack Jr et al., 2010). In order to identify important AD-related ROIs from the plethora of possible ROIs in brain MRI data, we proposes a novel feature selection method for nonparametric varying-coefficient models. Our idea is to combine the smoothing splines and an l_1-penalty in the penalized likelihood framework, which can simultaneously select AD-related ROIs and estimate their smooth heterogenous progressions. Our method is robust to the inconsistency among longitudinal scans and can be applied to general longitudinal studies with heterogeneous data structures. We introduce a computationally efficient algorithm to implement the proposed method.

Third, discovering linear relationships between high-dimensional covariates and an outcome remains a challenging problem when a significant fraction of covariates is important in predicting a response. Considering examples of human biology, it is sensible that more relevant predictors may be included when an increasing amount of genetic or microbiome information is leveraged, especially for gene-gene, gene-environment, or microbiome-environment interactions. In this setting it is crucial to provide variable selection methods which are able to yield high power in variable selection while controlling the

number of false discoveries.

Chapter 4 address this problem with a novel class of penalties where larger coefficients are subjected to attenuated penalization. The proposed penalty class results in estimators that are selection consistent and asymptotically minimax in high-dimensional scenarios under a linear sparsity regime. We show theoretically and through extensive simulations that our method gives higher power while controlling FDR under the cases of strong correlations and weak signals, compared with standard penalties.

Fourth, big data is marked by its massive size. To economize the computational cost, the stochastic gradient descent (SGD) method is almost ubiquitously used for optimization tasks, including the training of deep neural networks (DNNs). Standard gradient descent proceeds iteratively via the gradient of the objective function, while SGD adopts an unbiased but variable estimate of the true gradient. The stochasticity in SGD is proportional to the ratio of the step size and the batch size of samples used in gradient estimations and the effect of the batch size on generalization performance remains an elusive but critical problem. The understanding how geometry and generalization performance of models trained by SGD relate with the batch size of SGD is limited in the literature. Recently, a hypothesis by Keskar et al. (2016) that "*large batch SGD tends to converge to sharp minimizers of the training function*" has received increasing attentions.

Chapter 5 provides a theoretical justification to this conjecture, with the tools from empirical process theory and nonlinear partial differential equations. As part of my ongoing effort, I am working to extend the current work to explain the generalization mystery that large batch SGD tends to generalize less well on unseen data.

1.2 Uncertainty Quantification and A Non-Asymptotic Theory

Computer models constructed on partial differential equations and other mathematical physics tools are increasingly used to facilitate the study of complex systems. As George Box famously stated "*All models are wrong, but essentially some are useful*" – even the best

computer models are only approximations of reality and the model errors always exist. Optimal *predictions* for real systems are only possible by combining the information from expensive data and the insights from the complex but imperfect structure of computer models.

Chapter 6 proposes a new method for *quantifying uncertainties in computer models* by following this line of thinking. The uncertainties of computer models come from model errors and unknown calibration parameters that cannot be directly measured. As an example of a calibration parameter, the soil permeability in underground water simulations is important but its true value is rarely known. We propose to identify calibration parameters by minimizing the distance between computer models and collected data in the reproducing kernel Hilbert space (RKHS) norm. We provide justification of the use of RKHS norm as opposed to the L_2 norm, as it not only incorporates L_2-distance information, but also sensitivity information. Theoretically, our calibration method is shown to give the minimal predictive mean squared error for any *finite sample* with statistical guarantees. This result is based on a novel sharp bound for nonparametric estimation error in the finite-sample regime. We introduce an algorithm to carry out the proposed calibration method. Beyond calibration of computer models, our method can be applied to calibrate unknown parameters for misspecified models in statistics and engineering.

The dissertation is concluded by a few remarks on future works in Chapter 7.

Chapter 2

Minimax Optimal Rates of Estimation in Functional ANOVA Models with Derivatives

2.1 Introduction

Derivative observations for complex systems are available in many applications. In dynamic systems and traffic engineerings, real-time motion sensors can record velocity, acceleration in addition to positions. Economists estimate cost functions, where data on factor demands and costs are collected together, and the demand functions are partial derivatives of the cost function by Shephard's Lemma (Hall and Yatchew, 2007, 2010). In actuarial science, mortality force data can be obtained from demography, which together with samples for the survival distribution can yield derivatives for the survival distribution function. In computer experiments, partial derivatives are available by using differentiation mechanisms at little additional cost.

We consider the problem of nonparametric regression with data from the function itself and its first-order partial derivatives. Let $\partial f(\mathbf{t})/\partial t_j$ denote the jth first-order partial derivative of a scalar function $f(\mathbf{t})$ of d covariates $\mathbf{t} = (t_1, \ldots, t_d)$. Consider a multivariate

regression model

$$\begin{cases} Y^{(0)} = f_0(\mathbf{t}^{(0)}) + \epsilon^{(0)}, \\ Y^{(j)} = \partial f_0 / \partial t_j(\mathbf{t}^{(j)}) + \epsilon^{(j)}, \quad 1 \le j \le p. \end{cases} \tag{2.1}$$

Here, $Y^{(0)}$ is the observation of the function under design $\mathbf{t}^{(0)}$ and $Y^{(j)}$ is the observation of the jth first-order partial derivative under design $\mathbf{t}^{(j)}$. Suppose that $\mathbf{t}^{(0)}, \mathbf{t}^{(j)}s$ are supported on $\mathcal{X}_1^d$ with $\mathcal{X}_1 = [0,1]$. Assume the random errors $\epsilon^{(0)}$ and $\epsilon^{(j)}$s are uncentered and correlated. Let $p \in \{1, \ldots, d\}$ denote the number of different types of first-order partial derivatives available. Without loss of generality, we focus on the first p covariates in (2.1). Assume that $\{(\mathbf{t}_i^{(j)}, y_i^{(j)}) : i = 1, \ldots, n\}$ are copies of $(\mathbf{t}^{(j)}, Y^{(j)})$ for $j = 0, 1, \ldots, p$.

We use the smoothing spline analysis of variance (SS-ANOVA) (Wahba, 1990) for modeling $f_0(\mathbf{t})$ which assumes a tensor product structure and smoothness properties on lower dimensions. This framework is desirable for many applications with derivative data. For illustration, consider cost function estimation in economics (Hall and Yatchew, 2007). Write the cost as $f_0(t_1, \ldots, t_d)$, with t_d being the level of output and $(t_1, \ldots, t_{d-1})$ the prices of the $d-1$ factor inputs. The first order partial derivatives of f_0 with respect to input prices is the quantities of factor inputs, which are available together with the cost itself. The Cobb-Douglas production function (Varian, 1992) yields that $f_0(t_1, \ldots, t_d) = [c_0^{-1/c} \prod_{j=1}^{d-1} (c/c_j)^{c_j/c}] \prod_{j=1}^{d-1} t_j^{c_j/c} t_d^{1/c}$, which is an SS-ANOVA function, where c_0 is the efficiency parameter, $c_1, \ldots, c_{d-1}$ are elasticity parameters, and $c = c_1 + \cdots + c_{d-1}$.

2.1.1 Existing Work and Our Contributions

Our work is related to the pioneering work of Hall and Yatchew (2007) which proposed kernel estimators to incorporate derivative data and established improved rates of convergence. Their method replaces local averages with nonlocal averages from partial derivatives. Provided that data on sufficient mixed higher-order partial derivatives are available, local averaging can be avoided and the root-n consistency can be achieved. Since obtaining higher-order derivatives can be difficult in practice, this work focuses on data from first-

order partial derivatives and under a relaxed error structure. Hall and Yatchew (2010) consider series-type estimators to incorporate derivatives under deterministic designs. Main differences between Hall and Yatchew (2007, 2010) and ours are as follows:

- Function space. We consider SS-ANOVA functions that have a tensor product structure, which is not explored in Hall and Yatchew (2007, 2010). The tensor product structure in our model can improve the convergence rate exponentially with p types of first-order partial derivatives as in (2.5) and (2.7). For $p = d - 1$ in (2.1), we achieve the same rate as additive models. Our simulations in the Supplemental Materials corroborate this improvement of convergence rates. For situations where the true function cannot be well modeled by tensor product functions with squared approximation error $O(n^{-1})$, the first-order partial derivatives only cannot substantially improve the estimation error and use higher-order derivatives are necessary as shown in Hall and Yatchew (2007).

- Estimation approach. We propose a new estimator in the RKHS to incorporate first-order partial derivatives. See Theorems 2.2 and 2.4 for its minimax optimality under both deterministic and random designs. Its easy interpretability for estimation in SS-ANOVA provides a direct description of interactions (Wahba et al., 1995). In remarks after (2.11) and (2.12), we observe that the first-order partial derivatives have an effect on reducing interactions of a SS-ANOVA function in terms of the optimal rates. Since the first derivatives help achieve the root-n consistency in univariate estimation, the tensor product structure of SS-ANOVA allows the components with partial derivative data to be estimated with the root-n consistency and reduce the interactions.

- Error structure. Our approach broadens the i.i.d. error structure in Hall and Yatchew (2007, 2010) to allow the random errors to have certain bias and correlation. This relaxed assumption is in line with applications where derivatives are estimated from function observations.

The rest of the article is organized as follows. We provide additional notation and a summary of main results in Section 2.2. We give the main results on estimating functions with deterministic designs in Section 2.3 and random designs in Section 2.4. We consider the optimal rates of estimating first-order partial derivatives in Section 2.5. We describe results of a real example in Section 2.6. Another real application and extensive simulations together with all proofs are relegated to the Appendix.

2.2 Notation and Summary of Main Results

2.2.1 SS-ANOVA and Error Structure

The SS-ANOVA model has the following form:

$$f_0(\mathbf{t}) = \text{constant} + \sum_{k=1}^{d} f_{0k}(t_k) + \sum_{k<j} f_{0kj}(t_k, t_j) + \cdots , \tag{2.2}$$

where the f_{0k}s are the main effects, the f_{0kj}s are the two-way interactions, and so on. Components on the right hand side satisfy side conditions to assure identifiability. The series is truncated to some order r of interactions to enhance interpretability. Here, $f_0(t)$ is a *full* or *truncated* interaction SS-ANOVA model if $r = d$ or $1 \leq r < d$, respectively. We assume that $f_0 \in \mathcal{H}$, where $\mathcal{H}$ is a RKHS corresponding to the decomposition (2.2). Let $\mathcal{H}^k$ be a function space of functions of t_k over $\mathcal{X}_1$ such that $\int_{\mathcal{X}_1} f_{0k}(t_k)dt_k = 0$ for any $f_{0k}(t_k) \in \mathcal{H}^k$, and $\{1\}$ be the space of constant functions. Construct the tensor product space $\mathcal{H}$ as

$$\begin{aligned} \mathcal{H} &= \bigotimes_{k=1}^{d} \left[\{1\} \oplus \mathcal{H}^k\right] \\ &= \{1\} \oplus \sum_{k=1}^{d} \mathcal{H}^k \oplus \sum_{k<j} \left[\mathcal{H}^k \otimes \mathcal{H}^j\right] \oplus \cdots , \end{aligned} \tag{2.3}$$

where the second equality is the expansion of the tensor product. The components of the SS-ANOVA decomposition (2.2) are in the mutually orthogonal subspaces of $\mathcal{H}$ in (2.3). We further assume that all component functions come from a common RKHS $(\mathcal{H}_1, \|\cdot\|_{\mathcal{H}_1})$

given by $\mathcal{H}^k \equiv \mathcal{H}_1$ for $k = 1, \ldots, d$. Let $K : \mathcal{X}_1 \times \mathcal{X}_1 \mapsto \mathbb{R}$ be a Mercer kernel generating the RKHS $\mathcal{H}_1$ and write $K_d\left((t_1, \ldots, t_d)^\top, (t'_1, \ldots, t'_d)^\top\right) = K(t_1, t'_1) \cdots K(t_d, t'_d)$. Then K_d is the reproducing kernel of RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ (Aronszajn, 1950).

Suppose the random errors $\epsilon^{(0)}$ and $\epsilon^{(j)}$s in (2.1) satisfy

$$\begin{aligned} &\mathbb{E}[\epsilon_i^{(j)}] = o(n^{-1/2}), \;\; \text{Var}[\epsilon_i^{(j)}] = \sigma_j^2 < \infty, \\ &\text{Cov}[\epsilon_i^{(j)}, \epsilon_{i'}^{(k)}] = O\left(|i - i'|^{-\Upsilon}\right) \;\; \text{for some } \Upsilon > 1, \end{aligned} \tag{2.4}$$

where $i \neq i'$ and $j, k = 0, 1, \ldots, p$. Random errors in derivative data can be uncentered and correlated. The short-range correlations is assumed in (2.4) for some $\Upsilon > 1$ since partial derivatives are usually calculated using local function data. The error structure (2.4) is reasonable when derivatives are estimated by methods like the infinitesimal perturbation analysis (Glasserman, 2013).

2.2.2 Deterministic Design

We derive the minimax optimal convergence rates for estimating $f_0(\cdot)$ and its partial derivatives $\partial f_0 / \partial t_j(\cdot)$. First consider regular lattices, or called tensor product designs. Suppose that the eigenvalues of the K decay polynomially with the νth largest eigenvalue of the order ν^{-2m}. In Section 2.3, we show that the minimax optimal rate for estimating a full interaction $(r = d)$ SS-ANOVA model $f_0 \in \mathcal{H}$ is

$$\begin{aligned} &\infty_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{E} \int_{\mathcal{X}_1^d} \left[\tilde{f}(\mathbf{t}) - f_0(\mathbf{t})\right]^2 d\mathbf{t} \\ &\qquad = \begin{cases} C_1 \left[n(\log n)^{1+p-d}\right]^{-2m/(2m+1)} & \text{if } 0 \leq p < d, \\ C_2 \left\{n^{-1}(\log n)^{d-1} + n^{-2md/[(2m+1)d-2]}\right\} & \text{if } p = d, \end{cases} \end{aligned} \tag{2.5}$$

where the infimum is taken over all measurable estimators, and C_1, C_2 are constants not depending on n. If $0 \leq p < d$, the above rate is the minimax optimal rate for estimating a $(d - p)$ dimensional full interaction SS-ANOVA model with only function observations (Gu,

2013; Lin, 2000). If $p = d$ and $d \geq 3$, the minimax optimal rate in (2.5) becomes

$$\infty_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{E} \int_{\mathcal{X}_1^d} \left[\tilde{f}(\mathbf{t}) - f_0(\mathbf{t}) \right]^2 d\mathbf{t} = C_2 n^{-2md/[(2m+1)d-2]}. \tag{2.6}$$

The rate in (2.6) converges *faster* than the optimal rate $n^{-2m/(2m+1)}$ for additive models given in Hastie and Tibshirani (1990); Stone (1985). If $p = d$ and $d = 2$, the minimax optimal rate in (2.5) is $n^{-1} \log n$. If $p = d$ and $d = 1$, the root-n consistency is achieved in (2.5) and this phenomenon is observed in Hall and Yatchew (2007, 2010).

The results for truncated interaction SS-ANOVA models ($r < d$) with derivatives will be given in Section 2.3. In particular, for the additive model $r = 1$ and $p = d$, the minimax optimal rate is n^{-1}, which coincides with the *parametric* convergence rate.

2.2.3 Random Design

We are interested in obtaining sharp results for random designs. Suppose that design points $\mathbf{t}^{(0)}$ and $\mathbf{t}^{(j)}$s are independently drawn from distributions $\Pi^{(0)}$ and $\Pi^{(j)}$s with support on $\mathcal{X}_1^d$. In Section 2.4, we show that the minimax optimal rate for estimating a full interaction ($r = d$) SS-ANOVA model $f_0 \in \mathcal{H}$ is

$$\begin{aligned} \infty_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{P} \Bigg\{ \int_{\mathcal{X}_1^d} \left[\tilde{f}(\mathbf{t}) - f_0(\mathbf{t}) \right]^2 d\mathbf{t} \geq C_3 \Bigg(\left[n (\log n)^{1+p-d} \right]^{-2m/(2m+1)} 1_{0 \leq p < d} \\ + \left[n^{-1} (\log n)^{d-1} + n^{-2md/[(2m+1)d-2]} \right] 1_{p=d} \Bigg) \Bigg\} = 0, \end{aligned} \tag{2.7}$$

where C_3 is a constant scalar not depending on n. Results for truncated interaction ($r < d$) SS-ANOVA models will be given in Section 2.4. In addition, the minimax optimal rates are obtained for estimating $\partial f_0 / \partial t_j(\cdot)$ for $j \in \{1, \ldots, p\}$ and both full and truncated SS-ANOVA models. These rates are

$$\infty_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{P} \left\{ \int_{\mathcal{X}_1^d} \left[\tilde{f}(\mathbf{t}) - \partial f_0 / \partial t_j(\mathbf{t}) \right]^2 d\mathbf{t} \geq C_4 n^{-2(m-1)/(2m-1)} \right\} > 0, \tag{2.8}$$

where C_4 does not depend on n. This result holds regardless of the values of d and r. The rate is the same as the optimal rate for estimating $\partial f_0/\partial t_j(\cdot)$ if f_0 comes from a univariate function space $\mathcal{H}_1$ (Stone, 1980, 1982) instead of the d-variate function space $\mathcal{H}$.

2.3 Minimax Risks with Deterministic Designs

This section provides the minimax optimal rates of estimating $f_0(\cdot)$ with model (2.1) and regular lattices. A regular lattice of size $n = l_1 \times \cdots \times l_d$ on $\mathcal{X}_1^d$ is a collection of design points $\{\mathbf{t}_1, \ldots, \mathbf{t}_n\} = \{(t_{i_1,1}, t_{i_2,2}, \ldots, t_{i_d,d}) \mid i_k = 1, \ldots, l_k, k = 1, \ldots, d\}$, where $t_{j,k} = j/l_k$, $j = 1, \ldots, l_k, k = 1, \ldots, d$. This design is often used in statistics when the true function f_0 is a functional ANOVA model. Under the regular lattice design, it is reasonable to assume $f_0 : \mathcal{X}_1^d \mapsto \mathbb{R}$ to have a periodic boundary condition as any finite-length sequence $\{f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n)\}$ can be associated with a periodic sequence

$$
\begin{aligned}
& f^{\text{per}}(i_1/l_1, \cdots, i_d/l_d) \\
&= \sum_{q_1=-\infty}^{\infty} \cdots \sum_{q_d=-\infty}^{\infty} f(i_1/l_1 - q_1, \cdots, i_d/l_d - q_d), \quad \forall (i_1, \ldots, i_d) \in \mathbb{Z}^d
\end{aligned}
$$

by letting $f(\cdot) \equiv 0$ outside $\mathcal{X}_1^d$ and at the unobserved boundaries of $\mathcal{X}_1^d$. On the other hand, any finite-length sequence $\{f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n)\}$ can be recovered from the periodic sequence $f^{\text{per}}(\cdot)$.

Recall that K is the reproducing kernel for component RKHS $\mathcal{H}_1$, which is a symmetric positive semi-definite, square integrable function on $\mathcal{X}_1 \times \mathcal{X}_1$. We require an additional differentiability condition on kernel K:

$$
\frac{\partial^2}{\partial t \partial t'} K(t, t') \in C(\mathcal{X}_1 \times \mathcal{X}_1). \tag{2.9}
$$

A straightforward explanation on this condition is as follows. Denote by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ the inner

product of RKHS $\mathcal{H}$ in (2.3). Then, for any $g \in \mathcal{H}$,

$$\frac{\partial g(\mathbf{t})}{\partial t_j} = \frac{\partial \langle g, K_d(\mathbf{t}, \cdot)\rangle_{\mathcal{H}}}{\partial t_j} = \left\langle g, \frac{\partial K_d(\mathbf{t}, \cdot)}{\partial t_j} \right\rangle_{\mathcal{H}},$$

where the last step is by the continuity of $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. This implies that the composite functional of evaluation and partial differentiation $\partial g / \partial t_j(\mathbf{t})$ is a bounded linear functional in $\mathcal{H}$ and has a representer $\partial K_d(\mathbf{t}, \cdot)/\partial t_j$ in $\mathcal{H}$.

From Mercer's theorem (Riesz and Sz.-Nagy, 1955), K admits a spectral decomposition

$$K(t, t') = \sum_{\nu=1}^{\infty} \lambda_\nu \psi_\nu(t) \psi_\nu(t'),$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are its eigenvalues and $\{\psi_\nu : \nu \geq 1\}$ are the corresponding eigenfunctions.

We are now in the position to present our main results. We first state a minimax lower bound under regular lattices.

Theorem 2.1. *Assume that $\lambda_\nu \asymp \nu^{-2m}$ for some $m > 3/2$, and design points $\mathbf{t}^{(0)}$ and $\mathbf{t}^{(j)}, j = 1, \ldots, d$, are from the regular lattice. Suppose that $f_0 \in \mathcal{H}$ has periodic boundaries on $\mathcal{X}_1^d$ and is truncated up to r interactions in (2.2). Then under the error structure (2.4), as $n \to \infty$,*

$$\begin{aligned} &\infty_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{E} \int_{\mathcal{X}_1^d} \left[\tilde{f}(\mathbf{t}) - f_0(\mathbf{t})\right]^2 d\mathbf{t} \\ &\qquad = \begin{cases} C_1 \left[n (\log n)^{1-(d-p)\wedge r}\right]^{-2m/(2m+1)}, & \text{if } 0 \leq p < d \\ C_2 \{n^{-1} (\log n)^{r-1} + n^{-2mr/[(2m+1)r-2]}\}, & \text{if } p = d \end{cases} \end{aligned}$$

where constants C_1, C_2 do not depend on n.

For two scalars $\{a, b\}$, $a \vee b$ denotes their maximizer and $a \wedge b$ denotes their minimizer. We relegate the proof of Theorem 2.1 to Section A.2.1 in the Appendix. Next, we show the lower bounds of convergence rates in Theorem 2.1 are obtainable. We consider the method of regularization by simultaneously minimizing the empirical losses of function

observations and partial derivative observations with a single penalty:

$$\widehat{f}_{n\lambda} = \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{n(p+1)} \left[\frac{1}{\sigma_0^2} \sum_{i=1}^{n} \left\{ y_i^{(0)} - f(\mathbf{t}_i^{(0)}) \right\}^2 + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \left\{ y_i^{(j)} - \frac{\partial f}{\partial t_j}(\mathbf{t}_i^{(j)}) \right\}^2 \right] + \lambda J(f) \right\}, \tag{2.10}$$

where the weighted squared error loss may be replaced by other convex losses, and $J(\cdot)$ is a quadratic penalty associated with RKHS $\mathcal{H}$, and $\lambda \geq 0$ is a tuning parameter. By the representer lemma (Wahba, 1990), (2.10) has a closed-form solution. If the variances σ_j^2s are unknown, we can replace σ_j^2s in (2.10) by consistent estimators for the variances (Hall et al., 1990). For estimator (2.10), the empirical loss of partial derivatives adds a further regularity restriction to the estimation compared with the traditional smoothing splines in Wahba (1990). The following theorem shows $\widehat{f}_{n\lambda}$ in (2.10) is minimax rate optimal.

Theorem 2.2. *Under the conditions of Theorem 2.1, $\widehat{f}_{n\lambda}$ given by (2.10) satisfies*

$$\mathbb{E} \int_{\mathcal{X}_1^d} \left[\widehat{f}_{n\lambda}(\mathbf{t}) - f_0(\mathbf{t}) \right]^2 d\mathbf{t} = \begin{cases} C_1' \left[n(\log n)^{1-(d-p)\wedge r} \right]^{-2m/(2m+1)} & \text{if } 0 \leq p < d, \\ C_2' n^{-1} (\log n)^{r-1} + n^{-2mr/[(2m+1)r-2]} & \text{if } p = d, \end{cases}$$

where constants C_1', C_2' do not depend on n, if the tuning parameter λ is chosen by $\lambda \asymp \left[n(\log n)^{1-(d-p)\wedge r} \right]^{-2m/(2m+1)}$ when $0 \leq p < d$, and $\lambda \asymp n^{-(2mr-2)/[(2m+1)r-2]}$ when $p = d, r \geq 3$, and $\lambda \asymp (n \log n)^{-(2m-1)/2m}$ when $p = d, r = 2$, and $\lambda \lesssim n^{-(m-1)/m}$ when $p = d$, $r = 1$.

For two positive sequences a_n and b_n, $a_n \lesssim b_n$ (or $a_n \gtrsim b_n$) means that there exists a constant $c > 0$ (or $c' > 0$) such that $a_n \leq c b_n$ (or $a_n \geq c' b_n$) for all n. The proof of Theorem 2.2 is presented in Section A.2.2 in the Appendix. Theorems 2.1 and 2.2 together imply that

with model (2.1) and regular lattices, the minimax optimal rate for estimating $f_0 \in \mathcal{H}$ is

$$\mathbb{E}\int_{\mathcal{X}_1^d}\left[\widehat{f}(\mathbf{t})-f_0(\mathbf{t})\right]^2 d\mathbf{t} = \begin{cases} C_1\left[n(\log n)^{1-(d-p)\wedge r}\right]^{-2m/(2m+1)}, & \text{if } 0\le p<d, \\ C_2\{n^{-1}(\log n)^{r-1}+n^{-2mr/[(2m+1)r-2]}\}, & \text{if } p=d, \end{cases} \tag{2.11}$$

and the estimator in RKHS achieves (2.11). We make several remarks. First, suppose there is no derivative data. Then, (2.11) recovers $[n(\log n)^{1-d}]^{-2m/(2m+1)}$ and this rate is known (Gu, 2013). For a large n, the exponential term $(\log n)^{d-1}$ makes the full d-interaction SS-ANOVA model impractical for large d. On the contrary, suppose partial derivatives data are available. Then, (2.11) gives $n^{-2m/(2m+1)}$ for any $d \ge 1$, which coincides with the classical optimal rate for additive models (Hastie and Tibshirani, 1990; Stone, 1985) and is not affected by the dimension d.

Second, if partial derivative observations are available on all covariates with $p = d$, then the optimal rate can be improved. In addition to (2.6) for $r = d$ and $d \ge 3$, we point out some other interesting cases. For the additive model with $r = 1$ and $d \ge 1$, (2.11) provides the minimax rate n^{-1}. For the pairwise interaction model with $r = 2$ and $d \ge 1$, (2.11) provides the minimax rate $n^{-1}\log n$, which is different from n^{-1} only by a $\log n$ multiplier.

Third, we remark on an "interaction reduction" phenomenon in the sense that the optimal rate for estimating an unknown SS-ANOVA model by incorporating partial derivative data is the same as the optimal rate for estimating a *reduced interaction* SS-ANOVA without derivative data. For example, with $r = d$ and $p = 1$, (2.11) gives $[n(\log n)^{1-(d-1)}]^{-2m/(2m+1)}$, which is the same rate as $r = d-1$ and $p = 0$ involving no derivative data but a lower degree of interactions. And, with $r = d$ and $p = 2$, (2.11) gives $[n(\log n)^{1-(d-2)}]^{-2m/(2m+1)}$, which is the same rate as $r = d-2$ and $p = 0$ involving no derivative observations but two lower degrees of interactions. Similarly, we can extend the same discussion to $p = 3, \ldots, d-1$.

Fourth, the proofs for Theorems 2.1 and 2.2 indicate that when $p = d$, both the squared bias and variance are smaller in magnitude than $p < d$, and when $d - r < p < d$, only the

variance is smaller in magnitude than $0 \leq p \leq d - r$.

Finally, let n_0 denote the sample size on $(\mathbf{t}^{(0)}, Y^{(0)})$ and n_j denote the sample sizes on $(\mathbf{t}^{(j)}, Y^{(j)})$, where $1 \leq j \leq p$. If n_0 and n_js are not all identical to n, n in (2.11) can be replaced by $\min_{1 \leq j \leq p} n_j$.

2.4 Minimax Risks with Random Designs

We now turn to random designs for the minimax optimal rates of estimating $f_0(\cdot)$ with the regression model (2.1). Parallel to Theorem 2.1, we have the following minimax lower bound of estimation under random designs.

Theorem 2.3. *Assume that $\lambda_\nu \asymp \nu^{-2m}$ for some $m > 3/2$, and design points $\mathbf{t}^{(0)}$ and $\mathbf{t}^{(j)}, j = 1, \ldots, d$, are independently drawn from $\Pi^{(0)}$ and $\Pi^{(j)}$s, respectively. Suppose that $\Pi^{(0)}$ and $\Pi^{(j)}$s have densities bounded away from zero and infinity, and $f_0 \in \mathcal{H}$ is truncated up to r interactions in (2.2). Then under the error structure (2.4), as $n \to \infty$,*

$$\infty_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{P} \left\{ \int_{\mathcal{X}_1^d} \left[\tilde{f}(\mathbf{t}) - f_0(\mathbf{t}) \right]^2 d\mathbf{t} \geq C_3 \left(\left[n (\log n)^{1-(d-p)\wedge r} \right]^{-2m/(2m+1)} 1_{0 \leq p < d} \right. \right.$$
$$\left. \left. + \left[n^{-1} (\log n)^{r-1} + n^{-2mr/[(2m+1)r-2]} \right] 1_{p=d} \right) \right\} > 0$$

where constant C_3 does not depend on n.

The lower bound is established via Fano's lemma (Tsybakov, 2009). The proof is deferred to Section A.3.1. Next, we show the lower bounds of convergence rates in Theorem 2.3 can be achieved by using the estimator (2.10) in RKHS.

Theorem 2.4. *Under the conditions of Theorem 2.3, we assume that $\Pi^{(0)}$ and $\Pi^{(j)}$s are known, and*

$m > 2$. *Then,* $\widehat{f}_{n\lambda}$ *in (2.10) satisfies*

$$\lim_{C_3' \to \infty} \limsup_{n \to \infty} \sup_{f_0 \in \mathcal{H}} \mathbb{P} \left\{ \int_{\mathcal{X}_1^d} \left[\widehat{f}_{n\lambda}(\mathbf{t}) - f_0(\mathbf{t}) \right]^2 d\mathbf{t} \right.$$
$$> C_3' \left(\left[n(\log n)^{1-(d-p)\wedge r} \right]^{-2m/(2m+1)} 1_{0 \le p < d} \right.$$
$$\left. \left. + \left[n^{-1}(\log n)^{r-1} + n^{-2mr/[(2m+1)r-2]} \right] 1_{p=d} \right) \right\} = 0$$

if the tuning parameter λ *is chosen by* $\lambda \asymp \left[n(\log n)^{1-(d-p)\wedge r} \right]^{-2m/(2m+1)}$ *when* $0 \le p < d$, *and* $\lambda \asymp n^{-(2mr-2)/[(2m+1)r-2]}$ *when* $p = d, r \ge 3$, *and* $\lambda \asymp (n \log n)^{-(2m-1)/2m}$ *when* $p = d, r = 2$, *and* $\lambda \lesssim n^{-(m-1)/m}$ *when* $p = d$, $r = 1$. *In other words,* $\widehat{f}_{n\lambda}$ *is rate optimal.*

We use the linearization method in Cox and O'Sullivan (1990) to prove Theorem 2.4. The key ingredient of this method is to pick a suitable basis such that the expected loss of the regularization and the quadratic penalty $J(\cdot)$ can be simultaneously diagonalized. Our situation is unique in the sense that the loss function in (2.10) is the sum of squared error losses for both the function and partial derivatives but we are only interested in estimating the function itself in Theorem 2.4. This induces a third positive semi-definite functional, which is the squared error loss of function estimation. But three functionals are not guaranteed to be simultaneously diagonized, this making the direct application of the linearization method infeasible. We present a detailed proof in Section A.3.1.

Theorems 2.3 and 2.4 together demonstrate the fundamental limit rate of the squared error loss for estimating $f_0 \in \mathcal{H}$ with model (2.1) and random designs is

$$\left[n(\log n)^{1-(d-p)\wedge r} \right]^{-2m/(2m+1)} 1_{0 \le p < d} + \left[n^{-1}(\log n)^{r-1} + n^{-2mr/[(2m+1)r-2]} \right] 1_{p=d} \tag{2.12}$$

in a probabilistic sense, and the estimator in RKHS achieves (2.12). The minimax rate is the same as that with the regular lattice. We make several remarks on (2.12). First, the five remarks following (2.11) for the mean squared situation hold for (2.12) in a probabilistic

sense. Second, for the special case when $p = 0$, (2.12) recovers the minimax optimal rate of convergence $O_{\mathbb{P}}\left\{[n(\log n)^{1-r}]^{-2m/(2m+1)}\right\}$ for SS-ANOVA models, which is known in Lin (2000). Third, the squared error loss in Theorems 2.3 and 2.4 can be replaced by squared prediction error $\int\{\widehat{f}_{n\lambda}(\mathbf{t}) - f_0(\mathbf{t})\}^2 d\Pi^{(0)}(\mathbf{t})$ and it achieves the same minimax optimal rate as (2.12).

As a byproduct of Theorem 2.4, we obtain the following result of estimating the mixed partial derivatives $\partial^d f_0/\partial t_1 \cdots \partial t_d(\mathbf{t})$ by its natural estimator $\partial^d \widehat{f}_{n\lambda}/\partial t_1 \cdots \partial t_d(\mathbf{t})$.

Corollary 2.5. *Under the conditions of Theorem 2.4 and $m > 3$, then*

$$\lim_{D_1' \to \infty} \limsup_{n\to\infty} \sup_{f_0 \in \mathcal{H}} \mathbb{P}\left\{ \int_{\mathcal{X}_1^d} \left[\frac{\partial^d \widehat{f}_{n\lambda}(\mathbf{t})}{\partial t_1 \cdots \partial t_d} - \frac{\partial^d f_0(\mathbf{t})}{\partial t_1 \cdots \partial t_d} \right]^2 d\mathbf{t} \right.$$
$$> D_1' \left(\left[n(\log n)^{1-(d-p)\wedge r} \right]^{-2(m-1)/(2m+1)} 1_{0\le p<d} \right.$$
$$\left.\left. + \left[n^{-2(m-1)r/[(2m+1)r-2]} \right] 1_{p=d} \right) \right\} = 0,$$

if the tuning parameter λ is chosen by $\lambda \asymp \left[n(\log n)^{1-(d-p)\wedge r}\right]^{-2m/(2m+1)}$ when $0 \le p < d$, and $\lambda \asymp n^{-(2mr-2)/[(2m+1)r-2]}$ when $p = d$.

2.5 Minimax Risk for Estimating Partial Derivatives

Suppose noisy observations of data on the function and some partial derivatives in (2.1) are available. We are interested in the optimal rate for estimating first-order partial derivatives by using all observed data. For brevity, we only consider random designs although similar results can be obtained for regular lattices. The following theorem gives the minimax lower bound for estimating $\partial f_0/\partial t_j$, $1 \le j \le p$.

Theorem 2.6. *Assume that $\lambda_\nu \asymp \nu^{-2m}$ for some $m > 2$ and design points $\mathbf{t}^{(0)}$ and $\mathbf{t}^{(j)}, j = 1, \ldots, d$, are independently drawn from $\Pi^{(0)}$ and $\Pi^{(j)}$s, respectively. Suppose that $\Pi^{(0)}$ and $\Pi^{(j)}$s have densities bounded away from zero and infinity, and $f_0 \in \mathcal{H}$ is truncated up to r interactions in*

(2.2). Then under the error structure (2.4), for any $j \in \{1, \ldots, p\}$ and $1 \leq r \leq d$, as $n \to \infty$,

$$\infty_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{P}\left\{ \int_{\mathcal{X}_1^d} \left[\tilde{f}(\mathbf{t}) - \frac{\partial f_0(\mathbf{t})}{\partial t_j} \right]^2 d\mathbf{t} \geq C_4 n^{-2(m-1)/(2m-1)} \right\} > 0,$$

where constant C_4 does not depend on n.

We will prove this theorem in Section A.4.1 in the Appendix. As a natural estimator for $\partial f_0/\partial t_j$, $\partial \widehat{f}_{n\lambda}/\partial t_j$ achieves the lower bound of convergence rates in Theorem 2.6.

Theorem 2.7. *Under the conditions of Theorem 2.6, $\widehat{f}_{n\lambda}$ given by (2.10) satisfies that for any $j \in \{1, \ldots, p\}$ and $1 \leq r \leq d$,*

$$\lim_{C_4' \to \infty} \limsup_{n \to \infty} \sup_{f_0 \in \mathcal{H}} \mathbb{P}\left\{ \int_{\mathcal{X}_1^d} \left[\frac{\partial \widehat{f}_{n\lambda}(\mathbf{t})}{\partial t_j} - \frac{\partial f_0(\mathbf{t})}{\partial t_j} \right]^2 d\mathbf{t} > C_4' n^{-2(m-1)/(2m-1)} \right\} = 0,$$

if the tuning parameter λ is chosen by $\lambda \asymp n^{-2(m-1)/(2m-1)}$.

The proof of this theorem is given in Section A.4.2 in the Appendix. When $r = 1$, this result coincides with Corollary 2.5. Unlike Theorem 2.4 and Corollary 2.5, the distributions $\Pi^{(0)}$ and $\Pi^{(j)}$s are not assumed to be known.

Theorems 2.6 and 2.7 together give the minimax optimal rate for estimating $\partial f_0/\partial t_j$, provided in (2.8). Since the optimal rate in (2.8) holds regardless of the value of $p \geq 1$, first-order partial derivative data on different covariates do not improve the optimal rates for estimating each other. For example, given noisy data on $f_0(\cdot)$ and $\partial f_0/\partial t_j(\cdot)$, data on $\partial f_0/\partial t_k(\cdot)$ do not improve the minimax optimal rate for estimating $\partial f_0/\partial t_j(\cdot)$ if $1 \leq k \neq j \leq p$.

2.6 Real Data and Simulation Examples

This section consists of four examples. We give a real example on actuarial life table in Example 2.8 to demonstrate benefits of incorporating first-order partial derivatives for estimation. We provide another real application of multivariate estimation in manufacturing

in Example 2.9. We present simulations in Example 2.10 and 2.11 to corroborate the proposed theory and compare our estimator with the estimator in Hall and Yatchew (2007).

Example 2.8 (Survival distribution in actuarial life table)**.** *The life table in actuarial practice provides probabilities of survival and death at integer ages (Frees and Valdez, 1998) In order to value payments that are not at integer ages, actuaries need to make a fractional age assumption for probabilities of surviving at fractional ages. This is smoothing the data given at integer ages for survival function estimation. We consider a real data of U.S. 2015 period life table (*`www.ssa.gov/OACT/STATS/table4c6.html#fn2`*) for the Social Security area of male and female population separately. Write $f_0(t)$ as the survival distribution function and $u(t)$ as the force of mortality function. Then,*

$$f_0'(t) = -f_0(t)u(t).$$

Here, data on $f_0(t)$ can be calculated using the death probability in life table and $u(t)$ can be estimated by divided differences using $l(t)$ the number of people that survive at age t (Jones and Mereu, 2002),

$$u(t) = \frac{l(t-1) - l(t+1)}{2l(t)}, \;\; u(0) = \frac{3l(0) - 4l(1) + l(2)}{2l(0)}.$$

The data on $f_0(t)$, denoted by $Y^{(0)}$, and the estimate of $u(t)$ together yield the data on derivative $f_0'(t)$, denoted by $Y^{(1)}$. The random error of $Y^{(0)}$ and $Y^{(1)}$ for the current data satisfies our error structure (2.4).

Table 2.1: MSE of only incorporating $Y^{(0)}$ and MSE of incorporating $Y^{(0)}\&Y^{(1)}$ for Example 2.8. The MSEs are in the unit of 10^{-4}

		$n=5$	$n=10$	$n=15$	$n=20$
Male	Incorporating $Y^{(0)}$	15.3674	6.7944	1.7687	0.1745
	Incorporating $Y^{(0)}\&Y^{(1)}$	**7.4381**	**1.6488**	**0.3446**	**0.0227**
Female	Incorporating $Y^{(0)}$	23.0655	9.9948	2.2299	0.5925
	Incorporating $Y^{(0)}\&Y^{(1)}$	**9.4745**	**2.4790**	**0.4091**	**0.0755**

We compare our proposed estimator (2.10) by whether not incorporating derivative data. We use

Table 2.2: MSE of incorporating $Y^{(0)}\&Y^{(1)}$ relative to MSE of only incorporating $Y^{(0)}$ for Example 2.8

	$n=5$	$n=10$	$n=15$	$n=20$
Male	0.4840	0.2426	0.1948	0.1301
Female	0.4108	0.2480	0.1835	0.1274

the Matérn kernel $K(t,t') = (1+|t-t'|/\psi + |t-t'|^2/3\psi^2)\exp(-|t-t'|/\psi)$*, which satisfies the differentiability condition (2.9). Here, the scale parameter* ψ *is chosen by the five-fold cross-validation, and the tuning parameter* λ *in (2.10) is selected by GCV. The training data are selected as the equally spaced integers* t *in the range* $[0,119]$ *with varying sample sizes* $n = 5, 10, 15, 20$*. The boundaries* $\{0,119\}$ *are included in the training set. The MSE=* $\mathbb{E}[\widehat{f}_{n\lambda} - f_0]^2$ *is estimated by a test set consisting of all 120 samples with* $t \in \{0,1,\cdots,119\}$*. Table 2.1 summarizes the averaged MSEs over 200 experiments in each setting. A significant improvement of estimation is achieved by incorporating the derivative data. Table 2.2 provides the ratios of MSE of incorporating* $Y^{(0)}, Y^{(1)}$ *relative to MSE of only incorporating* $Y^{(0)}$*. The ratios decrease as the sample size increases, which confirms our theorem that incorporating derivatives accelerates the rate.*

Example 2.9 (Production time of CLFAS)**.** *The closed-loop flexible assembly system (CLFAS) in the design for manufacturing is known to be effective in lowering production cost and increasing flexibility; see, e.g., Suri and Leung (1987); Chen et al. (2013). A significant amount of cost is required for building a CLFAS. Hence, it is important to rapidly and accurately estimate the performance of CLFAS. We show in this example that first-order partial derivatives can be estimated at little cost and incorporating of them can significantly improve the estimation accuracy.*

As shown in Figure 2.1, consider a CLFAS consisting of six automatic workstations that is connected by a conveyor with six pallets in the system. Unfinished parts are loaded and unloaded through workstation 1 and proceed through CLFAS on the pallets. The operation time at workstation j*,* $1 \le j \le 6$*, is given by*

$$t_j + 1\{\textit{jam at station } j\} \cdot R_j,$$

where t_j *is the fixed machine production time (in minutes) and* R_j *is the additional random time*

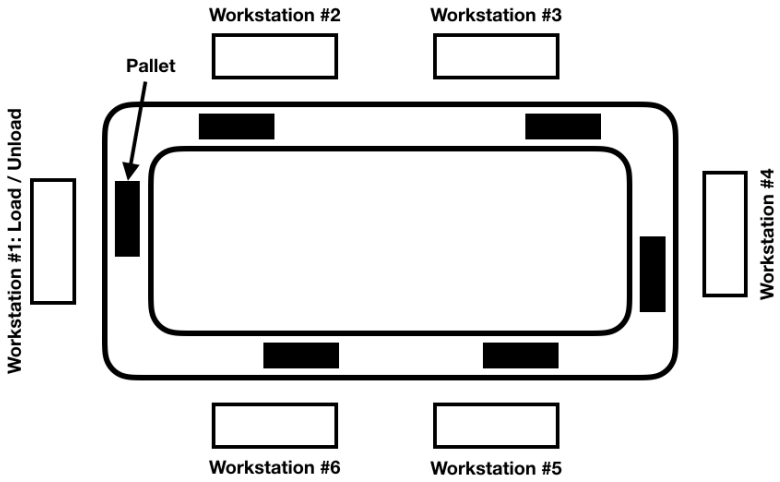

Figure 2.1: Diagram of the closed-loop flexible assembly system for Example 2.9

to clear the machine j if it jams. Let p_j be the probability that a part causes jam at workstation j. In our experiment, we set $p_j = 0.005$ and R_j to be i.i.d. uniformly drawn from $[6, 66]$. Because the operation times are random, queueing would occur. We are interested in predicting the expected production time of the first 5000 parts completed by the CLFAS, and we denote it by $f_0(t_1, \ldots, t_6)$ as a function of t_js. The estimation of f_0 helps identify the bottleneck workstations so that resources can be better distributed. If no queue occurs, $f_0(t_1, \ldots, t_6)$ is a SS-ANOVA function since it is additive in $t_1, \ldots, t_6$. The analysis below can be generalized to any number of workstations or pallets.

The following algorithm gives data on function f_0 and the unbiased estimators for partial derivatives $\partial f_0/\partial t_j$, $1 \le j \le 6$. The algorithm is based on the infinitesimal perturbation analysis (IPA) (Suri and Leung, 1987), which simply adds some accumulator variables A_{j_1,j_2} to be updated during the simulation.

1. *Initialize: $A_{j_1,j_2} \leftarrow 0$ for $j_1, j_2 = 1, \ldots, 6$.*

2. *At the end of an operation at station j, let $A_{j,j} \leftarrow A_{j,j} + 1$, $j = 1, \ldots, 6$.*

3. *If a pallet leaving station j_1 going to station j_1' terminates an idle period of station j_1', then set $A_{j_1',j_2} \leftarrow A_{j_1,j_2}$, $j_2 = 1, \ldots, 6$.*

4. *If a pallet leaving station j_1 going to station j_1' terminates a blocked period of station j_1, then set $A_{j_1,j_2} \leftarrow A_{j_1',j_2}$, $j_2 = 1, \ldots, 6$.*

5. *At the end of the simulation, let P denote the total number of parts completed and L be total length of simulation in time unites. The data on f_0 is given by $Y^{(0)}(\mathbf{t}) = L/P$. The IPA estimator for $\partial f_0/\partial t_j$ is $Y^{(j)}(\mathbf{t}) = A_{6,j}/P$, $j = 1, \ldots, 6$.*

The random noises exist in data $Y^{(0)}$ and $Y^{(j)}$s due to the stochastic nature of CLFAS. We compare estimation results of incorporating partial derivative data and not incorporating derivatives into our estimator (2.10), where the tuning parameter in (2.10) is selected by GCV. The tensor product Matérn kernel $\prod_{j=1}^{6}(1 + |t_j - t_j'|/\psi_j + |t_j - t_j'|^2/3\psi_j^2)\exp(-|t_j - t_j'|/\psi_j)$ is used, where ψ_js are chosen by the five-fold cross-validation. The experimental design is 100 uniform random points in $[3, 9]^6$. To address the impact of stochastic noises, we replicate 100 experiments of CLFAS at each design point with a run length of $P = 5000$ and average data. We note two facts of this data generating. First, obtaining function value at a new design point requires to conduct the experiment 100 more times. This is expensive compared with obtaining partial derivatives which only requires to record a small matrix A shown in above algorithm. Second, the error correlation only exists for function value and partial derivatives at the same design, not between components at different design points. Hence, this error structure satisfies our assumption (2.4).

The MSE$= \mathbb{E}[\widehat{f}_{n\lambda} - f_0]^2$ is estimated by a Monte Carlo sample of 1000 test points in $[3, 9]^6$. Because the true production costs f_0 at the test points are unknown, we approximated them by replicating 1000 experiments of CLFAS at each test point. The experiment of CLFAS is programmed in VBA. We replicate the above procedures for 100 times to compare the MSEs obtained by only incorporating function data $Y^{(0)}$ and by incorporating both function and derivatives $Y^{(0)}, Y^{(1)}, \ldots, Y^{(6)}$. Figure 2.2 gives the box plots of MSEs over these 100 macro-replications. It is evident that incorporating partial derivatives leads to a significant improvement of the estimation error.

Example 2.10 (Cost function in econometrics)**.** *In this example, we compare our estimator (2.10) with the estimator in Hall and Yatchew (2007). We adopt a similar simulation setting of Hall and*

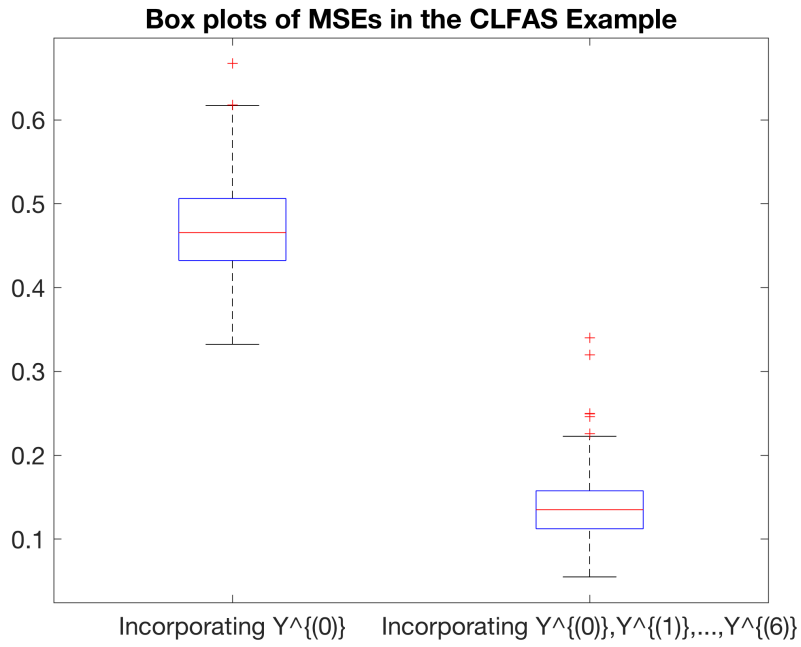

Figure 2.2: The box plots of MSEs for Example 2.9

Yatchew (2007) for estimating cost function. The true cost is

$$f_0(t_1,t_2,t_3,t_4) = c_0^{-\frac{1}{c_1+c_2+c_3}} \prod_{\nu=1}^{3} \left(\frac{c_1+c_2+c_3}{c_\nu}\right)^{\frac{c_\nu}{c_1+c_2+c_3}} \prod_{\nu=1}^{3} t_\nu^{\frac{c_\nu}{c_1+c_2+c_3}} t_4^{\frac{1}{c_1+c_2+c_3}},$$

where t_1, t_2, t_3 are the prices of three factor inputs, t_4 is the level of output produced, c_0 is the efficiency parameter, and c_1, c_2, c_3 are elasticity parameters. Clearly, f_0 has the tensor product structure (2.3). As in Hall and Yatchew (2007), we fix $t_3 = 1$ since the cost function is homogeneous of degree one in (t_1, t_2, t_3), that is, $f_0(t_1, t_2, t_3, t_4) = t_3 f_0(t_1/t_3, t_2/t_3, 1, t_4)$. Suppose data are given on

$$\begin{aligned} Y^{(0)} &= f_0(t_1,t_2,1,t_4) + \epsilon^{(0)} \\ Y^{(j)} &= \frac{\partial f_0(t_1,t_2,1,t_4)}{\partial t_j} + \epsilon^{(j)}, \qquad \text{for } j = 1,2. \end{aligned}$$

Set $c_0 = 1, c_1 = 0.8, c_2 = 0.7, c_3 = 0.6$. Let the designs for t_1, t_2 and t_4 be i.i.d. uniformly drawn from $[0.5, 1.5]$. Suppose that $\epsilon^{(j)}, j = 0, 1, 2$ are Gaussian with zero means, standard deviations 0.35, and correlation ρ.

Table 2.3: MSE of our estimator only incorporating $Y^{(0)}$, MSE of the estimator in Hall and Yatchew (2007) incorporating $Y^{(0)}\&Y^{(1)}\&Y^{(2)}$, and MSE of our estimator incorporating $Y^{(0)}\&Y^{(1)}\&Y^{(2)}$ for Example 2.10. The MSEs are in the unit of 10^{-4}

		Our estimator with only $Y^{(0)}$	Hall and Yatchew (2007) with $Y^{(0)}\&Y^{(1)}\&Y^{(2)}$	Our estimator with $Y^{(0)}\&Y^{(1)}\&Y^{(2)}$
$n = 100$	$\rho = 0$	127.1471	61.4098	**47.4739**
	$\rho = 0.4$	128.9210	63.1006	**49.8963**
	$\rho = 0.9$	129.6300	64.5989	**51.9224**
$n = 200$	$\rho = 0$	76.6199	33.3001	**24.1501**
	$\rho = 0.4$	77.7602	35.0696	**25.5342**
	$\rho = 0.9$	77.9138	36.2591	**27.0137**
$n = 500$	$\rho = 0$	36.1925	16.3861	**9.3499**
	$\rho = 0.4$	38.0683	18.2355	**10.4708**
	$\rho = 0.9$	38.9311	18.7698	**11.0498**
$n = 1000$	$\rho = 0$	21.8570	9.2788	**4.5364**
	$\rho = 0.4$	22.4943	10.4801	**5.1468**
	$\rho = 0.9$	22.9499	10.6193	**5.3288**

We compare the proposed estimator (2.10) with the estimator in Hall and Yatchew (2007) under varying sample size $n = 100, 200, 500, 1000$ *and correlation* $\rho = 0, 0.4, 0.9$. *For our estimator (2.10), the tensor product Matérn kernel* $\prod_{j=1,2,4}(1 + |t_j - t'_j|/\psi_j + |t_j - t'_j|^2/3\psi_j^2)\exp(-|t_j - t'_j|/\psi_j)$ *is used, where* ψ_j*s are chosen by the five-fold cross-validation and the* λ *in (2.10) is selected by GCV. For the estimator in Hall and Yatchew (2007), the kernel smoothing with tensor product Matérn kernel is used for local averaging in the* (t_1, t_4) *or* (t_2, t_4) *directions as the Example 3 of Hall and Yatchew (2007), and then estimators are averaged. The bandwidth parameters for the estimator in Hall and Yatchew (2007) are chosen by the five-fold cross-validation. The MSE=* $\mathbb{E}[\widehat{f}_{n\lambda} - f_0]^2$ *is estimated by a Monte Carlo sample of* 10^6 *test points in* $[0.5, 1.5]^3$.

Table 2.3 gives the MSEs of our estimator (2.10), the MSEs of estimator in Hall and Yatchew (2007), and additionally the MSE of (2.10) with only function data $Y^{(0)}$ *as the reference. In each combination of* n *and* ρ, *the MSEs are averaged over* 1000 *replicated simulations. It is clear from*

Table 2.3 that the MSEs of incorporating partial derivatives are significantly smaller than the MSEs without derivatives. Moreover, the performance of our estimator compares favorably with the estimator in Hall and Yatchew (2007).

Table 2.4 gives the MSEs of our estimator relative to MSEs of the estimator in Hall and Yatchew (2007) by incorporating $Y^{(0)}, Y^{(1)}, Y^{(2)}$. The ratio decreases with the sample size. This phenomenon is expected since our estimator converges at the rate of additive models (see, Theorem 2.4), which is faster than the convergence rate of nonparametric dimension not exceeding two by Hall and Yatchew (2007).

Table 2.4: MSE of our estimator incorporating $Y^{(0)}\&Y^{(1)}\&Y^{(2)}$ relative to MSE of the estimator in Hall and Yatchew (2007) incorporating $Y^{(0)}\&Y^{(1)}\&Y^{(2)}$

	$\rho = 0$	$\rho = 0.4$	$\rho = 0.9$
$n = 100$	0.7731	0.7907	0.8038
$n = 200$	0.7252	0.7281	0.7450
$n = 500$	0.5706	0.5742	0.5887
$n = 1000$	0.4889	0.4911	0.5018

Example 2.11 (The Black-Scholes call option pricing)**.** *In many stochastic simulations, first-order partial derivatives can be obtained with negligible effort compared to obtain the function observation itself. As an illustration, we study an example of pricing a call option. We would illustrate our theoretical results in this example.*

The Black-Scholes stochastic differential equation is commonly used to model the price S_t of a stock at time t through $dS_t = rS_t dt + \sigma S_t dW_t, t \geq 0$, where W_t is Wiener process, r is the risk-free rate, and σ is the volatility parameter of stock price. This equation has a closed-form solution: $S_t = S_0 \exp\{(r - \frac{1}{2}\sigma^2)t + \sigma\sqrt{t}Z\}$, where $Z \sim N(0,1)$. The European call option is a right to buy a stock at the presepcified date $t = T$ with a prespecified price K and the function value is

$$Y^{(0)} = e^{-rT}(S_T - K)_+.$$

The goal is to estimate the net present value of this option with fixed T, K:

$$f_0(S_0, r, \sigma; T, K) = E[Y^{(0)}]. \tag{2.13}$$

The sensitivities of interest are the partial derivatives of f_0 with respect to the parameters (S_0, r, σ) while holding (T, K) fixed. Partial derivative estimators for $\partial f_0/\partial S_0$, $\partial f_0/\partial r$, $\partial f_0/\partial \sigma$ obtained by the infinitesimal perturbation analysis (IPA) are, respectively,

$$\begin{aligned} Y^{(1)} &= e^{-rT}\frac{S_T}{S_0} \cdot 1\{S_T \geq K\}, \\ Y^{(2)} &= -TY^{(0)} + e^{-rT}TS_T \cdot 1\{S_T \geq K\}, \\ Y^{(3)} &= e^{-rT}\frac{1}{\sigma}[\log(S_T/S_0) - (r + \frac{1}{2}\sigma^2)T]S_T \cdot 1\{S_T \geq K\}. \end{aligned} \tag{2.14}$$

It can be shown that IPA estimators (2.14) are unbiased, that is e.g., $\mathbb{E}[Y^{(1)}] = \partial f_0/\partial S_0$. We refer to Glasserman (2013); L'Ecuyer (1990) for details.

In this experiment, we fix $T = 1$ and $K = 100$. The experiment design is as follows. Choose l equally spaced design points for each of three covariates: $S_0 \in [80, 120]$, $r \in [0.01, 0.05]$, and $\sigma \in [0.2, 1]$ with $l = 7, 14, 21$. The end points of each interval are always included. Hence the design has the tensor product structure with sample size $n = 7^3, 14^3, 21^3$. To address the impact of stochastic simulation noise, we simulate $q = 1000, 2000, 5000$ i.i.d. replications of S_T at each design point and average, and the independent sampling is used across design points. Here, a larger q corresponds to smaller noise variances of $Y^{(j)}$s.

Two facts of this data generating are noted. First, obtaining function value at a new design point requires to generate q new random numbers for getting S_T. However, obtaining a partial derivative estimate in (2.14) does not need any new random number. Second, the error correlation only exists for function value and partial derivatives at the same design, not between components at different design points. Hence, this error structure satisfies assumption (2.4).

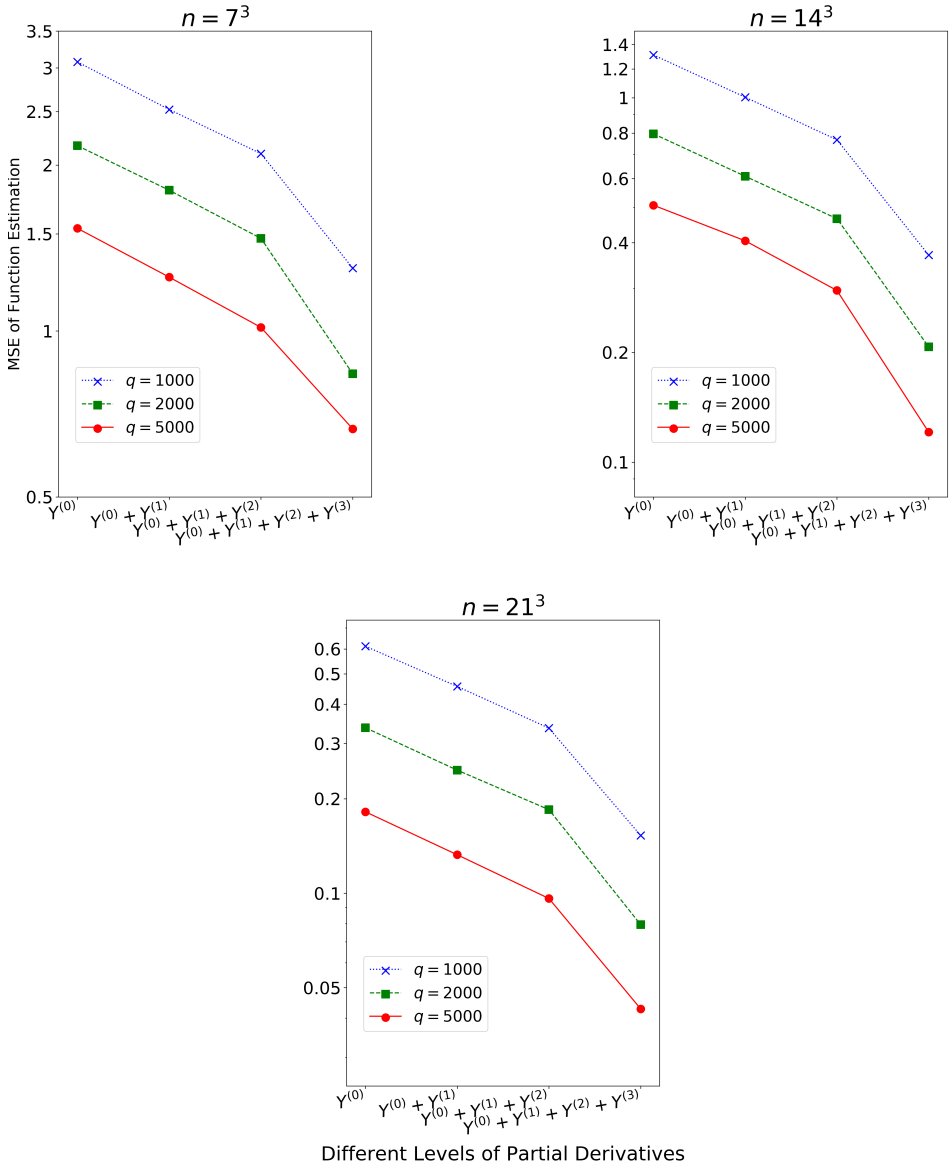

Figure 2.3: Estimation error of our regularized estimator incorporating different levels of partial derivatives for Example 2.11. The y-axis is in the log scale.

We compare estimation results of our proposed estimator (2.10) with different levels of partial derivative data. We use the tensor product Matérn kernel $\prod_{j=1}^{3}(1 + |t_j - t'_j|/\psi_j + |t_j - t'_j|^2/3\psi_j^2)\exp(-|t_j - t'_j|/\psi_j)$*, which satisfies the differentiability condition (2.9). Here, the scale parameter* ψ_j*s are chosen by the five-fold cross-validation, and the tuning parameter* λ *in (2.10) is selected by GCV. It is known that* $f_0(S_0, r, \sigma; T, K)$ *defined in (2.13) has an explicit solution*

$f_0(S_0, r, \sigma; 1, 100) = S_0\Phi(-d_1 + \sigma) - 100e^{-r}\Phi(-d_1)$ *when* $T = 1, K = 100$, *where* $d_1 = \sigma^{-1}[\log 100 - \log(S_0) - (r - \sigma^2/2)]$ *and* $\Phi(\cdot)$ *is the cdf of the standard normal distribution. The MSE=* $\mathbb{E}[\widehat{f}_{n\lambda} - f_0]^2$ *is estimated by a Monte Carlo sample of 10000 test points in* $[80, 120] \times [0.01, 0.05] \times [0.2, 1]$.

Figure 2.3 shows the estimation error, $\mathbb{E}[\widehat{f}_{n\lambda} - f_0]^2$, *when the sample size* $n = 7^3, 14^3, 21^3$ *for each combination of* q *and different levels of partial derivatives–only function data (i.e.,* $p = 0$*), function data with one type of first partial derivative (i.e.,* $p = 1$*), function data with two types of first partial derivatives (i.e.,* $p = 2$*), function data with three types of first partial derivatives (i.e.,* $p = 3$*). The results are averaged over 1000 simulations in each setting. The* y*-axis is in the log scale. Figure 2.3 suggests the estimation error converges exponentially with the number of types of first partial derivatives (i.e.,* p*), which agrees with our theoretical results. We also observe that the convergence rate increases when incorporating* $p = 3$ *partial derivatives compared with* $p \leq 2$. *This also confirms our theoretical finding that the faster rate* $n^{-1}(\log n)^{d-1} + n^{-2md/[(2m+1)d-2]}$ *is achieved when using all first partial derivatives* $p = d$, *compared to the rate* $[n(\log n)^{1+p-d}]^{-2m/(2m+1)}$ *when* $p < d$, *where* $d = 3$ *in this example. Furthermore, Figure 2.3 indicates that within each* n *the slopes are very close across different* q, *and the slopes get steeper when* n *increases. For example, we provide in Table 2.5 the ratios of MSE of incorporating* $Y^{(0)}, Y^{(1)}$ *and* $Y^{(2)}$ *(i.e.,* $p = 2$*) relative to MSE of only incorporating* $Y^{(0)}$ *(i.e.,* $p = 0$*). This further corroborates our derived results that incorporating derivatives leads to the faster convergence rates. Finally, it is clear that the estimation error decreases as the stochastic error decreases (i.e.,* q *increases).*

Table 2.5: MSE of incorporating $Y^{(0)}\&Y^{(1)}\&Y^{(2)}$ relative to MSE of only incorporating $Y^{(0)}$ for Example 2.11

n	$q = 1000$	$q = 2000$	$q = 5000$
$7^3 = 343$	0.6818	0.6789	0.6612
$14^3 = 2744$	0.5850	0.5848	0.5835
$21^3 = 9261$	0.5484	0.5483	0.5294

In this stochastic simulation example, partial derivatives can be easily estimated by (2.14) without additional cost. Although the function f_0 *does not have tensor product structure, our estimator*

with first-order partial derivatives gives substantial improvements in function estimation. For example, the MSE of $n = 7^3, q = 1000$ *with three types of partial derivatives included is even smaller than the MSE of* $n = 14^3, q = 1000$ *with no partial derivative included. This shows the use of derivatives saves the computational cost for sampling at new designs in order to achieve a same estimation accuracy.*

Chapter 3

High-Dimensional Smoothing Splines with Application to Alzheimer's Disease Prediction Using Longitudinal and Heterogeneous Magnetic Resonance Imaging

3.1 Introduction

Alzheimer's Disease (AD) is the most common cause of dementia in the aged population (Prince et al., 2013). In order to prevent disease progression and take therapeutic treatment in the earliest stage, it is vital to identify AD-related pathological biomarkers of progression and diagnose early-stage AD. A considerable amount of research has been devoted to the use of structured magnetic resonance imaging (MRI) for early-stage AD diagnosis; e.g., Jack Jr et al. (2010, 2013). The structural MRI provides measures of cerebral atrophy and it is shown to be most closely coupled with clinical symptoms in AD (Jack Jr et al., 2009).

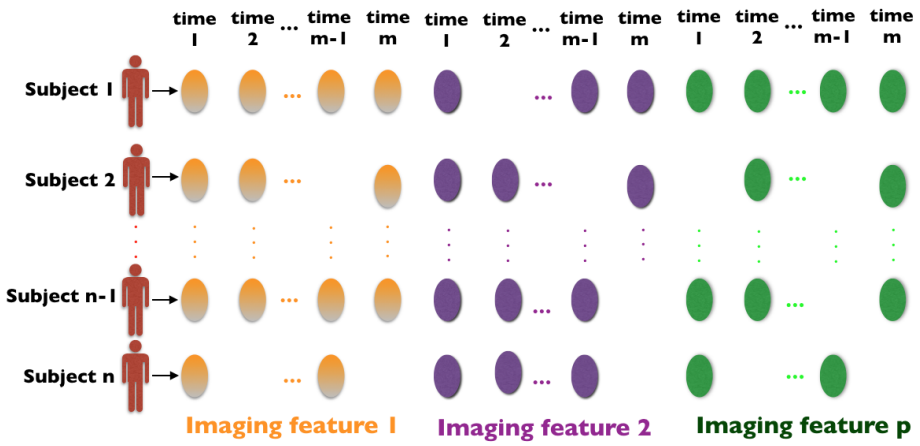

Figure 3.1: Illustration of heterogenous longitudinal data with p covariates.

Most work in the literature focus on the cross-sectional study with MRI collected at one single time point; see, e.g., Aguilar et al. (2013); Liu et al. (2016); Tzourio-Mazoyer et al. (2002). However, the cross-sectional study could be insensitive to early pathological changes. As an alternative, longitudinal analysis of structural abnormalities has recently attracted attentions (Chincarini et al., 2016; Yau et al., 2015; Zhang et al., 2012). Most of these existing longitudinal studies focus on the atrophy of a few well-known biomarkers such as the hippocampus, entorhinal cortex, and ventricular cortex. However, these prespecified regions of interest (ROIs) may be insufficient to capture the full morphological abnormality pattern of the brain MRI. Besides it, a few other issues remain as challenges in the longitudinal analysis. First, longitudinal scans across subjects are usually inconsistent. For example, subjects could have different scanning time and different total number of scans. Second, the total number of ROIs in the brain is large compared with the number of subjects, which poses a challenge to select AD-rated longitudinal biomarkers from the whole brain. Third, the rates of longitudinal change in different ROIs are different and this heterogeneity should be accounted by the modeling of progression.

The goal of this paper is to identify important AD-related ROIs in the whole brain MRI with longitudinal MRI data and use the selected ROIs for AD prediction. Specifically, we use the varying coefficient model (Hastie and Tibshirani, 1993) to characterize the heterogeneous

changes of different ROIs in structural MRI. This model also allows a nonlinear functional modeling between MRI and clinical cognition functions. We propose a novel feature selection method by combining the smoothing splines and a l_1-penalty, which can simultaneously select and estimate AD-related ROIs. We provide an efficient algorithm to implement the proposed feature selection method. Then the prediction is performed based on the selected longitudinal features and estimated varying coefficients. Our method is robust to the inconsistency among longitudinal scans and is adaptive to the heterogeneity of changes in different ROIs. The use of varying coefficient models is motivated by the hypothetical AD dynamic biomarkers curves proposed by Jack Jr et al. (2010, 2013), where their principle is that the rates of change over time for MRI and clinical cognition functions are in a temporally ordered manner. Hence, the functional relationship between the atrophy of MRI and the change in clinical cognition functions must be nonlinear in time.

To evaluate our method, we perform experiments using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). We predict future clinical changes of mild cognitive impairment (MCI) subjects with brain MRI data. The MCI is a prodromal stage of AD. The prediction of clinical changes help to determine whether a MCI subject will convert into AD at a future time point, which is vital for early diagnosis of AD.

Main differences between this paper and existing longitudinal studies in Chincarini et al. (2016); Yau et al. (2015); Zhang et al. (2012) are as follows.

- Different feature representations. We use the varying coefficient model to characterize nonlinear and smooth progression of longitudinal features, which is motivated by clinical findings and the dynamic biomarker curve in Jack Jr et al. (2010, 2013). On the other hand, Chincarini et al. (2016); Yau et al. (2015); Zhang et al. (2012) use linear representations for features.

- Different scalability to heterogenous longitudinal scans. Different from Chincarini et al. (2016); Yau et al. (2015); Zhang et al. (2012), our method does not require same scanning times and a same number of scans across samples.

- Different feature selections. We proposed a novel feature selection method by combining smoothing splines with a l_1-penalty, which allows to simultaneously select and estimate features. This is different from the two-step method in Zhang et al. (2012) by doing the selection and estimation separately and Chincarini et al. (2016); Yau et al. (2015) by only using pre-selected features.

The rest of the paper is organized as follows. We introduce our method in Section 3.2. We give experiment results in Section 3.3. Additional material and proofs are relegated to Appendix.

3.2 Methodology

The varying coefficient model (Hastie and Tibshirani, 1993) can describe time-dependent covariate effects on the responses. Given scaled time $t \in [0, 1]$, the response functional $Y(\cdot)$ is related to covariates $X_1(\cdot), \ldots, X_p(\cdot)$ through

$$Y(t) = b + \sum_{j=1}^{p} \beta_j(t) X_j(t) + \varepsilon(t), \quad b \in \mathbb{R}, \tag{3.1}$$

where the centered noise process $\varepsilon(\cdot)$ is independent of $X_j(\cdot)$s. The model (3.1) allows a nonlinear relationship between $X_j(\cdot)$s and $Y(\cdot)$ be letting the coefficients $\beta_j(\cdot)$s vary on t. On the other hand, (3.1) has an additive structure on covariates $X_j(\cdot)$s, which enables efficient estimations of coefficients $\beta_j(\cdot)$s.

In practice, data are obtained for subject $i = 1, \ldots, n$ at time $t_{i\nu}$, where $\nu = 1, 2, \ldots, m_i$, and $0 \leq t_{i1} \leq t_{i2} \leq \cdots \leq t_{im_i} \leq 1$. Note that m_i and $t_{i\nu}$s are allowed to be different for different subjects i. Denote $X_j(t_{ij}) = x_{ij}$ and let $y_{i\nu}$ be the response for subject i at time $t_{i\nu}$, then (3.1) implies

$$y_{i\nu} = b + \sum_{j=1}^{p} \beta_j(t_{i\nu}) x_{ij}(t_{i\nu}) + \varepsilon(t_{i\nu}), \quad b \in \mathbb{R}. \tag{3.2}$$

The structure of heterogenous longitudinal data is illustrated in Figure 3.1, where some subjects could have missing feature values at certain time point. The number of covariates p in

(3.2) can be larger than the sample size n, and then (3.2) becomes a high-dimensional model. Since some covariates might be irrelevant with the response, we want to select important covariates $X_j(\cdot)$s based on data (3.2) and use the selected covariates for prediction.

We propose a new method to simultaneously select covariates and estimate their corresponding varying coefficients as follows. Assume that varying coefficients $\beta_1(\cdot), \beta_2(\cdot), \ldots, \beta_p(\cdot)$ reside in a reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \|\cdot\|_{\mathcal{H}_K})$ with the reproducing kernel $K(\cdot,\cdot)$ (Wahba, 1990). Find $\beta_1(\cdot), \beta_2(\cdot), \ldots, \beta_p(\cdot) \in \mathcal{H}_K$ and $b \in \mathbb{R}$ to minimize

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{\nu=1}^{m_i}\left[y_{i\nu} - b - \sum_{j=1}^{p}\beta_j(t_{i\nu})x_{ij}(t_{i\nu})\right]^2 + \lambda\sum_{j=1}^{p}\|\beta_j\|_{\mathcal{H}_K}, \tag{3.3}$$

where $N = \sum_{i=1}^{n} m_i$ and $\|\cdot\|_{\mathcal{H}_K}$ is the RKHS norm. The first term in (3.3) measures the goodness of data fitting and the second term merits the selection property by the l_1-like penalty $\sum_{j=1}^{p}\|\beta_j\|_{\mathcal{H}_K}$. We first provide the following theorem to justify the existence of minimizer for (3.3).

Theorem 3.1. *There exists a minimizer of (3.3) that is in the domain $\beta_1(\cdot), \ldots, \beta_p(\cdot) \in \mathcal{H}_K$ and $b \in \mathbb{R}$.*

The proof of this theorem is given in Appendix B.3. The variable selection method (3.3) is new in the literature and (3.3) is efficient for optimization since it is convex in $\beta_j(\cdot)$s and it has only one tuning parameter λ. We provide an algorithm in Appendix B.4.

The following theorem gives further insights into (3.3) that it is indeed a combination of the smoothing splines (Wahba, 1990) and the Lasso (Tibshirani, 1996).

Theorem 3.2. *Consider the following optimization problem. Find $\beta_1(\cdot), \ldots, \beta_p(\cdot) \in \mathcal{H}_K$ and $\theta_1, \ldots, \theta_p, b \in \mathbb{R}$ to minimize*

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{\nu=1}^{m_i}[y_{i\nu} - b - \sum_{j=1}^{p}\beta_j(t_{i\nu})x_{ij}(t_{i\nu})]^2 + \tau_0\sum_{j=0}^{p}\theta_j^{-1}\|\beta_j\|_{\mathcal{H}_K}^2 + \tau_1\sum_{j=0}^{p}\theta_j, \tag{3.4}$$
$$s.t.\ \theta_j \geq 0, j = 0, 1, \ldots, p,$$

where τ_0 is a constant and τ_1 is a tuning parameter. Let $\tau_1 = \lambda^4/(4\tau_0)$. The following equivalence holds.

1) *If $(\widehat{\beta}_0, \widehat{\beta}_1(\cdot), \ldots, \widehat{\beta}_p(\cdot))$ minimizes (3.3), by letting $\widehat{\theta}_j = \tau_0^{1/2}\tau_1^{-1/2}\|\widehat{\beta}_j\|_{\mathcal{H}_K}$, we have that $(\widehat{\theta}_1, \ldots, \widehat{\theta}_p; \widehat{\beta}_0, \widehat{\beta}_1(\cdot), \ldots, \widehat{\beta}_p(\cdot))$ minimizes (3.4).*

2) *If there exists $(\widehat{\theta}_1, \ldots, \widehat{\theta}_p; \widehat{\beta}_0, \widehat{\beta}_1(\cdot), \ldots, \widehat{\beta}_p(\cdot))$ minimizes (3.4), then $(\widehat{\beta}_0, \widehat{\beta}_1(\cdot), \ldots, \widehat{\beta}_p(\cdot))$ minimizes (3.3).*

We give the proof of this theorem in Appendix B.5. Note that (3.4) is a combination of the smoothing splines and the Lasso since the first two terms:

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{\nu=1}^{m_i}[y_{i\nu} - b - \sum_{j=1}^{p}\beta_j(t_{i\nu})x_{ij}(t_{i\nu})]^2 + \tau_0\sum_{j=0}^{p}\theta_j^{-1}\|\beta_j\|_{\mathcal{H}_K}^2 + \tau_1\sum_{j=0}^{p}\theta_j$$

is actually the same as the smoothing splines in nonparametric statistics (Wahba, 1990), and the last term

$$\tau_1\sum_{j=0}^{p}\theta_j$$

is actually the same as the Lasso penalty (Tibshirani, 1996) for the weights θ_js.

Let $X_{j_1}, X_{j_2}, \ldots, X_{j_s}$ be s of selected features by (3.3), $1 \le j_1 \le j_2 \le \cdots \le j_s \le p$, and $\widehat{\beta}_{j_1}, \widehat{\beta}_{j_2}, \ldots, \widehat{\beta}_{j_s}$ be the corresponding estimated varying coefficients by (3.3). Then the prediction model for a new subject with features $X_{j_1}^*(t), X_{j_2}^*(t), \ldots, X_{j_s}^*(t)$ at time t is

$$\widehat{f}^*(t) = \widehat{\beta}_{j_1}X_{j_1}^*(t) + \widehat{\beta}_{j_2}X_{j_2}^*(t) + \cdots + \widehat{\beta}_{j_s}X_{j_s}^*(t).$$

3.2.1 Dataset for experiments

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging

(MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). More descriptions on the ADNI database is in Appendix B.1.

3.3 Experiment Results

We corroborate our method by predicting future clinical changes of MCI subjects. Generally, some MCI subjects will convert into AD after certain time (i.e., MCI converters, MCI-C for short), while others will not convert (i.e., MCI non-converters, MCI-NC for short) Zhang et al. (2012). The prediction of clinical change in a MCI subject help to determine whether the subject will convert into AD at a future time point, which is a central task for the early diagnosis of AD. We summarize the baseline demographic information of ADNI subjects studied here in Table 3.1.

Table 3.1: Demographics of ADNI subjects studied here

	MCI-C ($n = 74$)	MCI-NC ($n = 98$)
Male/Female	44 / 30	61 / 37
Age (years)	73.03 ± 6.65	74.35 ± 7.47
Edu. (years)	15.51 ± 3.05	15.59 ± 3.07

The preprocessing steps for brain MR imaging used here are described in Appendix B.2. Specifically, we have total 324 ROIs for each imaging. For MCI subjects, MRI scans were performed at baseline (bl), 6 months (M06), one year (M12), 18 months (M18), two years (M24), three years (M36), and four years (M48). However, some subjects may miss a few visit times and hence they do not have MRI scans at these time points. We choose $n = 172$ MCI subjects who have M48 imaging data. Table 3.2 lists the distributions of visit times for these 172 MCI subjects, where, e.g., 6 of MCI-C subjects make at most 3 visits

among the scheduled six times (bl, M06, M12, M18, M24, M36) such that they have at most 3 longitudinal MRI scans.

Table 3.2: Distribution of visit times for ADNI subjects studied here

	MCI-C ($n = 74$)	MCI-NC ($n = 98$)
≤ 3 scans	6	6
4 scans	8	14
5 scans	15	33
6 scans	45	45

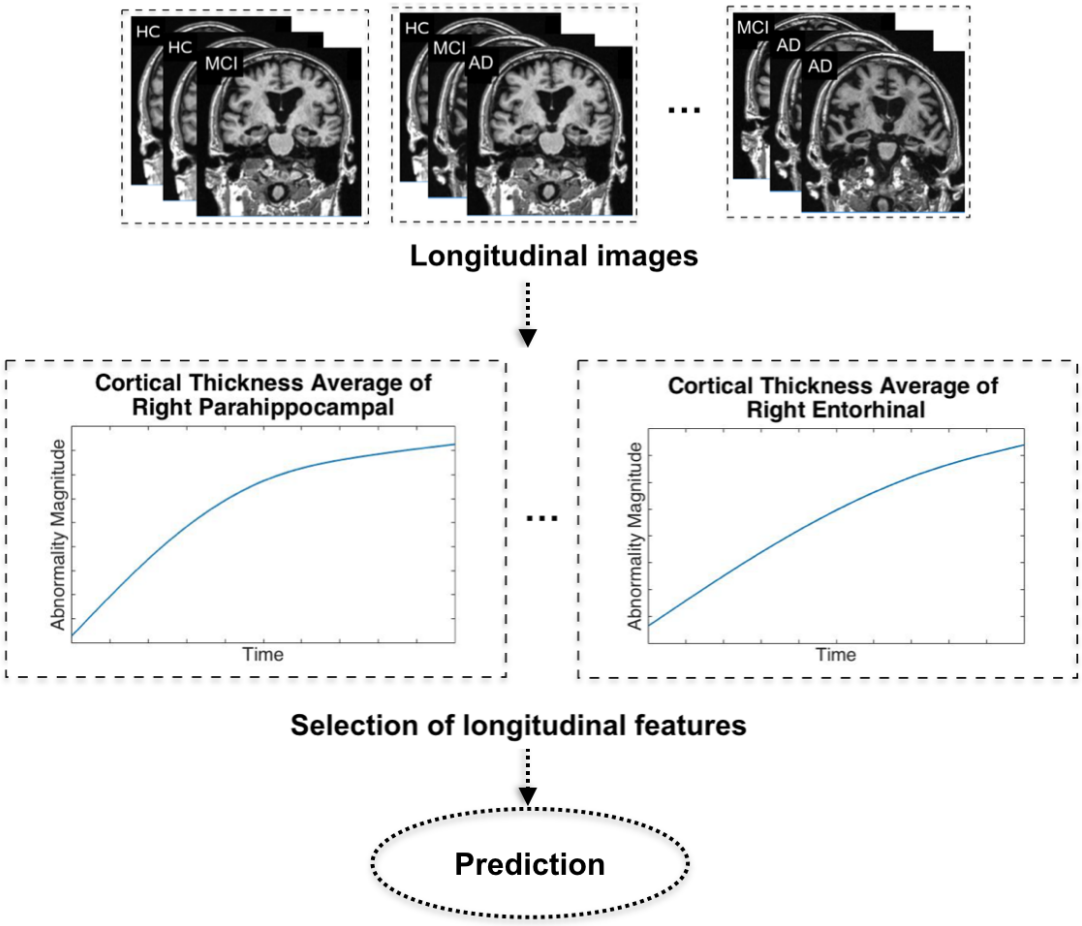

Figure 3.2: Flowchart of the proposed method.

Our goal is to use longitudinal information (from bl up to M36) to predict the clinical changes of MCI subjects at M48. Since the empirical evidences suggest that the rates of change over time for structural MRI and clinical cognition functions are in a temporally ordered manner (Jack Jr et al., 2010, 2013), a nonlinear modeling for the functional relationship between the atrophy of MRI and the change in clinical cognition functions is necessary. Hence, the varying coefficient model (3.1) is used. We choose the Alzheimer's Disease Assessment Scale – Cognitive Subscale (ADAS-Cog) as the response clinical cognitive test score $Y(\cdot)$ and it ranges from 70 (severe cognitive impairment) to 0 (no cognitive impairment). The ADAS-Cog measures disturbances of memory, language, and other cognitive abilities. The covariates $X_j(\cdot)$s include 324 MR imaging ROIs and 3 demographic covariates: age, gender, and education years. The index t in (3.1) should be identifiable and we let t be the scaled time relative to subjects enter the ADNI study. We normalize the time to the unit interval $[0, 1]$. Figure 3.2 gives the flowchart of our method, where the abnormality magnitude measures the shrinkage of a feature by comparing the average of normal subjects that progressed to AD over time relative to the average of normal subjects that did not progress to AD over time with ADNI dataset. For example, Figure 3.2 shows the *thickness of right parahippocampal cortex* and *thickness of right entorhinal cortex* significantly decrease over time for subjects progressed to AD compared to subjects did not progress. These two features are selected by our method for prediction.

We build six models by using six different levels of longitudinal information:

- Model 1: bl.
- Model 2: bl+M06 (including subjects have missings at bl).
- Model 3: bl+M06+M12 (including subjects have missings at bl or M06).
- Model 4: bl+M06+M12+M18 (including subjects have missings at bl, M06 or M12).
- Model 5: bl+M06+M12+M18+M24 (including subjects have missings at bl, M06, M12 or M18).

- Model 6: bl+M06+M12+M18+M24+M36 (including subjects have missings at bl, M06, M12, M18, or M24).

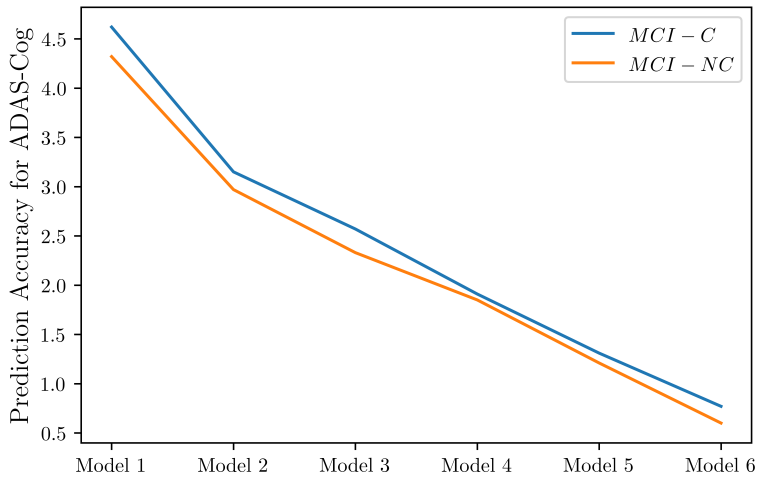

Figure 3.3: The prediction comparisons of our method using six levels of longitudinal data.

Following the flowchart in Figure 3.2, we first perform the feature selection method in (3.3) for each of the six models. In each experiment, we randomly leave out half of samples in both MCI-C and MCI-NC for prediction. For the training of each model, a 10-fold cross validation is performed to select the tuning parameter λ in (3.3). The experiments are replicated for 100 times. We summarized the mean squared prediction accuracy in Figure 3.3. It is clear that the longitudinal data can significantly improve the prediction results compared with only using baseline information. And the more longitudinal data included, the better prediction will be obtained. We also observe that the prediction results for MCI-NC are slightly better compared with MCI-C, which can be explained by the fact that MCI-NC subjects have more stable clinical status and less varied clinical scores.

We give examples of selected feature in Figure 3.4. These are four ROIs that consistently selected in Model 6 for 100 experiments. Figure 3.4 demonstrates the varying coefficients of the ROIs. Specifically, gender is an important factor and different ROIs have different functional relations with clinical functions (i.e., the maximum effect of each biomarker

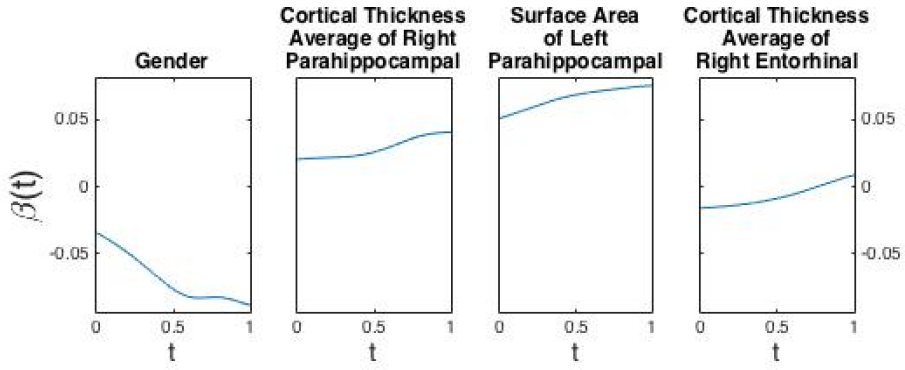

Figure 3.4: Examples of selected features for Model 6.

varies over the course of disease progression). This confirms the evidence and hypothesis in Sabuncu et al. (2011); Schuff et al. (2012) that atrophy does not affect all regions of the brain simultaneously, but perhaps in a sequential manner.

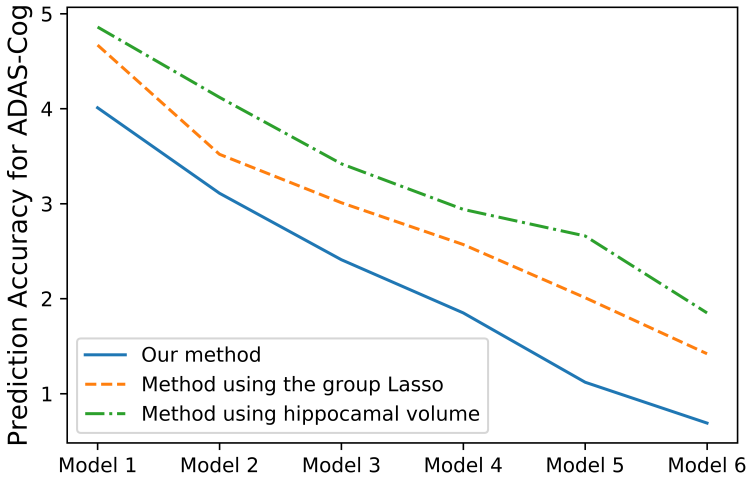

Figure 3.5: The prediction comparisons of three methods for MCI-C.

Now we compare our method (3.3) with other two state-of-the-art methods:

- The longitudinal analysis in Chincarini et al. (2016) which only uses the hippocampal volume shrinkage rate as the feature.
- The longitudinal analysis in Zhang et al. (2012) which use linear feature representations

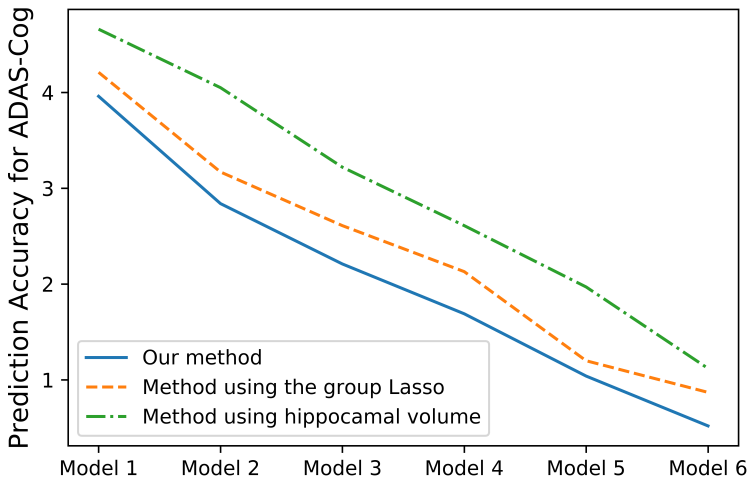

Figure 3.6: The prediction comparisons of three methods for MCI-NC.

and a group Lasso for variable selection (e.g., Yuan and Lin (2006)).

Since the methods in Chincarini et al. (2016); Zhang et al. (2012) require same scanning times and a same number of scans across samples, we perform Model 1–6 for AD prediction with samples having no missing visits. In each experiment, we randomly leave out half of samples in both MCI-C and MCI-NC for prediction. For the training of each model, a 10-fold cross validation is performed to select the tuning parameters in (3.3) and in Chincarini et al. (2016); Zhang et al. (2012). The experiments are replicated for 100 times. The mean squared prediction accuracy for MCI-C are summarized in Figure 3.5 and the mean squared prediction accuracy for MCI-NC are summarized in Figure 3.6. It is clear that our proposed method consistently achieves better prediction performances for both MCI-C and MCI-NC. The reason of the superior performance of our method is due to the modeling of nonlinear progression of longitudinal features and selecting important features from the whole brain instead of only using a prespecified feature for prediction.

Chapter 4

Selection and Estimation Optimality in High Dimensions with the TWIN Penalty

4.1 Introduction

Discovering relevant relationships between a large number of variables and an outcome continues to be an eminently challenging problem in statistics and a major interest in a wide variety of scientific disciplines. Decades of research has focused on variable selection techniques to identify relevant variables. Among these techniques, penalized regression-based methods such as the Lasso (Tibshirani, 1996), smoothly-clipped absolute deviation (SCAD) (Fan and Li, 2001), and the minimax concave penalty (MCP) (Zhang, 2010) have been widely explored, as they often perform well in practice, have computational advantages, and possess desirable variable selection properties. However, selection consistency results for penalized methods often require the imposition of relatively extreme levels of sparsity on the data generating mechanism and thus may not accurately describe real world data. For example, when modeling health outcomes of patients, such as hospitalization risk or human phenotypes, the relevant risk factors may be highly varied and numerous. As human biology

is extraordinarily complex, it is sensible that more relevant predictors may be included when an increasing amount of genetic or microbiome information is leveraged, especially when considering gene-gene, gene-environment, or microbiome-environment interactions (Nadeau and Topol, 2006; Martin et al., 2007; Bull and Plummer, 2014; Shreiner et al., 2015). As such, methodological and theoretical advances in variable selection commensurate with this possibility are needed.

In this paper we seek to address this gap with a novel class of penalties. The proposed penalty class results in estimators that are provably selection consistent and asymptotically minimax in high-dimensional scenarios under linear sparsity and relatively weak assumptions regarding the data-generating mechanism. We call our penalty class the two mountains penalty class, or TWIN (**T**Wo mounta**I**Ns) for short, as the shape of the penalty function resembles two mountains centered around the origin. The general shape of the two mountains penalty class makes it amenable to controlling the false discovery rate of variable selections (FDR) while retaining high power of selection and is thus instrumental to its desirable selection properties. Furthermore, the shape of TWIN penalty functions, illustrated in Figure 4.1a, results in sensible data-adaptive penalization where larger coefficients are subjected to attenuated penalization. Throughout this paper we show that this general pattern of penalization yields advantageous selection and estimation properties. Extensive simulations buttress our theoretical results and demonstrate the superior finite sample selection and estimation properties of our penalty in scenarios with strong correlations between relevant and irrelevant variables.

The core of this paper centers around the ubiquitous linear model, which posits that the relationship between a set of predictors and a response variable has the following linear form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}, \tag{4.1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is a vector of responses, $\boldsymbol{X} \equiv (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ is a random matrix with each column representing samples of a particular predictor, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ is a

vector which relates the predictors to a mean response value, and $\boldsymbol{z} \sim N(0, \sigma^2 \boldsymbol{I}_n)$ is an error term independent of $\boldsymbol{X}$. We adopt the familiar penalized regression framework, wherein sparse estimates $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ are achieved by minimizing a penalized least squares objective with penalty $P(\cdot)$:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}||^2 + \sum_{j=1}^{p} P(|b_j|) \right\}. \tag{4.2}$$

The Lasso falls under this framework with $P(|b|) = |b|$. The focus of this paper is on a new class of penalty functions $P(\cdot)$, which will be introduced in Section 4.2.

We highlight three main contributions of this work:

1. We propose a novel class of penalty functions for variable selection, which provide data-adaptive penalization in a manner which results empirically in favorable selection and prediction performance. We provide two examples of the penalty class which are amenable to computationally efficient algorithms.

2. We provide selection consistency results for the proposed class of penalty functions in both the high dimensional ($p > n$) and low-dimensional settings under linear sparsity. Similar to SLOPE (Bogdan et al., 2015), our penalty admits a finite sample bound for the FDR under orthogonality and is thus a candidate for future study of FDR control under more general designs.

3. We establish new minimax optimal risk under the linear sparsity. Moreover, we show that TWIN estimators are minimax optimal for both orthogonal and random designs.

The remainder of this paper is organized as follows. We introduce our proposed class of penalty functions in Section 4.2. In Section 4.3 we study the key selection properties of the TWIN penalty and in Section 4.4 we present minimax optimality results. Section 5.3 investigates the numerical properties of the TWIN penalty in comparison with other standard penalties using extensive simulation studies. In Section 4.6 we analyze a microarray study relating gene expression levels to a phenotype in mice with the TWIN penalty.

4.2 Methodology

4.2.1 The TWIN penalty class and examples

The "two mountains" penalty class is defined by a general shape, which has the appearance of two mountains centered around the origin. Figure 4.1a depicts the archetypal shape of TWIN with two examples of the penalty class in comparison with the shapes of the Lasso penalty and the MCP. The motivation of the two mountains shape is clear: it has a singularity at zero, thus allowing for variable selection, and it penalizes small coefficients more heavily and relaxes the amount of penalization for large coefficients, effectuating the idea that variables with larger coefficients are more likely to be related to our response. Thus, it provides data-adaptive penalization of coefficients. However, the relationship between the magnitude of penalization is not monotone with coefficient size, as it is potentially unreasonable to assume that all small coefficients are necessarily unimportant.

The TWIN penalty class $P_{\lambda,\tau}(t)$ is indexed by two parameters $\lambda, \tau > 0$ and satisfies the following criteria:

1. $P_{\lambda,\tau}(t)$ is continuous and nonnegative for $t \in \mathbb{R}^+$ with $P_{\lambda,\tau}(0) = 0$;
2. $\sup_{\lambda>0} P_{\lambda,\tau}(t) = \infty$ for any $t \neq 0$;
3. The derivative of the penalty is continuous except at the origin and satisfies
    - $P'_{\lambda,\tau}(0+) = \lambda$, which enables the selection of variables,
    - $P'_{\lambda,\tau}(t)$ is positive for $0 < t < \tau$ and decreases to 0 such that $P'_{\lambda,\tau}(\tau) = 0$,
    - $P'_{\lambda,\tau}(t)$ is nonpositive for $t > \tau$, first decreasing in a neighborhood after τ and then increasing to 0, yielding a "coefficient enlargement" effect for a range of t and (near) unbiasedness for large t,

When $P_{\lambda,\tau}$ is a member of the TWIN class, we call the minimizer of (4.2) a TWIN estimator. Penalties that meet all of the two mountains (TWIN) criteria resemble two symmetrical hill or mountain shapes centered around 0 when taken as a function of $|t|$. The tuning

parameter τ specifies the precise location of the peaks of the "mountains", i.e. where the penalty achieves its maximum value. The second criteria above guarantees that adjusting λ will eventually result in a large enough penalty to set any coefficient to zero. The third property in criterion 3 above results in what we call coefficient enlargement in the sense that some estimates are slightly biased away from zero; see Figures 4.1c and 4.2. The TWIN class can be further delineated based on the limiting behavior of $P_{\lambda,\tau}(t)$. The first subclass of TWIN penalties, which we call TWIN-a, is defined as all TWIN penalties which only achieve zero derivative in the limit. The second subclass, TWIN-b, has derivative equal to zero for all $t \geq d$ for some constant $d > 0$. This distinction results in different properties and our theoretical derivations will handle them separately.

The pattern of decreased penalization for $t > \tau$ is inspired by multiple testing procedures, wherein smaller p-values are compared with lower thresholds, for example Benjamini and Hochberg (1995). From the regression point of view (assuming equal variance of each coefficient estimate), smaller p-values correspond to stronger signals, i.e. variables with larger regression estimates. Thus the behavior of TWIN is opposite that of another recently proposed data-adaptive penalty, SLOPE (Bogdan et al., 2015), which penalizes coefficients whose estimates are larger more heavily than those whose estimates are smaller.

In the following we introduce two specific TWIN penalties that will be used throughout this paper for demonstration purposes. While the theoretical results in this paper apply to all TWIN penalties, our numerical examples and our data analysis focus on the following two specific penalties in the TWIN class.

Example 4.1 (TWIN-a)**.**

$$P_{\lambda,\tau}(t) = \begin{cases} \lambda c(1 - (1 - t/\tau)^2) & t \leq m_1\tau \\ \lambda c d_1 \tau / t & t > m_1\tau \end{cases}, \tag{4.3}$$

where $d_1 > 0$ and $m_1 > 0$ are calculated such that the function above is continuous and has matching derivatives at m_1 and c is a normalizing constant defined such that $P'_{\lambda,\tau}(0+) = \lambda$. The term c can

be dropped for clarity or ease of implementation. A direct calculation shows that $d_1 = 32/27$ *and* $m_1 = 4/3$*. Note that letting* $\tau \to 0$ *and* $\lambda\tau \to 1/(cd_1)$ *yields* $P_{\lambda,\tau}(t) = 1/t$*, which is the reciprocal Lasso of Song and Liang (2015).*

Example 4.2 (TWIN-b)**.**

$$P_{\lambda,\tau}(t) = \begin{cases} \lambda c(1-(1-t/\tau)^2) & t \le m_2\tau \\ \lambda c[(t-d_2)^2/\tau^2 + h] & m_2\tau < t < d_2 \\ \lambda ch & t > d_2 \end{cases}, \tag{4.4}$$

where $h \in (0,1)$ *and* $d_2 > 0$*,* $m_2 > 1$ *are calculated such that the function above is continuous and has matching derivatives at* $m_2\tau$ *and* d_2 *and again* c *is a normalizing constant defined such that* $P'_{\lambda,\tau}(0+) = \lambda$*. A straightforward calculation shows that* $d_2 = (1+\sqrt{2(1-h)})\tau$ *and* $m_2 = 1+\sqrt{(1-h)/2}$*. The parameter* h *can be chosen to balance convexity of the penalty, and hence computational stability, with effect enlargement, however we simply choose* $h = 1/2$*.*

Examples 2.1 and 2.2 differ only in their behavior for $t > \tau$.

Remark 4.3. *If* $\tau \to \infty$ *and* $\lambda c/\tau \to \lambda^*/2$*, both TWIN-a and TWIN-b become the Lasso penalty with tuning parameter* λ^**.*

To better understand the behavior of TWIN penalties, let us consider the following univariate penalized least squares problem

$$\frac{1}{2}(z-\theta)^2 + P_{\lambda,\tau}(|\theta|). \tag{4.5}$$

Fan and Li (2001) note that a good penalty function should meet three key criteria, namely i) (near) unbiasedness ii) sparsity, and iii) continuity of the minimizer of (4.5) with respect to z. TWIN meets the first two criteria, however, like for the hard-thresholding function (Antoniadis, 1997; Fan, 1997) and for the reciprocal Lasso (Song and Liang, 2015), it does not always meet the third. Specifically, for a range of values of τ, the minimizer of (4.5)

is not continuous in z; see Figure 4.1c. Thus, in some sense, the tuning parameter τ of TWIN offers a trade-off between continuity and computational stability. In spite of added computational instability, we find that TWIN with values of τ resulting in a discontinuous estimator often performs remarkably well in practice. Both examples TWIN-a (Example 2.1) and TWIN-b (Example 2.2) are computationally convenient, because they both admit closed-form solutions for univariate (4.5), allowing for faster coordinate-descent algorithms with simple updates.

Figure 4.2 displays the regularization paths of the Lasso, SCAD, MCP, TWIN-a and TWIN-b penalties from a simulated dataset with $n = 200$, $p = 1000$ among which only 10 active variables are related to the response, the covariates are generated independently from $N(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma}_{ij} = 0.5^{|i-j|}$, and $\boldsymbol{z} \sim N(0, I_n)$. The coefficients for the 10 active variables are given by $(-1/2, 2/3, -5/6, 1, -7/6, 4/3, -3/2, 5/3, -11/6, 2)$. The horizontal gray dashed lines are the oracle least squares estimates for the 10 active variables. Due to the low sample size, correlations between inactive variables and the response range between -0.21 and 0.22. The correlations between active variables and the response range in magnitude from 0.07 to 0.45 and are thus often dominated by random correlations with the response. Due in part to these correlations, the Lasso selects multiple inactive variables early on in the regularization path, a phenomenon studied rigorously in Su et al. (2017). Note that TWIN results in estimates which are inflated for a range of λ. Due to the fact that the derivative of the TWIN-a penalty is never exactly zero, it results in increased coefficient enlargement compared with TWIN-b. As we justify in Section 4.2.3, this added enlargement effect may be more beneficial in scenarios with strong correlations between covariates. Smaller coefficients, however, can still receive shrinkage towards zero by TWIN depending on the value of τ. This behavior can be helpful in scenarios where prediction is a priority.

4.2.2 Heuristics of TWIN

In this subsection, based on heuristic arguments, we provide insights into why the TWIN estimator yields reduced false discoveries compared with the Lasso, SCAD and MCP. The

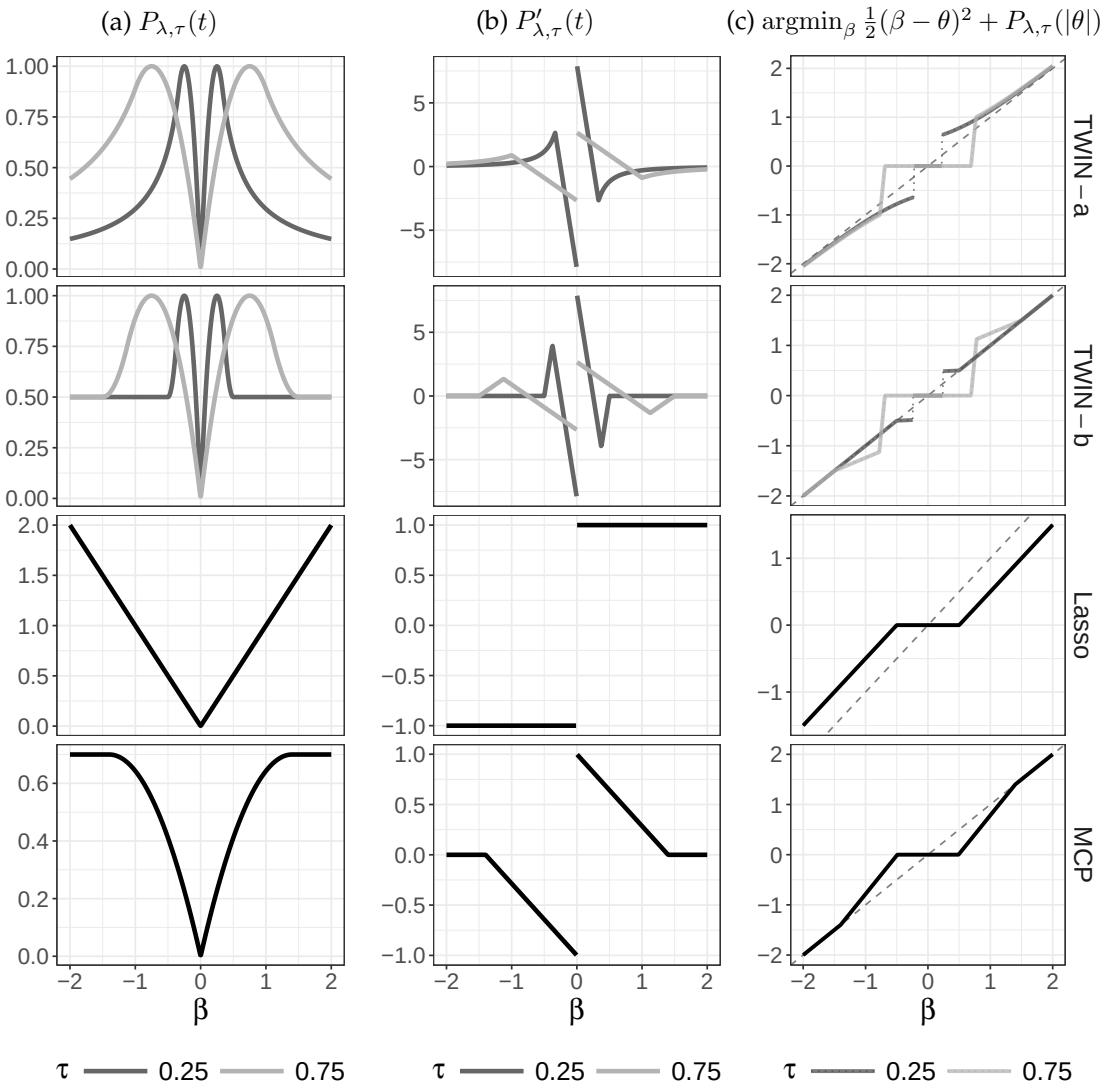

Figure 4.1: Panel (a) compares the penalty functions for TWIN-a and TWIN-b with the Lasso and MCP all with with $\lambda = 1$ (and $\lambda c = 1$ in the case of TWIN). The extra tuning parameter γ for MCP is set to 1.4. Panel (b) compares the corresponding derivative functions. Panel (c) compares the thresholding functions for all of the penalties.

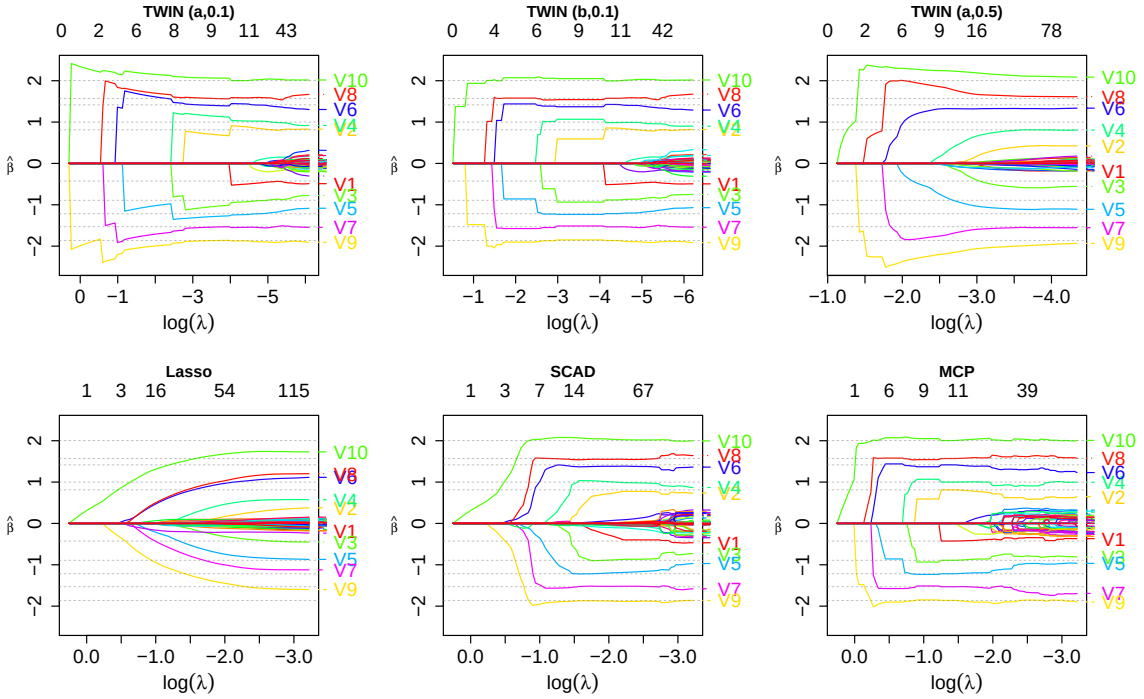

Figure 4.2: Plot of coefficient paths as the λ tuning parameter is varied for TWIN-a and -b in comparison with that of the Lasso, SCAD, and MCP. The top left plot is TWIN-a with $\tau = 0.1$, the top middle is TWIN-b with $\tau = 0.1$, and the top right is TWIN-a with $\tau = 0.5$. Only variables $V1 - V10$ have nonzero coefficients in this example and only these variables are labeled on the right of each plot if selected.

arguments in this section roughly follow and extend the arguments in Su et al. (2017). For simplicity, in this section we fix $\sigma = 0$ as the following can be extended to cases with noise. Consider a Gaussian random design matrix $\boldsymbol{X}$ which has i.i.d. $N(0, 1/n)$ entries and consider an oracle TWIN estimator with known true support $A^o = \{j : \beta_j \neq 0\}$ as obtained by

$$\widehat{\boldsymbol{\beta}}_{A^o} = \underset{\boldsymbol{b}_{A^o} \in \mathbb{R}^{\epsilon p}}{\operatorname{argmin}} \frac{1}{2} ||\boldsymbol{y} - \boldsymbol{X}_{A^o} \boldsymbol{b}_{A^o}||^2 + \sum_{j \in A^o} P_{\lambda,\tau}(|b_j|), \tag{4.6}$$

where A^o is of approximate size $\epsilon p, 0 < \epsilon < 1$, and $n, p \to \infty$. The matrix $\boldsymbol{X}_{A^o}$ is comprised of columns indexed by A^o from the full design matrix $\boldsymbol{X}$. If $|\mathbb{E}_{\boldsymbol{\beta}_{A^o}}[\boldsymbol{x}_i'(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}_{A^o})]| \leq \lambda$ for all $i \in \bar{A}^o$, where $\bar{A}^o = \{1, \ldots, p\} \backslash A^o$, the KKT condition (4.12) suggests in expectation that

extending $\widehat{\boldsymbol{\beta}}_{A^o}$ by adding zeros to $\bar{A}^o$ results in a solution of (4.2). If for some $j \in \bar{A}^o$,

$$|\mathbb{E}_{\boldsymbol{\beta}_{A^o}}[\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}_{A^o})]| > \lambda, \tag{4.7}$$

then we must consider the reduced problem (4.6) with support $A^o \cup \{j\}$ instead of A^o in order to yield an equivalent solution with (4.2). Hence, (4.7) provides evidence of false discoveries. Since $\widehat{\boldsymbol{\beta}}_{A^o}$ is independent of $\boldsymbol{X}_{\bar{A}^o}$, by conditioning on $\boldsymbol{X}_{A^o}$, $\mathbb{E}_{\boldsymbol{\beta}_{A^o}}[\boldsymbol{x}_j'(\boldsymbol{y}-\boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}_{A^o})]$ is normally distributed with mean zero and variance $n^{-1}\|\mathbb{E}_{\boldsymbol{\beta}_{A^o}}[\boldsymbol{X}_{A^o}(\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o})]\|^2$.

To compare TWIN with the Lasso, observe that when $n > k$, the largest singular value of $\boldsymbol{X}_{A^o}(\boldsymbol{X}_{A^o}'\boldsymbol{X}_{A^o})^{-1}$ is bounded, thus with probability approaching one,

$$\begin{aligned}
&n^{-1}\|\mathbb{E}_{\boldsymbol{\beta}_{A^o}}[\boldsymbol{X}_{A^o}(\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o})]\|^2 \\
=&n^{-1}\|\boldsymbol{X}_{A^o}(\boldsymbol{X}_{A^o}'\boldsymbol{X}_{A^o})^{-1}\mathbb{E}_{\boldsymbol{\beta}_{A^o}}[\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_{A^o})P_{\lambda,\tau}'(|\widehat{\boldsymbol{\beta}}_{A^o}|)]\|^2 \\
\leq& c_0 n^{-1}\left\{\lambda^2 \#\{j \in A^o, |\mathbb{E}_{\beta_j}[\widehat{\beta}_j]| < \gamma\lambda\} + \sup_{t \geq \gamma\lambda} |P_{\lambda,\tau}'(t)|^2 \#\{j \in A^o, |\mathbb{E}_{\beta_j}[\widehat{\beta}_j]| \geq \gamma\lambda\}\right\},
\end{aligned} \tag{4.8}$$

where c_0 is some constant and γ is defined in (4.13) which indicates the region where $P_{\lambda,\tau}'$ is approximately zero. For Lasso estimators, we know $P'(\cdot) \equiv \lambda$ and thus the right-hand side of (4.8) is of order λ when $|A^o|$ is linear in p. In other words, Lasso estimators satisfy (4.7) for a number of variables in $\bar{A}^o$ linear in p, which causes a non-vanishing false discovery proportion; see Su et al. (2017). TWIN estimators, however, yield (near) unbiasedness, which results in $\sup_{t \geq \gamma\lambda} |P_{\lambda,\tau}'(t)|^2$ close to 0. If the distribution of $\boldsymbol{\beta}_{A^o}$ is such that the minimal absolute value of true coefficients is larger than a certain threshold with a large probability (as in, e.g., Tibshirani (2011)), then $\#\{j \in A^o, |\mathbb{E}_{\beta_j}[\widehat{\beta}_j]| < \gamma\lambda\}/n \to 0$ and thus the right-hand side of (4.8) approaches 0 for TWIN estimators, resulting in a vanishing proportion of false discoveries.

To compare TWIN with SCAD and MCP, we note that although these penalties are all (nearly) unbiased, TWIN penalties possess an *enlargement* property for estimates with absolute values of a middling range; see, Figure 4.1 for illustration. The enlargement property

can compensate in some sense for the shrinkage error of estimates near zero. Specifically, we can bound the left-hand side of (4.8) as follows:

$$\begin{aligned} n^{-1}\|\mathbb{E}_{\boldsymbol{\beta}_{A^o}}[\boldsymbol{X}_{A^o}(\boldsymbol{\beta}_{A^o}-\widehat{\boldsymbol{\beta}}_{A^o})]\|^2 &= n^{-1}\|\boldsymbol{X}_{A^o}(\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}\mathbb{E}_{\boldsymbol{\beta}_{A^o}}[\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_{A^o})P'_{\lambda,\tau}(|\widehat{\boldsymbol{\beta}}_{A^o}|)]\|^2 \\ &\leq c_1 n^{-1}\|\mathbb{E}_{\boldsymbol{\beta}_{A^o}}[P'_{\lambda,\tau}(|\widehat{\boldsymbol{\beta}}_{A^o}|)]\|^2 \end{aligned} \tag{4.9}$$

for some constant $c_1 \geq 0$. Since SCAD, MCP and TWIN yield shrinkage for weak signals, $P'(|\widehat{\beta}_j|) > 0$ for small $\widehat{\beta}_j$. However, the enlargement property of TWIN enables $P'_{\lambda,\tau}(|\widehat{\beta}_j|) < 0$ for β_j with middling magnitudes, which compensates for positive $P'_{\lambda,\tau}(|\widehat{\beta}_j|)$'s and results in a smaller bound in (4.9). Thus for $j \in \bar{A}^o$, the conditional variance of $\mathbb{E}_{\boldsymbol{\beta}_{A^o}}[\boldsymbol{x}'_j(\boldsymbol{y}-\boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}_{A^o})]$ has a smaller upper bound for TWIN, implying that TWIN is likely to give a smaller proportion of false discoveries than SCAD and MCP. Moreover, it is evident from extensive simulations in Section 5.3 that TWIN can be significantly better than SCAD and MCP in the linear sparsity regime with strong positive and negative correlations between inactive and active variables.

4.2.3 The role of the tuning parameter τ

TWIN's tuning parameter τ has an important impact on the selection behavior of TWIN. We note that the reciprocal Lasso may yield overly sparse solutions when the underlying truth is not extremely sparse, and the Lasso may over-select variables when the underlying solution is indeed quite sparse. The tuning parameter τ balances between these two extremes. As τ tends to 0 and to ∞, TWIN becomes the reciprocal Lasso and the Lasso, respectively, allowing for a dynamic range of selection behavior. We now conduct a simulation study to investigate the finite sample properties of TWIN as τ is varied. Data are generated under model (4.1) where the data-generating setup is described in Section 5.3 and the coefficients in the linear model are generated as described in Model 3 in Section 5.3. We evaluate selection performance by investigating the average FDR versus true discovery rate of variable selection (TDR) curves as the tuning parameter λ is varied. The curves are displayed in Figure 4.3.

Generally, smaller values of τ tend to result in better selection characteristics as λ is

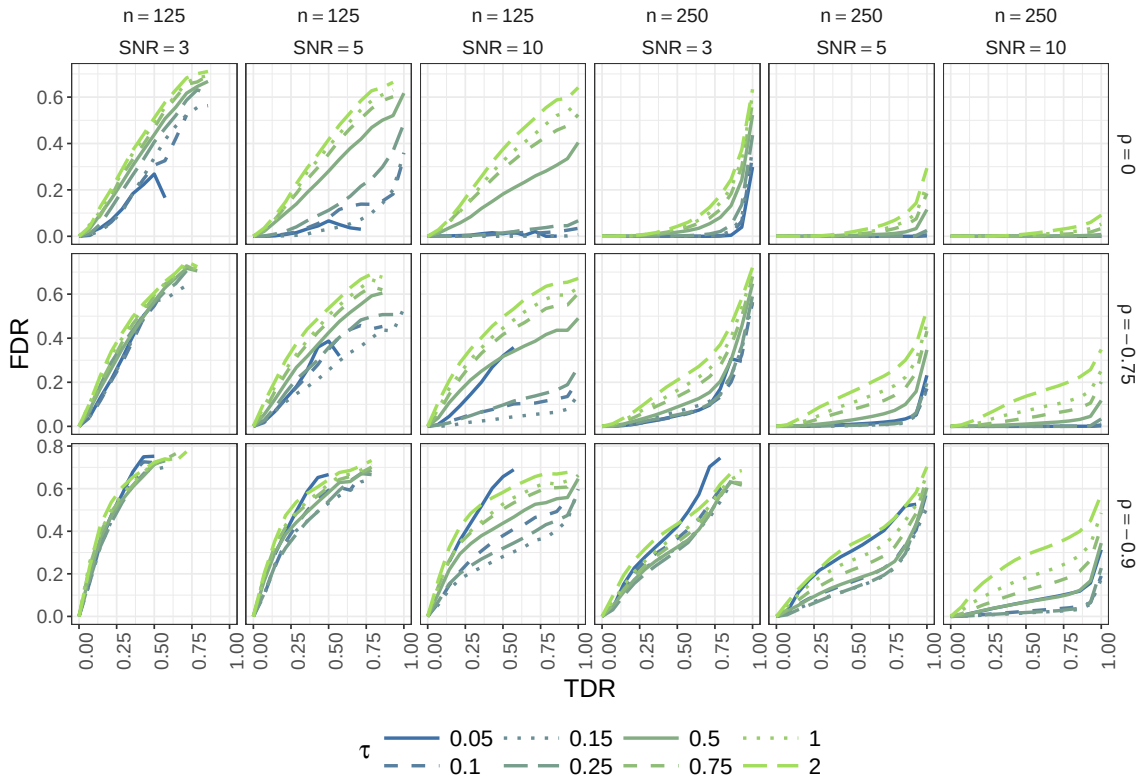

Figure 4.3: The results above are for a simulation with data generated under Model 3 described in Section 5.3. Models are fit using the TWIN-a penalty.

varied, however this comes at a cost of computational instability. The smallest value of τ considered works well in low correlation settings, but poorly with high correlations and when many covariates are selected. Slightly larger values of τ such as 0.25 to 0.75 tend to have better performance in low signal settings with high correlations. Over all settings, including a more complete set of simulations presented in the Supplementary Material, values of 0.1 and 0.15 tend to work the best. However, in practice, it may be the case that τ must be increased or decreased to some degree for ideal performance. In the Supplementary Material we further investigate the role of τ on prediction performance. The message is similar for prediction, however in scenarios with very low signal, larger values of τ are preferable if prediction is the primary goal. As τ increases, the model which minimizes the mean squared prediction error tends to be larger in size. It is important to bear in mind that these investigations only span a small number of possibilities and thus do not always reflect

how selection and estimation performance vary with τ.

4.3 Selection Properties

In this section we investigate the selection properties of TWIN estimators. In particular, we show that TWIN is selection consistent when a non-vanishing fraction of variables are important. Further, TWIN yields a finite sample FDR bound under orthogonal designs. We also provide universal values for both tuning parameters λ and τ for which the selection consistency results hold. For low-dimensional regimes, these values do not depend on any unknown quantities other than the noise level. We begin by studying the selection properties for orthogonal designs and then extend these results to random Gaussian designs. Hereafter, we denote $\widehat{\boldsymbol{\beta}}$ as a TWIN estimator (distinctions between TWIN-a and TWIN-b will be made when warranted), $\boldsymbol{\beta}$ as the true coefficient vector, and

$$\widehat{A} \equiv \{j : \widehat{\beta}_j \neq 0\}, \quad A^o \equiv \{j : \beta_j \neq 0\}, \quad \text{and} \quad k \equiv |A^o| = \#\{j : \beta_j \neq 0\}. \tag{4.10}$$

4.3.1 Orthogonal designs

To gain insights about the TWIN estimator, we first consider orthogonal designs. Under orthogonality, the optimality conditions for TWIN results in the following thresholding rule as the solution to

$$\widehat{\boldsymbol{\beta}} = \text{sgn}(\boldsymbol{X}'\boldsymbol{y}) \left(|\boldsymbol{X}'\boldsymbol{y}| - P'_{\lambda,\tau}(|\widehat{\boldsymbol{\beta}}|) \right)_+ ,$$

where the sign function $\text{sgn}(t) \equiv I\{t > 0\} - I\{t < 0\}$. See Figure 4.1c for an illustration. We note that when $|\widehat{\beta}_j| > \tau$, the absolute value of the resulting estimator is larger than the absolute value of the data. We call this effect the enlargement property since TWIN amplifies estimates for moderately large $|\beta_j|$. However, TWIN yields (nearly) unbiased estimates for sufficiently large $|\beta_j|$. This overall behavior is different from the "unbiasedness" property of SCAD (Fan and Li, 2001) and MCP (Zhang, 2010), and is also different from the "shrinkage" property of the Lasso. We now present an upper bound of the FDR of TWIN

under orthogonal designs.

Proposition 4.4. *Suppose that the data are generated from the linear model* (4.1) *with an orthogonal design* $\boldsymbol{X}$ *and* $\boldsymbol{z} \sim N(0, \sigma^2 I_p)$. *Then for any* $\alpha \in [0, 1]$ *the false discovery rate (FDR) and the family-wise error rate (FWER) for TWIN estimators obey,*

$$FDR = \mathbb{E}\left[\frac{\#\{j \in \widehat{A} \backslash A^o\}}{|\widehat{A}| \vee 1}\right] \leq \alpha\left(1 - \frac{k}{p}\right), \;\; FWER = \mathbb{P}\left\{\exists j \in \widehat{A} \backslash A^o\right\} = \alpha,$$

by choosing

$$\min_{t \in \mathbb{R}} \{|t| + P'_{\lambda,\tau}(|t|)\} = \sigma \Phi^{-1}(1 - \alpha/2p). \tag{4.11}$$

If there are multiple pairs of (τ, λ) *satisfying (4.11), we select the pair resulting in the largest number of selected variables so as to increase power.*

There are significant challenges in showing similar finite sample bounds for TWIN with a random design due to the estimation error of regression coefficients. See, for example, Bogdan et al. (2015). Instead, we show that the FDR asymptotically approaches zero in Theorem 4.9.

4.3.2 Random designs

In this section we study the selection properties of TWIN under random Gaussian designs where the columns of $\boldsymbol{X}$ have i.i.d. $N(0, 1/n)$ entries so that the columns are approximately normalized. Random designs are widely utilized in the statistics literature for studying regression methods. See, for example, Candès et al. (2006); Zou (2006); Meinshausen and Yu (2009); Van de Geer and Bühlmann (2009); Su and Candès (2016). Such designs are a sensible starting point for theoretical analysis of model selection properties due to weak correlations between the different predictors, as they obey restricted isometry properties (Candès and Tao, 2005) or restricted eigenvalue conditions (Bickel et al., 2009) with high probability. However, based on our numerical experiments, we suspect similar results may hold for designs with significant correlations and we leave this for future work.

The rest of this section is organized as follows. We first introduce main assumptions in Section 4.3.2.1 and then provide probability bounds of correct selection for TWIN in two cases: the global minimizer of (4.1) in the regular case where rank$(\boldsymbol{X}) = p$ in Section 4.3.2.2 and the local solution in the degenerate case where rank$(\boldsymbol{X}) < p$ in Section 4.3.2.3.

4.3.2.1 Working assumptions and linear sparsity

We assume throughout Section 4.3.2 that $p, n \to \infty$ and $n/p \to \delta$ for some constant $\delta > 0$. Further, as in Su et al. (2017), we assume that $\beta_1, \ldots, \beta_p$ are independent copies of a random variable Π which satisfies $\mathbb{E}\Pi^2 < \infty$ and $\mathbb{P}(\Pi \neq 0) = \epsilon$ where $\epsilon \in (0, 1)$ is some constant. Hence, our assumptions accommodate linear sparsity where the expected value of k equals to $\epsilon \cdot p$. An asymptotic regime such as is discussed in Wainwright (2009), among other works, where the proportion of nonzero coefficients vanishes in the limit of p does not allow for linear sparsity. As noted in Su et al. (2017), studying penalized regression methods in the linear sparsity regime yields theoretical results which accurately describe variable selection and estimation performance across a wide range of practical settings, as it can accommodate scenarios with relatively high dimension and a moderately low level of sparsity in addition to scenarios with very sparse signals. See Bayati and Montanari (2012); Su et al. (2017) for extended discussion on the merits of the linear sparsity assumption.

For notational simplicity, we consider in Section 4.3.2 and Section 4.4 that $\min_{t\in\mathbb{R}}\{|t| + P'_{\lambda,\tau}(|t|)\} = P'_{\lambda,\tau}(0+) = \lambda$, however the results in these two sections can be straightforwardly generalized to the case $0 < \min_{t\in\mathbb{R}}\{|t| + P'_{\lambda,\tau}(|t|)\} < \lambda$. A TWIN estimator $\widehat{\boldsymbol{\beta}}$ follows

$$\begin{cases} \boldsymbol{x}'_j(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \operatorname{sgn}(\widehat{\beta}_j)P'_{\lambda,\tau}(|\widehat{\beta}_j|), & \widehat{\beta}_j \neq 0, \\ |\boldsymbol{x}'_j(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})| \leq \lambda, & \widehat{\beta}_j = 0. \end{cases} \tag{4.12}$$

Equations (4.12) are the Karush-Kuhn-Tucker (KKT) conditions for the global minimization of (4.2). In general, solutions of (4.12) include all local minimizers of (4.2).

4.3.2.2 Probability bounds for selection consistency

We first provide probability bounds for selection consistency when $n > p$ and n and p both tend to infinity. To clarify the distinction between TWIN-a and TWIN-b members of the TWIN class and to aid the presentation of theoretic results, we introduce an additional parameter γ that describes the limiting behavior of $P'_{\lambda,\tau}(t)$ as follows:

$$P'_{\lambda,\tau}(t) \begin{cases} < 0 \text{ and } |P'_{\lambda,\tau}(t)| = o(\lambda), & \text{when } t \geq \gamma\lambda, \text{ for TWIN-a;} \\ = 0, & \text{when } t \geq \gamma\lambda, \text{ for TWIN-b.} \end{cases} \tag{4.13}$$

In particular, TWIN-b becomes flat beyond a certain region while TWIN-a only has a 0 derivative beyond a certain range in the limit; see the illustration in Figure 4.1b. We consider the TWIN-a and TWIN-b variants of TWIN separately, as they exhibit slightly different behavior. Recall that our theoretical exposition applies to all TWIN-a and TWIN-b penalties, not just the specific examples introduced in Section 4.2.1. We first present a non-asymptotic bound for selection consistency with TWIN-a penalties.

Theorem 4.5. *Suppose that $n > p$, $\widehat{A}$ and A^o are defined in (4.10). Let $\widehat{\boldsymbol{\beta}}$ be the TWIN-a estimator in (4.2) for $\lambda \geq \{[(1-\vartheta)\sqrt{\delta/\epsilon} - 1]^{-1}(1+\vartheta) + 1\}(1+\vartheta)\sigma\sqrt{2\log p}$ and $\tau \geq (1 - \delta^{-1/2} - \vartheta)^{-2}\lambda$ with any $\vartheta > 0$. Then if $|\beta_j| > \gamma\lambda + \sigma\sqrt{(2+4\vartheta)\log k}(1 - \epsilon^{1/2}\delta^{-1/2} - \vartheta)^{-1}$ for all $j \in A^o$, we have*

$$\begin{aligned} \mathbb{P}\left\{\widehat{A} \neq A^o\right\} &\leq \mathbb{P}\left\{\widehat{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}^o \text{ or } sgn(\widehat{\boldsymbol{\beta}}) \neq sgn(\boldsymbol{\beta})\right\} \\ &\leq e^{-n\sigma^2\vartheta^2/2} + e^{-k\vartheta^2/2} + 3e^{-n\vartheta^2/2} + \sqrt{\pi\vartheta}k^{-\vartheta}. \end{aligned}$$

In particular for large n, TWIN-a can arbitrarily control both type I and type II errors to low levels under the linear sparsity regime, which yields $\mathbb{P}\{\widehat{A} = A^o\} \to 1$.

Corollary 4.6. *Suppose that $n > p$ and $\epsilon \leq 0.25$. Let $\widehat{\boldsymbol{\beta}}$ be the TWIN-a estimator in (4.2) for $\lambda_{a,univ} = (1 + \delta^{-1/2})\sigma\sqrt{2\log p}$ and $\tau_{a,univ} = (0.99 - \delta^{-1/2})^{-2}\lambda_{a,univ}$. Then if $|\beta_j| \geq \gamma\lambda_{a,univ} + \sigma\sqrt{2\log k}(1 - \epsilon^{1/2}\delta^{-1/2})^{-1}$ for all $j \in A^o$, $\mathbb{P}\{\widehat{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}^o \text{ or } sgn(\widehat{\boldsymbol{\beta}}) \neq sgn(\boldsymbol{\beta})\} \to 0$.*

The universal parameters $\lambda_{a,\text{univ}}$ and $\tau_{a,\text{univ}}$ do not require knowledge of the sparsity

level. The condition $\epsilon \leq 0.25$ is only a technical requirement for the proof, however, it is a reasonable assumption in many applications. Now, we consider the TWIN-b penalty and provide a similar non-asymptotic bound for its selection consistency.

Theorem 4.7. *Suppose that $n > p$, $\widehat{A}$ and A^o are defined in (4.10). Let $\widehat{\boldsymbol{\beta}}$ be the TWIN-b estimator in (4.2) for $\lambda \geq (1+3\vartheta)\sqrt{1-\epsilon\delta^{-1}}\sigma\sqrt{2\log p}$ and $\tau \geq (1-\delta^{-1/2}-\vartheta)^{-2}\lambda$ with any $\vartheta > 0$. Then if $|\beta_j| > \gamma\lambda + \sigma\sqrt{(2+4\vartheta)\log k}(1-\epsilon^{1/2}\delta^{-1/2}-\vartheta)^{-1}$ for all $j \in A^o$, we have*

$$\begin{aligned}\mathbb{P}\left\{\widehat{A} \neq A^o\right\} &\leq \mathbb{P}\left\{\widehat{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}^o \text{ or } sgn(\widehat{\boldsymbol{\beta}}) \neq sgn(\boldsymbol{\beta})\right\} \\ &\leq e^{-\vartheta^2(n-k)\sigma^2/2} + 2e^{-n\vartheta^2/2} + \sqrt{\pi\vartheta}(p-k)^{-\vartheta} + \sqrt{\pi\vartheta}k^{-\vartheta}.\end{aligned}$$

In particular for large n, TWIN-b can arbitrarily control both type I and type II errors to low levels under the linear sparsity regime, which yields $\mathbb{P}\{\widehat{A} = A^o\} \to 1$.

Corollary 4.8. *Suppose that $n > p$. Let $\widehat{\boldsymbol{\beta}}$ be the TWIN-b estimator in (4.2) for $\lambda_{b,univ} = \sigma\sqrt{2\log p}$ and $\tau_{b,univ} = (0.99-\delta^{-1/2})^{-2}\lambda_{b,univ}$. Then if $|\beta_j| \geq \gamma\lambda_{b,univ} + \sigma\sqrt{2\log k}(1-\epsilon^{1/2}\delta^{-1/2})^{-1}$ for all $j \in A^o$, we have $\mathbb{P}\{\widehat{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}^o \text{ or } sgn(\widehat{\boldsymbol{\beta}}) \neq sgn(\boldsymbol{\beta})\} \to 0$.*

Similar to Corollary 4.6, the universal parameters $\lambda_{b,\text{univ}}$ and $\tau_{b,\text{univ}}$ do not require knowledge of the sparsity level. Extensive simulation studies demonstrating the effectiveness of the universal parameters and extended discussion on handling unknown noise level are presented in the Supplementary Material.

4.3.2.3 Selection consistency for high-dimensional regression

Now we consider the high-dimensional case where $p > n$ and $k < n$ and show the selection consistency of TWIN. For brevity, we only present results for TWIN-b as the following theorem can be generalized to the TWIN-a similarly as Section 4.3.2.2.

Theorem 4.9. *Suppose that $p > n$, $\widehat{A}$ and A^o are defined in (4.10). Let $\widehat{\boldsymbol{\beta}}$ be the TWIN-b estimator in (4.2) for $\lambda \geq \max\{(1+3\vartheta)\sqrt{1-\epsilon\delta^{-1}}\sigma\sqrt{2\log p}, 2[1+\vartheta+\sqrt{(\epsilon/\delta+1)/2}]\sigma\sqrt{2\tilde{c}+1}\}$ and $\tau \geq (1-\sqrt{(\epsilon/\delta+1)/2}-\vartheta)^{-2}\lambda$ with any $\vartheta > 0$ and $\tilde{c} \equiv [(1-\epsilon)\log(1-\epsilon) - (\delta-\epsilon)\log(\delta-$*

$\epsilon) - (1-\delta)\log(1-\delta)]/\delta$. *Then if* $|\beta_j| > \gamma\lambda + \sigma\sqrt{(2+4\vartheta)\log k}(1-\epsilon^{1/2}\delta^{-1/2}-\vartheta)^{-1}$ *for all* $j \in A^o$ *and* $\epsilon/\delta \leq 0.12$, *we have*

$$\begin{aligned}\mathbb{P}\left\{\widehat{A} \neq A^o\right\} &\leq \mathbb{P}\left\{\widehat{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}^o \text{ or } sgn(\widehat{\boldsymbol{\beta}}) \neq sgn(\boldsymbol{\beta})\right\} \\ &\leq e^{-\vartheta^2(n-k)\sigma^2/2} + 2e^{-n\vartheta^2/2} + \sqrt{\pi\vartheta}(p-k)^{-\vartheta} \\ &\quad + \sqrt{\pi\vartheta}k^{-\vartheta} + \left\{[\tilde{c} + (n-k)^{-1}]\sqrt{2\pi(n-k)}\right\}^{-1}.\end{aligned}$$

Corollary 4.10. *Suppose that* $p > n$. *Let* $\widehat{\boldsymbol{\beta}}$ *be the TWIN-b estimator in (4.2) for* $\lambda_{b,univ} = \sigma\sqrt{2\log p}$ *and* $\tau'_{univ} \geq [0.99 - \sqrt{(\epsilon/\delta+1)/2}]^{-2}\lambda_{b,univ}$. *Then if* $|\beta_j| \geq \gamma\lambda_{b,univ} + \sigma\sqrt{2\log k}(1 - \epsilon^{1/2}\delta^{-1/2})^{-1}$ *for all* $j \in A^o$ *and* $\epsilon/\delta \leq 0.12$, *we have* $\mathbb{P}\{\widehat{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}^o \text{ or } sgn(\widehat{\boldsymbol{\beta}}) \neq sgn(\boldsymbol{\beta})\} \to 0$.

The parameter $\lambda_{b,\text{univ}}$ is the same as in Corollary 4.8 and does not require knowledge about the sparsity level. For τ'_{univ} to avoid a requirement of exact knowledge of the sparsity level, we can use a prior upper bound on ϵ, denoted by ϵ', and set $\tau'_{\text{univ}} = [0.99 - \sqrt{(\epsilon'/\delta+1)/2}]^{-2}\lambda_{b,\text{univ}}$, which satisfies the condition of Corollary 4.10.

Theorem 4.9 and Corollary 4.10 show that in the case of high-dimensionality and linear sparsity, TWIN estimators have false discovery rate and true discovery rate (TDR) obeying

$$\lim_{n\to\infty} \text{FDR} = \lim_{n\to\infty} \mathbb{E}\left[\frac{\#\{j \in \widehat{A} \backslash A^o\}}{|\widehat{A}| \vee 1}\right] = 0, \quad \lim_{n\to\infty} \text{TDR} = \lim_{n\to\infty} \mathbb{E}\left[\frac{\#\{j \in \widehat{A} \cap A^o\}}{k \vee 1}\right] = 1.$$

Theorem 4.9 also implies that $n = (\delta/\epsilon + o(1))k > 8.33k$ is sufficient for perfect recovery. It is known in the compressed sensing literature that in the no noise case, n Gaussian samples with $n \geq 2(1+o(1))k\log(p/k) = 2(1+o(1))k\log(1/\epsilon)$ are required for perfect support recovery using l_1-based methods; see, e.g., Donoho and Tanner (2010). Stricter conditions are usually assumed in the statistics literature for perfect recovery, for example, $k/p \to 0$ in Song and Liang (2015) and $(k\log p)/n \to 0$ in Su and Candès (2016).

4.4 Estimation Properties

In this section, we investigate the minimax optimality of estimation with TWIN estimators under random Gaussian designs and linear sparsity. In the Supplementary Material, we present corresponding results for minimax optimality under orthogonal designs. As noted in the literature (Su and Candès, 2016), minimax optimality results for orthogonal designs do not in general imply similar results for Gaussian designs because of the sample correlations among the columns of Gaussian designs. The goal of this section is to establish the minimax optimality of TWIN estimators under Gaussian designs and linear sparsity.

4.4.1 Risk lower bound under linear sparsity

The following result gives an explicit lower bound of asymptotic risk under the linear sparsity and random Gaussian designs.

Theorem 4.11. *Suppose that $k/p \to \epsilon \in (0,1)$ as $p \to \infty$. Let $\boldsymbol{\beta}$ be from the model (4.1) and the columns of $\boldsymbol{X}$ have i.i.d. $N(0,1/n)$ entries. Then for any $\vartheta \in (0,1)$, we have*

$$\infty_{\widetilde{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left\{ \frac{\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2\sigma^2 k \log(1/\epsilon)} > 1 - \vartheta \right\} = 1,$$

where the infimum is taken over all measurable estimators.

Similar results for random designs can be found in the literature; see, for example, Ye and Zhang (2010); Raskutti et al. (2011); Su and Candès (2016). However, the main difference of such results and Theorem 4.11 is that instead of assuming $k/p \to 0$ and $(k \log p)/n \to 0$, Theorem 4.11 considers the linear sparsity regime $k/p \to \epsilon$ with unknown constant $\epsilon \in (0,1)$ and provides the exact constant in front of the rate.

4.4.2 Risk upper bounds for TWIN estimators

We first give a probabilistic bound on the asymptotic risk for TWIN-a estimators.

Theorem 4.12. *Suppose that $p, n \to \infty$ with $n/p \to \delta$ for some constant $\delta > 1$ and $k/p \to \epsilon$ for some constant $0 < \epsilon < 1$. Let $\widehat{\boldsymbol{\beta}}$ be the TWIN-a estimator in (4.2) for $\lambda = \{[(1-\vartheta)\sqrt{\delta/\epsilon} - 1]^{-1}(1+\vartheta) + 1\}(1+\vartheta)\sigma\sqrt{2\log p}$ and $\tau \geq (1 - \delta^{-1/2} - \vartheta)^{-2}\lambda$ for arbitrary $\vartheta > 0$. Then,*

$$\sup_{\|\boldsymbol{\beta}\|_0 \leq k} \mathbb{P}\left\{ \frac{\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2}{C_1(\epsilon, \delta) \cdot 2\sigma^2 k \log p} \leq 1 \right\} \to 1, \tag{4.14}$$

where the constant $C_1(\epsilon, \delta) = \left\{ \frac{\sqrt{3}}{[(1-\lambda\tau^{-1})\delta^{1/2}\epsilon^{-1/2} - 2]^{1/2}} + 1 \right\}^2 [(\delta^{1/2}\epsilon^{-1/2} - 1)^{-1} + 1]^2$.

We make the following remarks on the above theorem. First, comparing Theorem 4.12 with the lower bound result Theorem 4.11, there is a difference in their logarithm terms, which is actually due to the unknown sparsity level. More specifically, when k is unknown, a tight upper bound for $1/\epsilon$ is p. Hence, TWIN-a estimators are minimax rate optimal. Second, $C_1(\epsilon, \delta)$ is close to one when ϵ is small, which meets the constant in Theorem 4.11. Third, we have shown in Corollary 4.6 that universal tuning parameters $\lambda_{a,\text{univ}}$ and $\tau_{a,\text{univ}}$ yield selection consistency. The following result shows further that these universal tuning parameters yield asymptotic estimation risk with the minimax optimal rate.

Corollary 4.13. *Suppose that $p, n \to \infty$ with $n/p \to \delta$ for some constant $\delta > 1$ and $k/p \to \epsilon$ for some constant $0 < \epsilon \leq 0.25$. Let $\widehat{\boldsymbol{\beta}}$ be the TWIN-a estimator in (4.2) for $\lambda_{a,univ} = (1 + \delta^{-1/2})\sigma\sqrt{2\log p}$ and $\tau_{a,univ} = (0.99 - \delta^{-1/2})^{-2}\lambda_{a,univ}$. Then, $\sup_{\|\boldsymbol{\beta}\|_0 \leq k} \mathbb{P}\{\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2 / [C_1'(\epsilon, \delta) \cdot 2\sigma^2 k \log p] \leq 1\} \to 1$ with constant $C_1'(\epsilon, \delta) = \{\sqrt{3}/[(1.98 - \delta^{-1/2})\epsilon^{-1/2} - 2]^{1/2} + 1\}^2(1 + \delta^{-1/2})^2$.*

A similar probabilistic bound on the asymptotic risk holds for TWIN-b estimators.

Theorem 4.14. *Suppose that $p, n \to \infty$ with $n/p \to \delta$ for some constant $\delta > 1$ and $k/p \to \epsilon$ for some constant $0 < \epsilon < 1$. Let $\widehat{\boldsymbol{\beta}}$ be the TWIN-b estimator in (4.2) for $\lambda = (1+3\vartheta)\sqrt{1 - \epsilon\delta^{-1}}\sigma\sqrt{2\log p}$ and $\tau \geq (1 - \delta^{-1/2} - \vartheta)^{-2}\lambda$ for arbitrary $\vartheta > 0$. Then,*

$$\sup_{\|\boldsymbol{\beta}\|_0 \leq k} \mathbb{P}\left\{ \frac{\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2}{C_2(\epsilon, \delta) \cdot 2\sigma^2 k \log p} \leq 1 \right\} \to 1, \tag{4.15}$$

where the constant $C_2(\epsilon, \delta) = \left[\frac{\sqrt{3}}{(\delta^{1/2}\epsilon^{-1/2}-2)^{1/2}} + 1\right]^2 (1 - \epsilon\delta^{-1})$.

We make the following remarks on the above theorem. First, similar to the discussion after Theorem 4.12, TWIN-b estimators are minimax rate optimal. Second, $C_2(\epsilon, \delta)$ is close to one when ϵ is small, which also meets the constant in Theorem 4.11. Third, we note $C_1(\epsilon, \delta) > C_2(\epsilon, \delta)$, which implies TWIN-b estimators achieve a smaller upper bound of asymptotic risk than TWIN-a estimators when $\epsilon > 0$. Heuristically, this is due to the unbiasedness property of the TWIN-b estimators, whereas TWIN-a estimators are only nearly unbiased and often result in stronger enlargement effects. Fourth, Corollary 4.8 shows that universal tuning parameters $\lambda_{b,\text{univ}}$ and $\tau_{b,\text{univ}}$ yield selection consistency and now the following result shows they also yield the minimax optimal rate.

Corollary 4.15. *Suppose that* $p, n \to \infty$ *with* $n/p \to \delta$ *for some constant* $\delta > 1$ *and* $k/p \to \epsilon$ *for some constant* $0 < \epsilon < 1$. *Let* $\widehat{\boldsymbol{\beta}}$ *be the TWIN-b estimator in (4.2) for* $\lambda_{b,univ} = \sigma\sqrt{2\log p}$ *and* $\tau_{b,univ} = (0.99 - \delta^{-1/2})^{-2}\lambda_{b,univ}$. *Then,* $\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\{\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2/[C_2'(\epsilon, \delta) \cdot 2\sigma^2 k \log p] \le 1\} \to 1$ *with constant* $C_2'(\epsilon, \delta) = [\sqrt{3}/(\delta^{1/2}\epsilon^{-1/2} - 2)^{1/2} + 1]^2$.

Finally, we remark that results in Theorem 4.12 and 4.14 can be generalized to the high-dimensional case where $p > n$ and $k < n$ as in Section 4.3.2.3.

4.5 Numerical Studies

In this section we seek to demonstrate the variable selection properties of the TWIN penalty under various challenging and realistic high dimensional scenarios. In this section we simulate data under model (4.1) where the number of non-zero elements in $\boldsymbol{\beta}$ is very small relative to the dimension p. We generate $\boldsymbol{X}$ from a multivariate Gaussian distribution with covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ with $\boldsymbol{\Sigma}_{ij} = \rho^{|i-j|}$. Larger $|\rho|$ indicates stronger correlations between predictors. The correlation parameter ρ is varied from $(0, -0.75, -0.90)$, the sample size is set to 125 and 250, and p is set to 1000. We focus on $\rho \le 0$, as most data for regression tasks exhibit both positive and negative correlations. We set the variance of the error term

such that the signal-to-noise ratio (SNR), defined as $SNR = \sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}}/\sigma$, is 3, 5, and 10. Given the number of active variables in the models considered below, this range of the signal-to-noise ratio makes it very difficult to recover the active variables. In all of the above settings the k active coefficients are chosen uniformly at random from all p covariates with magnitudes of the active coefficients generated under the following two schemes: i.) independent random variates from a uniform distribution on $[-2, 0.5] \cup [0.5, 2]$ and ii.) $(-c)^{j-1}$ for the jth of k active variables. Under Models 1 and 3, we generate coefficients from scheme i.) with $k = 50$ and $k = 25$, respectively, and under Models 2 and 4 we generate coefficients from scheme ii.) with $c = 0.95$ and $c = 0.8$, respectively, and $k = 50$ and $k = 25$, respectively. The beta-min condition is not satisfied under scheme ii.), as the smallest nonzero coefficients are close to 0 and much smaller than the largest coefficients, whereas under scheme i.) coefficients are bounded away from 0.

We compare TWIN-a and TWIN-b with the Lasso, SCAD, and MCP. We use the `R` package `ncvreg` (Breheny and Huang, 2011) to implement SCAD and MCP and use the `R` package `glmnet` (Friedman et al., 2010) to implement the Lasso. Throughout the simulations, we set the γ tuning parameter for MCP to be 1.4 as recommended in Zhang (2010) and for SCAD to be 3.7, as recommended in Fan and Li (2001). The bandwidth tuning parameter τ of TWIN-a and TWIN-b is set to be 0.1 throughout the simulations. In the Supplementary Material we introduce two algorithms for computation for the TWIN penalty. The first algorithm is a modification of coordinate descent and is denoted as CD and the second algorithm is a hybrid local linear approximation (Zou and Li, 2008) and coordinate descent algorithm, which we denote as MCLLA for mixed coordinate local linear approximation. We investigate the performance of TWIN using both CD and MCLLA using random coordinate updates instead of cyclical updates, as described in the Supplementary Material.

As we wish to understand the underlying operating characteristics of all methods with respect to FDR and TDR, we evaluate each method by investigating the relationship between FDR and TDR as the selection tuning parameter λ is varied. In Figures 4.4 and 4.5 we display average FDR versus TDR curves under Models 1-4 averaged over 100 independent datasets.

To demonstrate predictive performance under the same simulation settings, we display in Figure 4.6 the average square root of the mean squared prediction error (RMSE) versus the number of nonzero coefficients for each method. Due to space concerns, prediction results for Models 3 and 4 are included in the Supplementary Material. The RMSE is evaluated on an independent dataset of size 5000. The independent dataset is generated anew for each replication of the simulation study.

We first evaluate the variable selection results. In settings with more active variables (Models 1 and 2), both TWIN-a and TWIN-b outperform all other methods when there are correlations between covariates. In the no correlation setting ($\rho = 0$), TWIN-a and TWIN-b both outperform SCAD and the Lasso, but have similar albeit slightly worse performance than MCP in high SNR and/or sample size settings. However, TWIN-a and TWIN-b tend to perform better than MCP in most low-signal and/or low sample size settings. In settings with 25 active variables (Models 3 and 4), the comparisons are similar, except SCAD performs nearly as well as TWIN-a and TWIN-b when $\rho = -0.9$ under Model 3 and slightly better than TWIN-a and TWIN-b when $\rho = -0.9$ under Model 4.

Regarding prediction performance, we first consider results under Models 1 and 2. In low SNR settings, the Lasso and SCAD tend to perform the best, with the Lasso achieving the smallest minimum RMSE, albeit with models which are on average much larger than models which minimize RMSE under different penalties. Like MCP and unlike SCAD and the Lasso, both TWIN-a and TWIN-b tend to achieve their minimum RMSE with models that are of approximately the correct size of the underlying data-generating model. In high correlation settings and large signals, TWIN tends to have the best minimum RMSE of all methods including the Lasso.

Comparing the MCLLA and CD algorithms for TWIN-a and TWIN-b, we find that MCLLA tends to outperform CD in small sample size settings, however when the sample size is larger, CD performs better. This trend holds in additional simulation studies presented in the Supplementary Material. In the Supplementary Material we present results with $p = 2000$ under Models 1 and 2 and under two similar models with an increased number of

active variables ($k = 100$). The results are quite similar, further substantiating our theoretical results.

4.6 Analysis of Polymerase Chain Reaction (PCR) Study

Lan et al. (2006) conducted an experiment to investigate the relationship between gene expression and gene function in mice. In the study gene expression levels were measured on 22,575 genes of 29 male and 31 female mice using Affymetrix MOE430 microarrays. To examine gene function, three phenotypes phosphoenopyruvate carboxykinase (PEPCK), glycerol-3-phosphate acyltransferase (GPAT), and stearoyl-CoA desaturase 1 (SCD1) were measured for each of the mice by quantitative real-time PCR. The data are publicly available from the Gene Expression Omnibus (GEO) project (`http://www.ncbi.nlm.nih.gov/geo` via accession number GSE3330).

For ease of presentation we restrict our focus to analysis of the SCD1 phenotype, which is a key enzyme in the metabolism of fatty acids. As there is no natural validation data available for this study, we compare different methods by repeatedly drawing random splits of the 60 samples into 55 training samples and 5 testing samples. As a preprocessing step we take a log transformation of the gene expression levels. Using each comparator method we fit a model predicting SCD1 using all 22,575 gene expression levels. The sample correlations of the design matrix range from -0.83 to 0.99 with 10th and 90th quantiles of -0.22 and 0.24, respectively. Each method is evaluated by the average out-of-sample mean squared prediction error (MSPE) on the testing samples ($MSPE = S^{-1}\sum_{s=1}^{S}\sum_{i\in I_{test,s}}(y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}_{train,s})^2/|I_{test,s}|$, where $I_{test,s}$ are the indices of the testing samples for the sth replication and $\widehat{\boldsymbol{\beta}}_{train,s}$ is an estimate of $\boldsymbol{\beta}$ using the training samples from the sth replication). We repeat this procedure $S = 100$ times. We consider the Lasso, MCP, SCAD, and TWIN penalties in our analysis and for all methods use 10-fold cross validation for selection of the tuning parameter λ. The additional tuning parameters for all methods were chosen as described in Section 5.3. Due to the small sample size, we utilize the MCLLA algorithm for

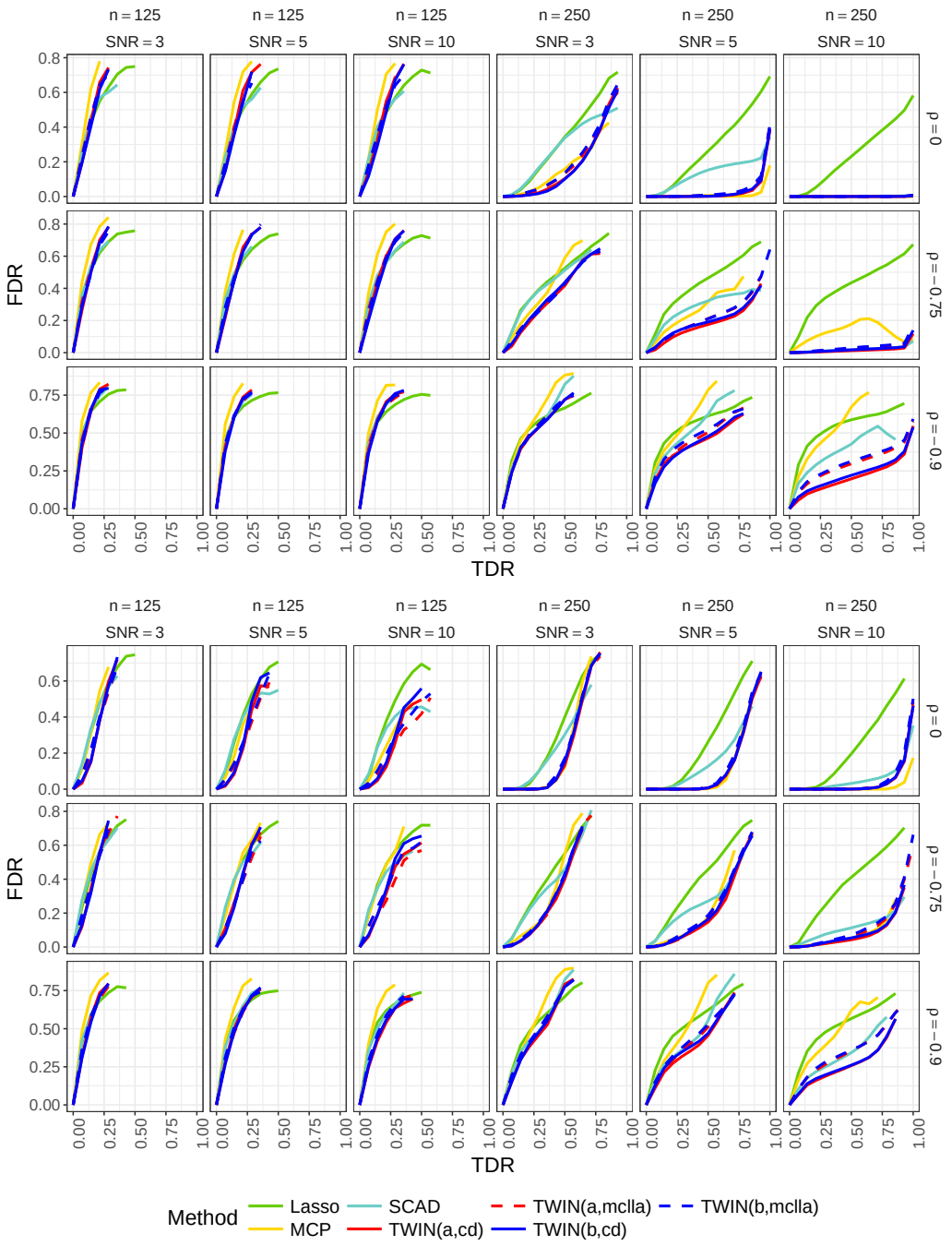

Figure 4.4: The results above are for a simulation with data generated under Model 1 (top panel) and Model 2 (bottom panel).

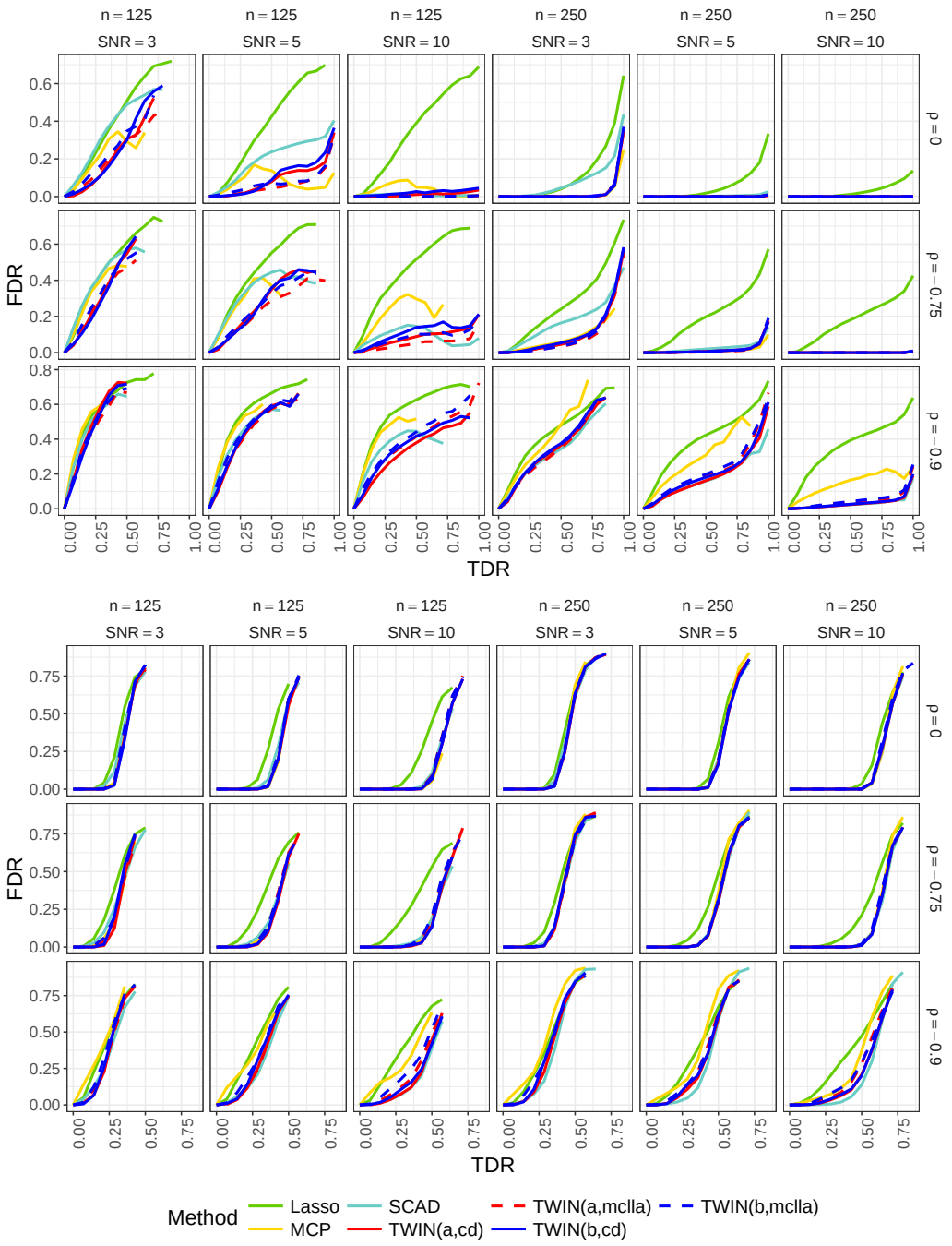

Figure 4.5: The results above are for a simulation with data generated under Model 3 (top panel) and Model 4 (bottom panel).

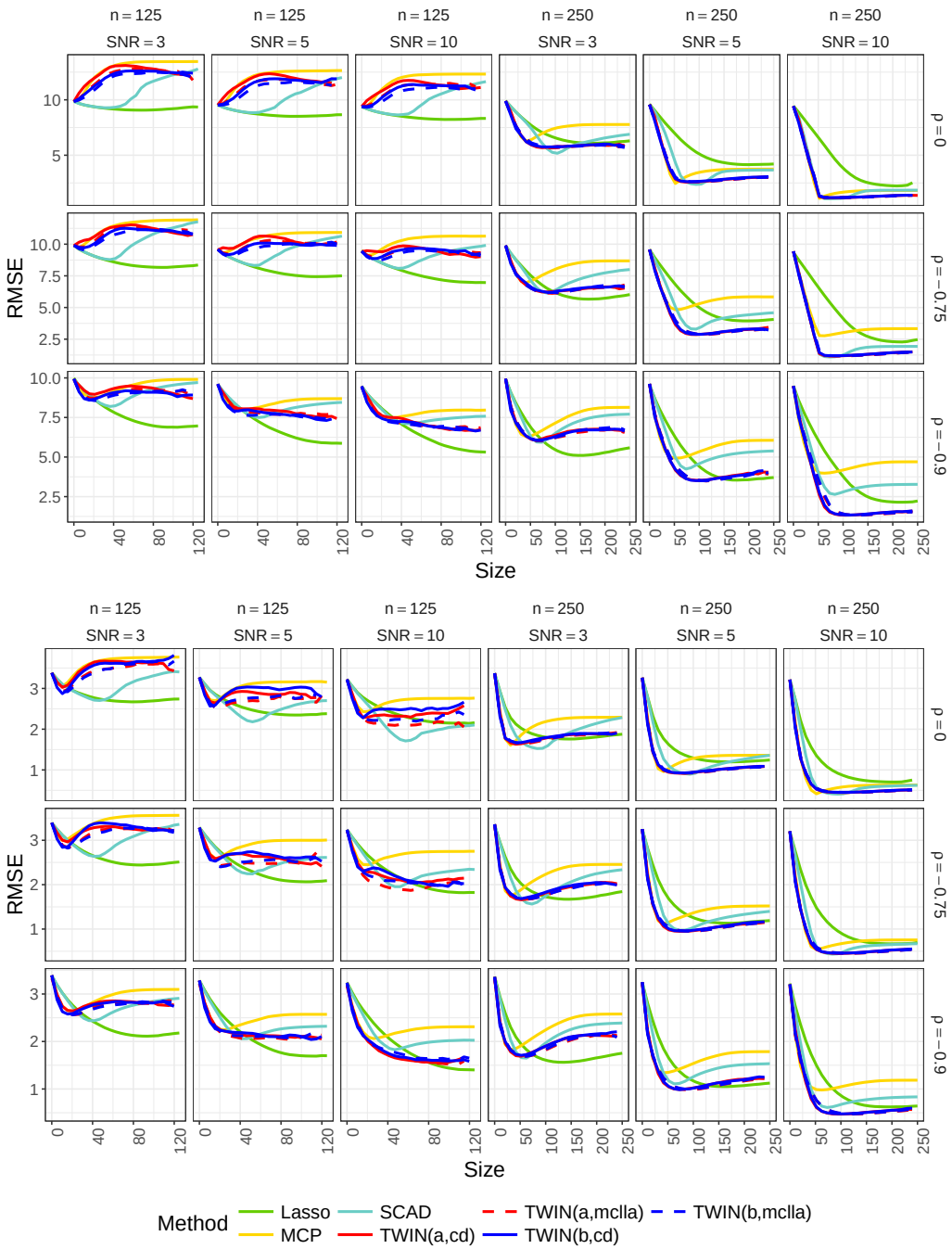

Figure 4.6: The results above are for a simulation with data generated under Model 1 (top panel) and Model 2 (bottom panel).

TWIN. We also investigated TWIN with $\tau = 0.15$ and the results were similar.

The average MSPE and number of selected variables for each method are reported in Table 4.1. Both TWIN-a and TWIN-b have better predictive performance than all other methods except the Lasso while retaining very parsimonious models. MCP selects about half has many genes as TWIN on average, but its MSPE is significantly worse than that of both TWIN-a and TWIN-b. Both TWIN penalties also yield stable results across the replications. The top two genes selected by both TWIN-a and TWIN-b are the same genes and are selected in all 100 replications by both penalties. The Lasso, MCP, and SCAD all selected one of these two genes for all replications. The gene selected second most often by TWIN was selected in all replications for the Lasso, but was only selected 10 times by SCAD and was never selected by MCP. The third most commonly selected gene for the TWIN penalties was the same gene for both TWIN-a (selected 44 times) and TWIN-b (selected 56 times). This gene was selected 88 times by the Lasso, 30 times by SCAD, and was never selected by MCP.

Method	Lasso	MCP	SCAD	TWIN-a	TWIN-b
MSPE	$0.613(0.058)$	$0.760(0.048)$	$0.740(0.048)$	$0.609(0.040)$	$0.651(0.049)$
Number Selected	$40.58(1.23)$	$1.66(0.10)$	$26.16(0.74)$	$3.16(0.14)$	$3.58(0.16)$

Table 4.1: Average test set MSPE and number of variables selected by Lasso, MCP, SCAD, TWIN-a, and TWIN-b. Standard errors are in parentheses. Note that $n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 2.090$, where $\bar{y}$ is the average of the response values.

Chapter 5

Towards Theoretical Understanding of Large Batch Training in Stochastic Gradient Descent

5.1 Introduction

Deep neural networks are typically trained by stochastic gradient descent (SGD) and its variants. These methods update the weights using an estimated gradient from a small fraction of large training data. Although deep neural networks are highly complex and non-convex, the SGD training models possess good properties in the sense that saddle points can be avoided (Ge et al., 2015) and "bad" local minima vanish exponentially (Choromanska et al., 2015; Dauphin et al., 2014). However, a central challenge remains about why and when SGD training neural networks tend to generalize well to unseen data despite the fact of heavily over-parameterization and overfitting (Zhang et al., 2017).

Recently, Keskar et al. (2016) proposed a hypothesis based on empirical experiments that (i) large-batch methods tend to converge to sharp minimizers of the training function and (ii) the sharp minimum causes a worse generalization. These two parts of the hypothesis are important for understanding the SGD in the deep neural networks. In this paper, we

focus on the first part of the hypothesis. Extensive numerical results corroborate the positive correlation between large-batch methods and sharp minimizers; see, e.g., Dinh et al. (2017); Hoffer et al. (2017). However, the theoretical result for supporting this observation is limited in the literature. Our work fills some gap in this important direction by providing new results on the properties of SGD in both *finite-time* regime where the number of SGD iterations is finite and *asymptotic* regime where the number of SGD iterations is sufficiently large. As a result, we can justify and provide new insights into the first part of the hypothesis by Keskar et al. (2016).

The main contributions of this paper are summarized as follows:

- We manage to use the finite-time escaping time of SGD from one local minimum to its nearest local minimum as an approach for justifying the hypothesis by Keskar et al. (2016).
- We prove that SGD tends to converge to flatter minima in the asymptotic regime regardless of the batch size. However, it may take exponential time to converge. This result provides new insights into the hypothesis by Keskar et al. (2016).
- We derive new results showing that the SGD with a larger learning rate to batch size ratio tends to converge to a flat minimum faster, however, its generalization performance could be worse than the SGD with a smaller learning rate to batch size ratio.

5.2 Main Results

Suppose the training set consists of N samples. we define $L_n(\cdot)$ as the loss function for the sample $n \in \{1, \dots, N\}$. Then $L(\cdot) = \mathbb{E}[L_n(\cdot)]$ is the risk function, where the expectation is taken with respect to the population of data. Let $\mathbf{w}$ be the vector of unknown model parameters in $\mathbb{R}^d$.

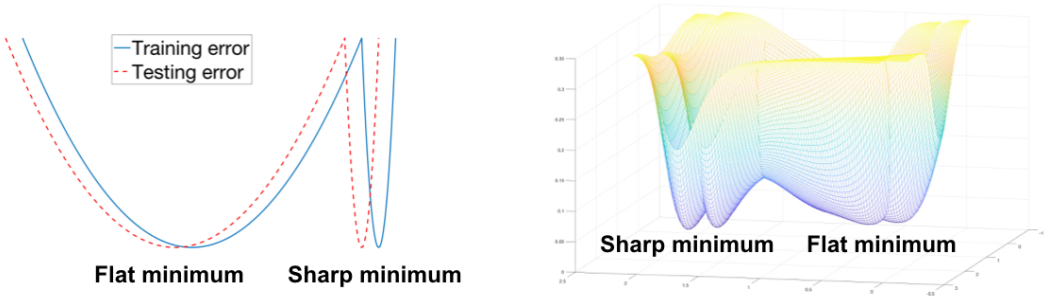

Figure 5.1: A sketch of "flat" and "sharp" minima for one-dimensional case (left panel) and two-dimensional case (right panel). The vertical axis indicates the value of the loss function.

The mini-batch SGD estimates the gradient $\mathbf{g}$ with some mini-batch B, a set of M randomly selected sample indices from $\{1, \ldots, N\}$, by $\widehat{\mathbf{g}}^{(B)}(\mathbf{w}) = \frac{1}{M}\sum_{n\in B}\nabla L_n(\mathbf{w})$. We consider the stochastic gradient descent with learning rate γ_k and mini-batch batch size M_k, and it gives the update rule

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{\gamma_k}{M_k}\sum_{n\in B_k}\nabla L_n(\mathbf{w}_k). \tag{5.1}$$

Here, k indexes the update step, and $|B_k| = M_k$. We call (5.1) a small batch training if $M_k << N$ and typically $M_k \in \{64, 128, 256\}$. In contrast, we call (5.1) a *large batch training* if M_k/N is some non-negligible positive constant and typically $M_k/N = 10\%$. We allow the diminishing learning rate γ_k and varying batch size M_k in (5.1), which is motivated from practice that SGD converges to the optimum by decreasing the learning rate.

KMNST hypothesis We call the following hypothesis proposed by Keskar et al. (2016) as the KMNST hypothesis since K-M-N-S-T is the collection of author initials in Keskar et al. (2016): *Large batch training tends to converge to the sharp minimizer of the training function*. A conceptual sketch of "sharp" (and relatively, "flat") minima is plotted in Figure 5.1. The theory built in this Section 5.2 aims to justify the KMNST hypothesis.

5.2.1 Stochastic Differential Equation for SGD

We consider SGD as a discretization of stochastic differential equations. Let $\text{Var}[\nabla L_n(\mathbf{w})] \equiv \boldsymbol{\sigma}^2(\mathbf{w})$, which is finite and positive definite for typical loss functions. In Appendix D.1, we show that for independent and identically distributed (iid) samples and $\mathbf{w}$ in any bounded domain,

$$\mathbb{E}[\widehat{\mathbf{g}}^{(B)}(\mathbf{w})] = \nabla L(\mathbf{w}), \quad \text{Var}[\widehat{\mathbf{g}}^{(B)}(\mathbf{w})] = M^{-1}\boldsymbol{\sigma}^2(\mathbf{w}).^1 \tag{5.2}$$

We can write the mini-batch SGD (5.1) as

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \gamma_k \nabla L(\mathbf{w}_k) + \frac{\gamma_k}{\sqrt{M_k}}\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ has zero mean and variance $\boldsymbol{\sigma}(\mathbf{w})$ by (5.2). We consider a stochastic differential equation (SDE):

$$d\mathbf{W}(t) = -\nabla L(\mathbf{W}(t))dt - \sqrt{\frac{\gamma(t)}{M(t)}}\boldsymbol{\sigma}(\mathbf{W}(t))d\mathbf{B}(t), \;\; \mathbf{W}(0) = \mathbf{w}_0. \tag{5.3}$$

By the Euler scheme, the SDE (5.3) can be discretized to obtain the mini-batch SGD (5.1); see, e.g., Mandt et al. (2017); Jastrzebski et al. (2017); Li et al. (2017). The stochastic Brownian term $\mathbf{B}(t)$ in (5.3) accounts for the random fluctuations due to the use of mini-batches for gradient estimation in (5.1). Note that (5.3) allows the batch size and step size to be time-dependent.

We consider the gradient covariance to be isotropic:

$$\boldsymbol{\sigma}^2(\mathbf{w}) = \beta(\mathbf{w}) \cdot \mathbf{I}, \tag{5.4}$$

where $\beta(\mathbf{w})$ may depend on $\mathbf{w}$. A similar assumption has been made in the literature, see e.g., Jastrzebski et al. (2017); Chaudhari et al. (2018), where they assume $\beta(\mathbf{w}) \equiv \beta$ is a

[1] Hoffer et al. (2017); Jastrzebski et al. (2017) obtain a similar result as the (5.2) but in a different sense. Specifically, (5.2) takes the expectation and variance with respect to the underlying population, however, Hoffer et al. (2017); Jastrzebski et al. (2017) take the expectation and variance with respect to the sampling distribution of $B \in \{1, \ldots, N\}$. Note that (5.2) is preferable if we want to analyze the risk function $L(\cdot)$ instead of the sample average loss $N^{-1}[L_1(\cdot) + \cdots + L_N(\cdot)]$ and if we regard the training data only a subset of the true underlying population.

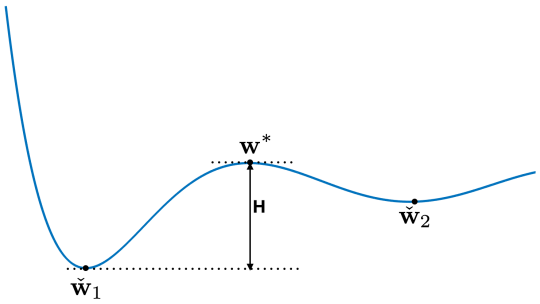

Figure 5.2: A sketch of two local minimizer $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$ of a risk function. The $\mathbf{w}^*$ is the saddle point between $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$ and the H is the relative height of $\mathbf{w}^*$ to $\check{\mathbf{w}}_1$.

constant. Let $p(\mathbf{w}, t)$ be the probability density function of the solution $\mathbf{W}(t)$ to the SDE (5.3). We derive the following characteristics for $p(\mathbf{w}, t)$ in Appendix D.2.

Lemma 5.1. *The $p(\mathbf{w}, t)$ satisfies the following Fokker-Planck equation:*

$$\partial_t p = \nabla \cdot \left(\left[\nabla \left(L(\mathbf{w}) + \frac{\gamma(t)\beta(\mathbf{w})}{2M(t)} \right) \right] p + \frac{\gamma(t)\beta(\mathbf{w})}{2M(t)} \nabla p \right), \quad p(\mathbf{w}, 0) = \delta(\mathbf{w}_0), \tag{5.5}$$

where $\delta(\cdot)$ denotes the delta function.

Note that the drift term in (5.5) is $-\nabla[L(\mathbf{w}) + \gamma(t)\beta(\mathbf{w})/\{2M(t)\}] \neq -\nabla L(\mathbf{w})$, which implies the SGD does not follow the mean drift $-\nabla L(\mathbf{w})$ to be its update direction. Specifically, a larger $\gamma(t)/M(t)$ ratio corresponds to a drift term deviate more from the mean drift $-\nabla L(\mathbf{w})$. This sheds light on the possible case that even the SGD with a larger $\gamma(t)/M(t)$ ratio tends to converge to a flat minimum faster (to be justified in Section 5.2.3), its generalization performance could be worse than the SGD with a smaller $\gamma(t)/M(t)$ ratio (to be illustrated in Section 5.3).

The results derived in this Section 5.2.1 can be related with the KMNST hypothesis in the following sense: the dynamics of SGD would depend on the $\gamma(t)/M(t)$ ratio instead of the γ or M separately, which is clear from the experiments in Section 5.3.

5.2.2 KMNST Hypothesis in the Finite-Time Regime

We first consider the behavior of SGD in the finite-time regime $t < \infty$, which is typical in the practice. Specifically, we are interested in the escape time of SGD from one local minimizer $\check{\mathbf{w}}_1$ to its nearest local minimizer $\check{\mathbf{w}}_2$. Refer to the Figure 5.2 as an illustration. Let $\mathbf{w}^*$ be the saddle point[2] between $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$. By the definition of $\mathbf{w}^*$, the Hessian $\Delta L(\mathbf{w}^*)$ can be shown to have a single negative eigenvalue $-\lambda^*$ (e.g., Berglund (2013)). By the Eyring-Kramers formula, we have the following theorem.

Theorem 5.2. *Let $\tau_{\check{\mathbf{w}}_1 \to \check{\mathbf{w}}_2}$ be the transition time from $\check{\mathbf{w}}_1$ to $\check{\mathbf{w}}_2$ for $\mathbf{W}(t)$, then*

$$\mathbb{E}[\tau_{\check{\mathbf{w}}_1 \to \check{\mathbf{w}}_2}] = \frac{2\pi}{\lambda^*} \sqrt{\frac{|\Delta L(\mathbf{w}^*)|}{|\Delta L(\check{\mathbf{w}}_1)|}} e^{H \cdot 2M(\check{\mathbf{w}}_1)/[\gamma(\check{\mathbf{w}}_1)\beta(\check{\mathbf{w}}_1)]} \{1 + O\left(\sqrt{\epsilon}\log(\epsilon^{-1})\right)\}$$

where $|\Delta L(\cdot)|$ represents for the determinate of the Hessian, $H = H(\mathbf{w}^*, \check{\mathbf{w}}_1) \equiv L(\mathbf{w}^*) - L(\check{\mathbf{w}}_1)$ is the relative height of $\mathbf{w}^*$ to $\check{\mathbf{w}}_1$, $M(\check{\mathbf{w}}_1)$ is the batch size of the SGD at $\check{\mathbf{w}}_1$, $\gamma(\check{\mathbf{w}}_1)$ is the learning rate of the SGD at $\check{\mathbf{w}}_1$, and β is defined in (5.4).

The above theorem is proved by Bovier et al. (2004, 2005) and in a more general case by Berglund (2013). From this theorem, one can see that the transition time depends on three factors, the diffusion factor $\gamma\beta/M$ in the SGD, the potential barrier $H(\mathbf{w}^*, \check{\mathbf{w}}_1)$ that SGD has to climb in order to escape $\check{\mathbf{w}}_1$, and the determinant of the Hessian at $\check{\mathbf{w}}_1$ and $\mathbf{w}^*$.

The results shown in this Section 5.2.2 can explain the KMNST hypothesis as follows. A larger batch size M of SGD at local minimizer $\check{\mathbf{w}}_1$ corresponds to a longer escaping time from $\check{\mathbf{w}}_1$, which is modeled by $\mathbb{E}[\tau_{\check{\mathbf{w}}_1 \to \check{\mathbf{w}}_2}]$. Hence, even if $\check{\mathbf{w}}_1$ corresponds to a sharp minimum with a large $|\Delta L(\check{\mathbf{w}}_1)|$, the exponential term $\exp[H \cdot 2M(\check{\mathbf{w}}_1)/[\gamma(\check{\mathbf{w}}_1)\beta(\check{\mathbf{w}}_1)]]$ could dominate the escaping time. As a result, the large batch training will be trapped at a sharp minimizer in the finite-time regime, which is the same as observed by Keskar et al. (2016) that *large batch training tends to converge to the sharp minimizer of the training function*. On the other hand, if

[2]There are possibly multiple saddle points between $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$. We define $\mathbf{w}^*$ as the saddle point with the minimal height among all saddle points in the following sense. Let $w(t), 0 \le t \le 1$, be any continuous path from $\check{\mathbf{w}}_1$ to $\check{\mathbf{w}}_2$. Denote by $\widehat{w} = \arg\inf_{w:w(0)=\check{\mathbf{w}}_1, w(1)=\check{\mathbf{w}}_2} \sup_{t\in[0,1]} L(w(t))$ the path with the minimal saddle point height among all continuous path. We define that $\mathbf{w}^* = \max_{t\in[0,1]} \widehat{w}(t)$.

the batch size is small, then $\exp[H \cdot 2M(\check{\mathbf{w}}_1)/[\gamma(\check{\mathbf{w}}_1)\beta(\check{\mathbf{w}}_1)]]$ is small. As a result, only when $|\Delta L(\check{\mathbf{w}}_1)|$ is small enough, then SGD can be trapped at this minimizer, which implies that small batch training tends to converge to flatter minima.

However, these phenomena will change in the asymptotic regime $t \to \infty$ as explained in Section 5.2.3.

5.2.3 KMNST Hypothesis in the Asymptotic Regime

Main assumptions. In this section, we consider the asymptotic regime that $t \to \infty$ and suppose the following three assumptions[3]:

(A.1) $L(\mathbf{w})$ is *confinement*: $\lim_{\|\mathbf{w}\|\to+\infty} L(\mathbf{w}) = +\infty$ and $\int e^{-L(\mathbf{w})} d\mathbf{w} < +\infty$.

(A.2) $\lim_{\|\mathbf{w}\|\to+\infty} \left\{\|\nabla L(\mathbf{w})\|^2/2 - \nabla \cdot \nabla L(\mathbf{w})\right\} = +\infty$, where $\nabla \cdot \nabla L$ denotes the trace of the Hessian for L. Moreover, $\lim_{\|\mathbf{w}\|\to+\infty} \left\{\nabla \cdot \nabla L(\mathbf{w})/\|\nabla L(\mathbf{w})\|^2\right\} = 0$.

(A.3) There exists a constant M, such that $\left|e^{-L(\mathbf{w})}\left(\|\nabla L(\mathbf{w})\|^2 - \nabla \cdot \nabla L(\mathbf{w})\right)\right| \leq M$.

We show in Appendix D.3 that (A.1) – (A.3) hold for typical loss functions such as the regularized mean cross entropy and the square loss functions. These assumptions appear commonly in the diffusion process literature, see, e.g., Pavliotis (2014). In particular, (A.1) ensures the Gibbs density function $p_G(\mathbf{w}) = e^{-L(\mathbf{w})}$ is well defined, and (A.2) is sufficient for the measure $\mu(\mathbf{w}) = \int p_G(\mathbf{w}) d\mathbf{w} = \int e^{-L(\mathbf{w})} d\mathbf{w}$ to satisfy the Poincaré inequality (e.g., Pavliotis (2014); Raginsky et al. (2017)):

$$\int \|\nabla f(\mathbf{w})\|^2 \, d\mu(\mathbf{w}) \geq C_P \int \left(f(\mathbf{w}) - \int f(\mathbf{w}) d\mu(\mathbf{w})\right)^2 d\mu(\mathbf{w}), \text{ for some } C_P > 0, \quad (5.6)$$

holds for any f satisfying $\int f^2(\mathbf{w}) d\mathbf{w} < \infty$.

We first give the stationary solution for the Fokker-Planck equation (5.5) when $t \to \infty$.

[3]We note that if the parameter vector $\mathbf{w}$ lies in a bounded region, then the Gibbs density is well defined only if $\int e^{-L(\mathbf{w})} d\mathbf{w} < \infty$, the Poincaré inequality is always true, and the assumption (A.3) is always true. Thus, although the mean cross entropy loss with bounded parameters does not satisfy (A.1) or (A.2), our results in this section still hold for the mean cross entropy loss.

Lemma 5.3. *Under the assumption (A.1) and suppose $\beta(\mathbf{w}) \equiv \beta$, then (5.5) has a stationary solution*

$$p_\infty(\mathbf{w}) = \kappa e^{-\eta_\infty L(\mathbf{w})},$$

where

$$\eta_\infty = 2M/[\gamma\beta(\check{\mathbf{w}})]$$

with the limiting batch size $M = \lim_{t\to\infty} M(t)$, the limiting learning rate $\gamma = \lim_{t\to\infty}\gamma(t)$, and the convergent local minimizer $\check{\mathbf{w}}$. The constant κ in the above formula is a normalization factor such that $\int p_\infty(\mathbf{w}) = 1$.

Proof for this lemma is given in Appendix D.4. We remark that for general $\beta(\mathbf{w})$ depending on $\mathbf{w}$, the existence and an explicit form of stationary solution for (5.5) remain an open question in the literature. Hence, we focus on $\beta(\mathbf{w}) \equiv \beta$ in this section.

Similar results as Lemma 5.3 can be found in related work, e.g., Jastrzebski et al. (2017). However, it is not clear whether $p(\mathbf{w}, t)$ converges to $p_\infty(\mathbf{w})$, not to mention how fast that $p(\mathbf{w}, t)$ would converge to $p_\infty(\mathbf{w})$. The following theorem gives a positive answer to this question, which later provides a new insight into the justification of KMNST hypothesis.

Theorem 5.4. *Under assumptions (A.1) – (A.3), the probability density function $p(\mathbf{w}, t)$ of $\mathbf{W}(t)$ converges to the stationary solution $p_\infty(\mathbf{w})$. Moreover, there exists $T > 0$ such that for any $t > T$,*

$$\left\| \frac{p(\mathbf{w},t) - p_\infty(\mathbf{w})}{\sqrt{p_\infty(\mathbf{w})}} \right\|^2_{L^2(\mathbb{R}^d)} \le C(t,T)e^{-C_P\cdot(t-T)/\eta_\infty},$$

where C_P is a constant define in (5.6) and $C(t,T) = C_P\cdot(t-T)/\eta_\infty + \left\|(p(T) - p_\infty)/\sqrt{p_\infty}\right\|^2_{L^2(\mathbb{R}^d)}$.

The proof for this theorem is given in Appendix D.5. We also give a quantification of constant T in Appendix D.6. Three remarks on Theorem 5.4 are as follows. *First*, this theorem shows that $p(\mathbf{w}, t)$ always converges to p_∞ with an exponential rate regardless of the initial value. This theorem provides a theoretical ground for the work that manages to understand $p(\mathbf{w}, t)$ based on analysis of the stationary distribution p_∞ (see, e.g., Jastrzebski

et al. (2017)). *Second*, it is known (Raginsky et al., 2017) the Poincaré constant $C_P \propto e^d$, where d is the dimension of the parameter $\mathbf{w}$. In the setting of the deep neural networks, C_P can be very large and it takes exponential time $t > e^d$ such that $p(\mathbf{w}, t)$ would approach to the stationary distribution p_∞. Therefore, the results only based on the stationary solution do not reveal information in the finite-time regime. *Third*, the convergence rate is relatively faster with a larger γ/M since it corresponds to a smaller η_∞. The last two remarks are illustrated by experiments in Section 5.3.

We now characterize $\mathbf{W}(t)$ in the asymptotic regime $t \to \infty$ based on the stationary distribution p_∞, and we give the proof of the following theorem in Appendix D.7.

Theorem 5.5. *Let $\check{\mathbf{w}}$ be a local minimizer. Then,*

$$\lim_{\epsilon \to 0} \mathbb{P}(|\mathbf{W}(\infty) - \check{\mathbf{w}}| \leq \epsilon) = \frac{\kappa e^{-2\eta_\infty L(\check{\mathbf{w}})}}{\eta_\infty^{d/2} \sqrt{|\Delta L(\check{\mathbf{w}})|}} \lim_{\epsilon \to 0} \left[e^{\eta_\infty \epsilon^2} \prod_{j=1}^{d} \sqrt{1 - e^{-\epsilon^2 \eta_\infty \lambda_j / \pi}} \right],$$

where d is the dimension of $\mathbf{w}$, λ_js and $|\Delta L(\check{\mathbf{w}})|$ represent the eigenvalues and the determinant of loss function Hessian $\Delta L(\check{\mathbf{w}})$, respectively, and the constants κ, η_∞ are defined in Lemma 5.3.

Given the complex form of the probability in Theorem 5.5, we give numerical illustrations in Appendix D.8. To appreciate the implication of Theorem 5.5, we consider any two local minimizers $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$ with the same value of $L(\check{\mathbf{w}}_1) = L(\check{\mathbf{w}}_2)$. Then,

$$\lim_{\epsilon \to 0} \frac{\mathbb{P}(|\mathbf{W}(\infty) - \check{\mathbf{w}}_1| \leq \epsilon)}{\mathbb{P}(|\mathbf{W}(\infty) - \check{\mathbf{w}}_2| \leq \epsilon)} = \sqrt{\frac{|\Delta L(\check{\mathbf{w}}_2)|}{|\Delta L(\check{\mathbf{w}}_1)|}}. \tag{5.7}$$

The ratio of probability (5.7) implies that in the asymptotic regime $t \to \infty$, the probability of SGD converging to a flatter minimum with a smaller determinant $|\Delta L(\cdot)|$ is always larger than the probability of SGD converging to a sharper minimum with a larger determinant $|\Delta L(\cdot)|$. Moreover, (5.7) does not depend on the batch size or learning rate, but it only depends on the determinant of Hessian at the local minimum.

The results derived in this Section 5.2.3 provide some new insights into the KMNST hypothesis: SGD tends to converge to flatter minima regardless of the batch size M (or

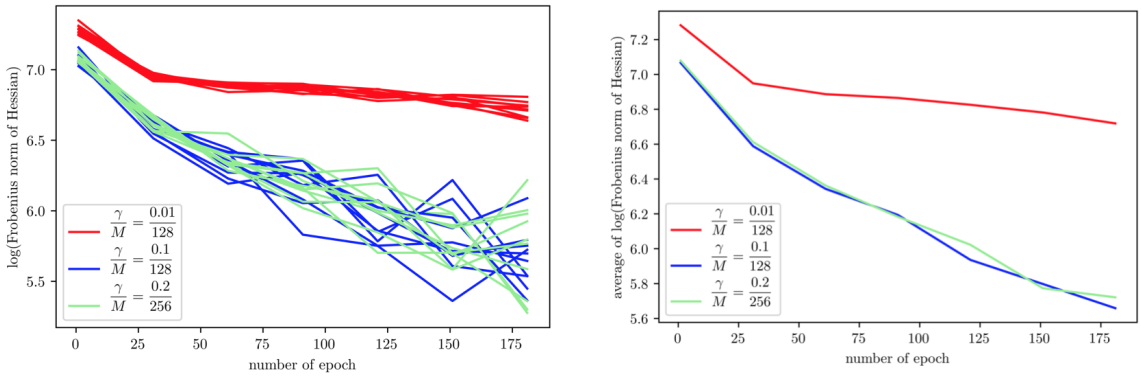

Figure 5.3: Log of Frobenius norm of Hessian as a function of epochs. Three (γ, M) pairs $(0.01, 128)$, $(0.1, 128)$ and $(0.2, 256)$ are studied, which are denoted in red, blue and green, respectively. The left plot shows 10 experiments for each of three (γ, M) pairs and the right plot shows the average of 10 experiments. Total 180 epochs are trained.

the ratio γ/M) in the asymptotic regime $t \to \infty$ as shown by (5.7). However, it may take exponential time e^d to converge, where d is the dimension of the model parameter $\mathbf{w}$. The experiments in Section 5.3 further corroborate these theoretical finding.

5.3 Numerical Experiments

In this section, we show experiments to corroborate the theoretical findings in the previous section. We train 4-layer batch-normalized ReLU MLPs on MNIST with different learning rate γ and batch size M. Specifically, we use three γ/M ratios: $\gamma/M = 0.01/128, 0.1/128, 0.2/256$. As is common for such tasks, the mean cross entropy loss is used as the loss function. We discussed in Section 5.2.1 that this loss satisfies our assumptions for theoretical analysis.

Geometry of SGD updates. Figure 5.3 shows the log of Frobenius norm of Hessian for minima obtained by SGD. Due to the high computational cost for computing the determinant of the Hessian, we use the Frobenius norm of the Hessian as a substitute. Note that a larger Frobenius norm of Hessian corresponds to a sharper minimum. The Frobenius norm is approximated using the method in Wu and Zhu (2017). Note that the dynamics of SGD behave similar across 10 experiments for each of three γ/M ratios as shown in the left plot

of Figure 5.3. Hence we focus on the averaged dynamics as in the right plot of Figure 5.3. Three main results can be observed from Figure 5.3:

- First, for the same γ/M ratio (e.g., $\gamma/M = 0.1/128$ and $0.2/256$), the minima obtained by SGD have the very similar norm of the Hessian. This illustrates the Lemma 5.1, 5.3 and Theorem 5.2 that the dynamics and geometry of the minima obtained by SGD would depend on the ratio γ/M instead of individual γ or M separately. A similar phenomenon is also observed by Jastrzebski et al. (2017).

- Second, since the SGD is trained using 180 epochs, the dynamics of SGD in Figure 5.3 fall in the finite-time regime. It is clear that the rate of SGD tending to a flatter minimum (i.e., with a smaller norm of the Hessian) with a larger γ/M ratio (e.g., $\gamma/M = 0.1/128$) is faster compared to with a smaller γ/M ratio (e.g., $\gamma/M = 0.01/128$). This illustrates the finite-time analysis in Theorem 5.2 that the SGD with a smaller γ/M ratio is easier to be trapped around a minimum and hence the SGD tends to other minima slower. As a result, the Hessian of minima changes slower for SGD with a smaller γ/M ratio.

- Third, Figure 5.3 also sheds light on the dynamics of SGD in the asymptotic regime. The SGD tends to converge to a flatter minimum regardless of the γ/M ratio, which demonstrates Theorem 5.5 and its corollary (5.7). However, the convergence rate is slow, in particular for the SGD with a small γ/M ratio, which is theoretically shown in Theorem 5.4 and its following remarks.

Training and generalization of SGD. Figure 5.4 shows the training accuracy and loss for the model trained by SGD. We run 10 experiments. It is clear that the training accuracy and loss are very close across 10 experiments for each of three γ/M ratios. Thus, we focus on interpreting the training and generalization performance of the model obtain from one experiment, which is shown in Figure 5.5. Three main results can be observed from Figure 5.5:

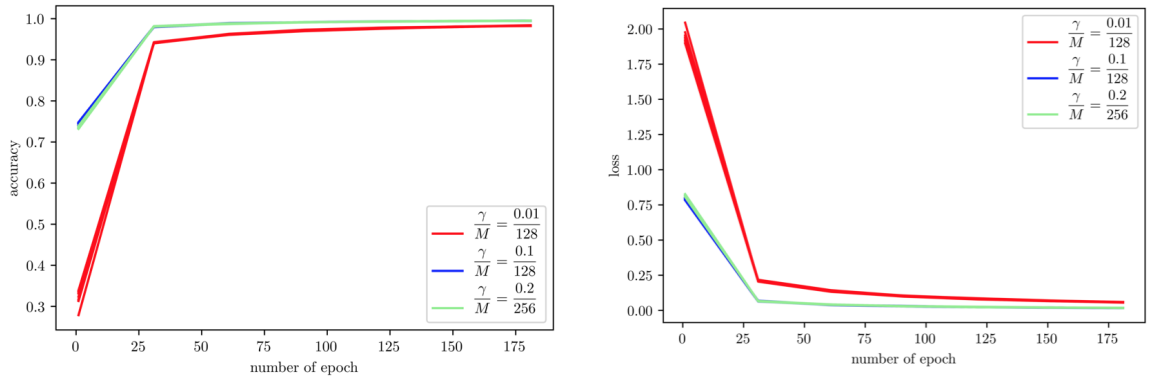

Figure 5.4: The left plot shows the training accuracy as a function of epochs and the right plot shows the cross entropy loss as a function of epochs. Three (γ, M) pairs $(0.01, 128)$, $(0.1, 128)$ and $(0.2, 256)$ are studied, which are denoted in red, blue and green, respectively. Both plots show 10 experiments for each of three (γ, M) pairs. Total 180 epochs are trained.

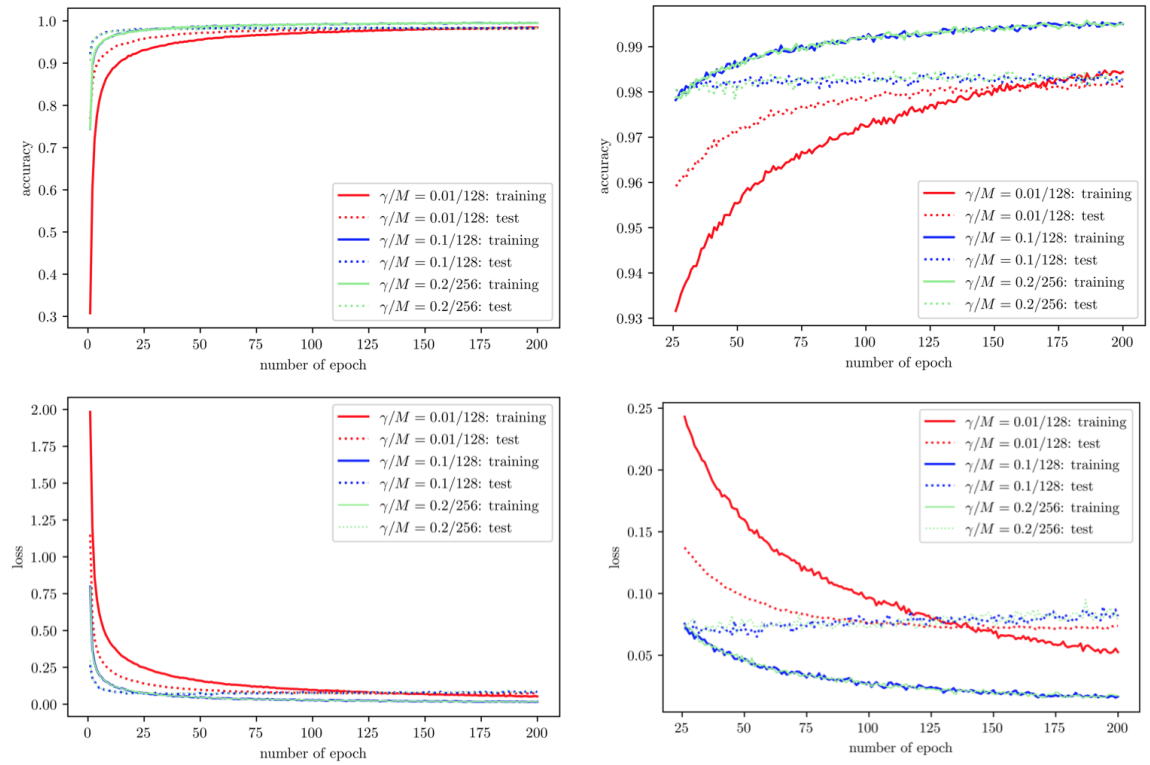

Figure 5.5: The top left plot shows the training and test accuracy as a function of epochs. The top right plot gives the zoomed in performance of the accuracy when epochs are no less than 25. The bottom left plot shows the cross entropy loss as a function of epochs. The bottom right plot gives the zoomed in performance of the loss when epochs are no less than 25. Three (γ, M) pairs $(0.01, 128)$, $(0.1, 128)$ and $(0.2, 256)$ are studied, which are denoted in red, blue and green, respectively. Total 200 epochs are trained.

- First, for the same γ/M ratio (e.g., $\gamma/M = 0.1/128$ and $0.2/256$), the training error and test error are very close. This meets our expectation since the dynamics of SGD only depends on the ratio γ/M as discussed above and the models trained by SGD with the same γ/M ratio should behave similarly.

- Second, the model obtained with a larger γ/M ratio (e.g., $\gamma/M = 0.1/128$) gives a better training accuracy and a smaller training loss compared with the case of a smaller γ/M ratio (e.g., $\gamma/M = 0.01/128$). This can be partially justified by our finite-time analysis in Theorem 5.2 that the SGD with a larger γ/M ratio is easier to escape a local minimum.

- Third, the model obtained with a smaller γ/M ratio gives a *smaller* test loss after a certain time (it is after 100 epochs in the bottom right plot of Figure 5.5). This can be explained by Lemma 5.1 and its following remark. Specifically, a smaller γ/M ratio corresponds a mean drift deviates less from the mean drift $-\nabla L(\mathbf{w})$, where $-\nabla L(\mathbf{w})$ is the drift for a global minimizer of the risk function $L(\mathbf{w})$. This shows a tradeoff between the large and small γ/M ratio in the sense of the training and test loss.

5.4 Related Work

The modeling of SDE for approximating SGD is well studied in the literature. See, e.g., Mandt et al. (2017); Poggio et al. (2017); Li et al. (2017); Jastrzebski et al. (2017); Chaudhari and Soatto (2018) and the references therein. Different from these work, we give a new result in Lemma 5.1, which not only gives the dynamics of SDE solution but also connects with the generalization performance. We also derive the theory for the SDE solution in the asymptotic regime, especially the convergence rate.

We clarify the definition of the sharpness in multi-dimensional cases. We find that the production of eigenvalues, or equivalently, the determinant of the risk function Hessian at minimizers is appropriate. Similar results have been derived in Jastrzebski et al. (2017); Dziugaite and Roy (2017).

The work by Jastrzebski et al. (2017) remarkably emphasize how the learning rate to batch size ratio affects the SGD and they also relate with the KMNST hypothesis. Here are some differences between Jastrzebski et al. (2017) and ours.

- Jastrzebski et al. (2017) use the stationary probability $p_\infty(\mathbf{w})$ to explain that the behavior of the SGD. However, we show that it takes the exponential time for $p(\mathbf{w}, t)$ to converge to $p_\infty(\mathbf{w})$ in the setting of the deep neural network. Hence, $p_\infty(\mathbf{w})$ cannot fully explain the behavior of SGD in the practical finite-time regime. Our work adds new elements to this picture by studying the escaping time of SGD from a local minimum in the sense of finite-time regime and we also give a new result on the convergence rate of $p(\mathbf{w}, t) \to p_\infty(\mathbf{w})$.

- In particular, the stationary probability in Jastrzebski et al. (2017) can not explain the KMNST hypothesis when two local minimizers $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$ having a same risk $L(\check{\mathbf{w}}_1) = L(\check{\mathbf{w}}_2)$. In this case, the result of Jastrzebski et al. (2017) coincides with (5.7) and it is independent of M or γ, which is undesired in explaining the KMNST hypothesis. On the other hand, there may exist many minima with a same risk value but different Hessians for a deep neural network. Therefore, our finite-time results can give a better explanation to the KMNST hypothesis in this case.

Chapter 6

Another Look at Statistical Calibration: A Non-Asymptotic Theory and Prediction-Oriented Optimality

6.1 Introduction

In engineering and sciences, computer models are increasingly used for studying complex physical systems such as cosmology, weather forecasting, material science, and shock physics (Santner et al., 2003). Let Y denote the output from a physical system $\zeta(\cdot)$ with the input X. Let $\{(X_i, Y_i) : i = 1, \ldots, n\}$ be independent copies of random pair (X, Y) from a regression model:

$$Y = \zeta(X) + \varepsilon, \tag{6.1}$$

where the random error ε follows a $\mathcal{N}(0, \sigma^2)$ distribution and the design point X has support on $\Omega = [0, 1]^d$. Let $\eta(x, \theta)$ denote a computer model for approximating $\zeta(x)$ with inputs $x \in \Omega$ and *calibration parameters* $\theta \in \Theta \subset \mathbb{R}^p$. The values of θ cannot be directly observed and

typically unknown in the physical data. As George Box famously stated *"All models are wrong, but essentially some are useful"*, even the best computer models are only approximations of reality. It is possible to enhance the quality of the computer model $\eta(x, \theta)$ by tuning or calibrating the calibration parameters θ. But in most practical scenarios, the computer output $\eta(x, \theta)$ cannot fit the physical response $\zeta(x)$ perfectly, regardless of how the calibration parameters θ are best tuned (Kennedy and O'Hagan, 2001; Santner et al., 2003). Another practical fact is that only a *limited* number n of training data are available from the physical experiment in (6.1) to tune θ.

By simultaneously following these two practical considerations, we take a new look at the calibration problem. Our purpose is to optimally predict $\zeta(\cdot)$ by calibrating θ and estimating the model discrepancy $\zeta(\cdot) - \eta(\cdot, \theta)$ based on a finite number of physical data. To this end, we use nonparametric approaches to modeling the physical system and the discrepancy function. We establish a non-asymptotic minimax estimation risk for nonparametric regression and achieve the optimal risk by using regularized estimators in the reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950; Wahba, 1990). We show that our prediction oriented calibration is equivalent to finding the minimizer of model discrepancy under the RKHS norm. We further establish an exact statistical guarantee in the sense that for a finite sample of physical observations, the prediction error is minimized by using the computer model calibrated with the proposed method. Furthermore, we provide an algorithm to estimate the optimal calibration parameters and the model discrepancy.

6.1.1 Comparison to Existing Work

We discuss the differences between our calibration method and other frequentist calibration methods in the literature. Joseph and Melkote (2009) considers calibration using a parametric form of discrepancy. We use a nonparametric approach in RKHS to better modeling the physical system and the model discrepancy. Nonparametric methods have been used in related works including Tuo and Wu (2015) and Wong et al. (2017). Specifically, Tuo and Wu (2015) proposes the L_2-calibration by minimizing the model discrepancy under the

L_2-norm. Wong et al. (2017) performs calibration by minimizing the model discrepancy under the empirical l_2-norm. Below are the main differences between Tuo and Wu (2015); Wong et al. (2017) and ours.

- Different methods. Calibration in Tuo and Wu (2015), Wong et al. (2017) and our work minimize different norms of the model discrepancy: the L_2-norm in Tuo and Wu (2015), the empirical l_2-norm in Wong et al. (2017) and the RKHS norm in our method.

- Different analyses. Theoretical results in Tuo and Wu (2015) and Wong et al. (2017) are based on asymptotics assuming the number of physical observations go to infinity. Our theory is based on finite-sample properties of calibration and prediction following the fact that usually, only a finite number of physical data are available.

- Different results. The L_2-calibration in Tuo and Wu (2015) minimizes the distance between the physical system and the imperfect computer model, but not directly for predicting the physical system. Wong et al. (2017) performs the least square calibration and then estimates the model discrepancy in the RKHS. For a finite number of physical observations, the estimation error of discrepancy can be large. To overcome this difficulty, our calibration method minimizes the predictive mean squared error for a finite sample of physical data with statistical guarantees.

Bayesian calibration was studied by Kennedy and O'Hagan (2001); Oakley and O'Hagan (2004); Higdon et al. (2004, 2008); Joseph and Yan (2015); Plumlee (2017); Tuo and Wu (2018), among others. Our frequentist calibration method is easier to compute and complements these Bayesian methods. Furthermore, we will discuss a connection between our method and these Bayesian methods in Section 6.4.

Our non-asymptotic minimax theory is inspired by recent developments of concentration inequalities that provide valid statistical inference and estimation results for finite samples. Existing research on concentration inequalities typically addresses finite dimensional parameters for parametric models. Because our interest is computer model calibration, we develop a non-asymptotic minimax theory for nonparametric models. The novelty of our

work is to find an explicit form of the constant besides the well-known rate for nonparametric estimations in the finite-sample regime. Moreover, we prove the constant is minimax optimal by showing it exactly matches the minimax lower bound. These results are new in the nonparametric statistics literature and they are the key to our new calibration method.

The remainder of the article is organized as follows. In Section 6.2, we discuss the identifiability issue and formulate a prediction-oriented optimal calibration method. In Section 6.3, we establish a non-asymptotic minimax theory and apply it to the prediction-oriented calibration method. In Section 6.4, we develop an algorithm for computing our calibration procedure and build a connection between our method and the Bayesian calibration method. In Section 6.5, we provide synthetic and real examples to corroborate the derived theory and illustrate some advantages of the proposed calibration method. Technical proofs are delegated to the Appendix.

6.2 Prediction-Oriented Calibration

Since the computer model is imperfect for modeling the physical system, $\eta(\cdot,\theta) \not\equiv \zeta(\cdot)$ for any $\theta \in \Theta$. A model discrepancy function $\delta(\cdot,\theta) \stackrel{\text{def}}{=} \zeta(\cdot) - \eta(\cdot,\theta)$ is commonly used (Kennedy and O'Hagan, 2001; Martins-Filho et al., 2008). Equivalently, write

$$\zeta(x) \equiv \eta(x,\theta) + \delta(x,\theta), \ \ \forall x \in \Omega, \ \theta \in \Theta. \tag{6.2}$$

The goal is to accurately predict $\zeta(\cdot)$ using the computer model, which requires calibrating θ and estimating $\delta(\cdot,\theta)$ simultaneously with a finite physical sample in (6.1). Two main difficulties arise. The identifiability issue of θ to be discussed in Section 6.2.1 and the non-negligible estimation error of $\delta(\cdot,\theta)$ due to the finite sample of the physical data. These two issues motivate our calibration method in Section 6.2.2.

6.2.1 The Identifiability Issue

Suppose that $\zeta(\cdot)$ resides in a RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ on $\Omega = [0,1]^d$. One example of $\mathcal{H}$ is the mth order Sobolev space $\mathcal{W}_2^m(\Omega)$ with $2m > d$:

$$\mathcal{W}_2^m(\Omega) = \left\{ g(\cdot) \in L_2(\Omega) \Big| \frac{\partial^{\alpha_1+\cdots+\alpha_d}}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d} g(\cdot) \in L_2(\Omega), \right. \\ \left. \forall \alpha_1, \ldots \alpha_d \in \mathbb{N} \text{ with } \alpha_1 + \cdots + \alpha_d \leq m \right\}.$$

See Wahba (1990) for an explicit form of kernel associated with $\mathcal{W}_2^m(\Omega)$. Since $\eta(\cdot, \theta)$ approximates the physical system $\zeta(\cdot)$, we assume the following regularity condition for $\eta(\cdot, \theta)$.

Assumption 1. *For any $\theta \in \Theta$, the computer model $\eta(\cdot, \theta) \in \mathcal{H}$.*

An analogy of Assumption 1 was already used in Plumlee (2017). Unlike our assumption of the RKHS function space, Plumlee (2017) considers a function space of bounded mixed derivatives. In practice, one can choose the RKHS $\mathcal{H}$ by studying the smoothness of computer model $\eta(\cdot, \theta)$. Assumption 1 implies that for any $\theta \in \Theta$, $\delta(\cdot, \theta) = \zeta(\cdot) - \eta(\cdot, \theta) \in \mathcal{H}$. This observation leads to a potential identifiability issue for θ. For example, for two different $\theta_1 \neq \theta_2 \in \Theta$, their corresponding model discrepancies $\delta(\cdot, \theta_1) = \zeta(\cdot) - \eta(\cdot, \theta_1)$ and $\delta(\cdot, \theta_2) = \zeta(\cdot) - \eta(\cdot, \theta_2)$ are both in $\mathcal{H}$. This implies that both $(\theta_1, \delta(\cdot, \theta_1)), (\theta_2, \delta(\cdot, \theta_2)) \in \Theta \times \mathcal{H}$ are true for model (6.2). There are infinitely many pairs $(\theta, \delta(\cdot, \theta)) \in \Theta \times \mathcal{H}$ true for (6.2) by arbitrarily choosing $\theta \in \Theta$ and using $\delta(\cdot, \theta) = \zeta(\cdot) - \eta(\cdot, \theta)$. This identifiability issue was first noticed by K. Beven and P. Diggle in their discussion of Kennedy and O'Hagan (2001).

6.2.2 Definition of Prediction-Oriented Calibration

Denote by Π the sampling distribution of X_i in (6.1) which is independent of ε_i and satisfies $\Pi(\Omega) = 1$. Here, Π is assumed to be absolutely continuous with respect to Lebesgue's measure. Let X^* be drawn from Π and $Y^* = \zeta(X^*) + \varepsilon^* = \eta(X^*, \theta) + \delta(X^*, \theta) + \varepsilon^*$ with $\varepsilon^* \sim \mathcal{N}(0, \sigma^2)$. Then for a fixed $\theta \in \Theta$, the minimal predictive mean squared error for

predicting Y^* is

$$\begin{aligned} &\infty_{\widetilde{\delta}_n} \mathbb{E}\left\{Y^* - [\eta(X^*, \theta) + \widetilde{\delta}_n(X^*, \theta)]\right\}^2 \\ &= \sigma^2 + \infty_{\widetilde{\delta}_n} \|\delta(\cdot, \theta) - \widetilde{\delta}_n(\cdot, \theta)\|^2_{L_2(\Pi)}, \end{aligned} \tag{6.3}$$

where the infimum is taken over all estimators $\widetilde{\delta}_n$ that are measurable functions of $\{(X_i, Y_i - \eta(X_i, \theta)) : i = 1, \ldots, n\}$ for a given θ.

The identifiability issue discussed in Section 6.2.1 indicates that there are infinitely many pairs of $(\theta, \delta(\cdot, \theta)) \in \Theta \times \mathcal{H}$ satisfying model (6.2). For a finite sample size n, the minimal estimation error $\infty_{\widetilde{\delta}_n} \|\delta(\cdot, \theta) - \widetilde{\delta}_n(\cdot, \theta)\|_{L_2(\Pi)}$ does not vanish (Cover and Thomas, 2006). Since $\eta(\cdot, \theta)$ is generally nonlinear, different choices of $\theta \in \Theta$ correspond to $\delta(\cdot, \theta)$ with distinct minimal estimation errors $\infty_{\widetilde{\delta}_n} \|\delta(\cdot, \theta) - \widetilde{\delta}_n(\cdot, \theta)\|_{L_2(\Pi)}$. This heuristic argument is justified in Section 6.3.

This observation also motivates us to define *optimal calibration* $\theta^{\text{opt-pred}}$ to minimize the minimal predictive mean squared error (6.3) uniformly for $\zeta \in \mathcal{H}$ over $\theta \in \Theta$. Equivalently, we define

$$\theta^{\text{opt-pred}} \stackrel{\text{def}}{=} \arg\min_{\theta \in \Theta} \left\{ \infty_{\widetilde{\delta}_n} \|\delta(\cdot, \theta) - \widetilde{\delta}_n(\cdot, \theta)\|_{L_2(\Pi)} \right\}, \tag{6.4}$$

which holds uniformly for $\zeta \in \mathcal{H}$. Here, the superscript "opt-pred" denotes "optimal for prediction". By definition and (6.3), $\theta^{\text{opt-pred}}$ in (6.4) is optimal for predicting Y^*. We assume the unicity of the minimizer in the definition (6.4) as used in Tuo and Wu (2015); Plumlee (2017). Our theoretical results can be extended to the non-unicity case with similar arguments. We will introduce an algorithm in Section 6.4 to estimate $\theta^{\text{opt-pred}}$ from the training data. The formulation of $\theta^{\text{opt-pred}}$ in (6.4) is frequentist. We will discuss in Section 6.4.1 the differences between $\theta^{\text{opt-pred}}$ and other frequentist calibration methods, including the L_2-calibration method in Tuo and Wu (2015) and the least square calibration method in Wong et al. (2017).

6.3 Main Results

Existing theory for calibration, including Tuo and Wu (2015) and Wong et al. (2017), adopts the asymptotic arguments that the number of observations of physical experiment goes to infinity. In Section 6.3.1, we present a non-asymptotic minimax risk for nonparametric models and apply it to the model calibration problem in Section 6.3.2. In Section 6.3.3, we show the improvement in prediction achieved by incorporating data from computer models.

6.3.1 Non-Asymptotic Minimax Theory for Nonparametric Regressions

We consider the nonparametric regression in (6.1), where $\zeta(\cdot)$ resides in the RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$. Let $K : \Omega \times \Omega \to \mathbb{R}$ be a Mercer kernel generating $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$. By the spectral theorem, K admits the eigenvalue decomposition:

$$K(x, x') = \sum_{\nu \geq 1} \lambda_\nu \phi_\nu(x) \phi_\nu(x'),$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues and $\{\phi_\nu : \nu \geq 1\}$ are the corresponding eigenfunctions such that $\langle \phi_\nu, \phi_{\nu'} \rangle_{L_2(\Pi)} = \delta_{\nu\nu'}$. Here, $\delta_{\nu\nu'}$ is the Kronecker delta. We assume the polynomial decay rate of eigenvalues in Assumption 2. This assumption is commonly used and holds for Sobolev space $\mathcal{H} = \mathcal{W}_2^m(\Omega)$ with Lebesgue measure on Ω (Wahba, 1990).

Assumption 2. *For $2m > d$, suppose that for any $\nu \geq 1$, the eigenvalues satisfy $c_\lambda \nu^{-2m/d} \leq \lambda_\nu \leq C_\lambda \nu^{-2m/d}$ with constants $0 < c_\lambda < C_\lambda < \infty$, and the eigenfunctions are uniformly bounded: $\max_{x \in \Omega} |\phi_\nu(x)| \leq c_\phi$ with a constant c_ϕ for any $\nu \geq 1$.*

Our main results in Theorems 6.1 and 6.2, which find an explicit form of the constant besides the well-known rate in the finite-sample regime. We first show an upper bound of non-asymptotic estimation error for regularized estimators in the RKHS:

$$\widehat{\zeta}_{n\lambda} = \underset{g \in \mathcal{H}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(X_i)]^2 + \lambda \|g\|_{\mathcal{H}}^2 \right\}, \tag{6.5}$$

where $\lambda > 0$ is the smoothing parameter.

Theorem 6.1. *Under the regression model (6.1) where $\zeta \in \mathcal{H}$ and Assumption 2 holds, there exists a constant $0 < C_* < \infty$ not depending on $n, \sigma, m, d, \|\zeta\|_{\mathcal{H}}$ such that for any $n \geq 1$ and $\alpha_0 = 3.36$, with probability at least 99.99%,*

$$\|\widehat{\zeta}_{n\lambda} - \zeta\|^2_{L_2(\Pi)} \leq C_* \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{\|\zeta\|_{\mathcal{H}}}\right)^{-\frac{2d}{2m+d}}\right]^2 \cdot \alpha_0^{\frac{4m}{2m+d}} n^{-\frac{2m}{2m+d}} \left(\|\zeta\|_{\mathcal{H}} + \sigma\right)^2 \left(1 + \frac{\sigma}{\|\zeta\|_{\mathcal{H}}}\right)^{-\frac{2d}{2m+d}}.$$

Here, $\widehat{\zeta}_{n\lambda}$ is defined by (6.5) and λ is chosen appropriately.

We relegate the proof of Theorem 6.1 to the Appendix. The proof is established by using results from empirical processes such as the maximal inequalities and concentration inequalities (Kosorok, 2008), and deriving some new techniques to address the finite sample and key quantities of interest $\|\zeta\|_{\mathcal{H}}$ and σ. The probability 99.99% can be improved to any probability that closer to 100% by increasing α_0. The formula for appropriately chosen λ is given in the Appendix. In practice, λ can be estimated by the method of generalized cross-validation (GCV) (Craven and Wahba, 1978), which does not need knowledge of σ or the RKHS norm of ζ. We give details on using GCV in Section 6.4.

We now establish that the non-asymptotic risk achieved by $\widehat{\zeta}_{n\lambda}$ in Theorem 6.1 is minimax optimal. Here, the minimax optimality is in the sense that there exists a data generating process, for which the lower bound of non-asymptotic risk in the worst-case scenario exactly matches the upper bound derived in Theorem 6.1.

Theorem 6.2. *Under the regression model (6.1) where $\zeta \in \mathcal{H}$ and Assumption 2 holds, there exist*

constants $\sigma_0, n_0 \in (0,\infty)$ such that for any $\sigma \geq \sigma_0$ and $n \geq n_0$,

$$\infty_{\widetilde{\zeta}_n} \sup_{\zeta\in\mathcal{H}} \mathbb{P}\left\{ \|\widetilde{\zeta}_n - \zeta\|^2_{L_2(\Pi)} \geq C_* \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1+\frac{\sigma}{\|\zeta\|_{\mathcal{H}}}\right)^{-\frac{2d}{2m+d}}\right]^2 \right.$$
$$\left. \cdot \alpha_0^{\frac{4m}{2m+d}} n^{-\frac{2m}{2m+d}} \left(\|\zeta\|_{\mathcal{H}}+\sigma\right)^2 \left(1+\frac{\sigma}{\|\zeta\|_{\mathcal{H}}}\right)^{-\frac{2d}{2m+d}} \right\} > 0.$$

Here, the infimum is taken over all estimators $\widetilde{\zeta}_n$ that are measurable functions of data $\{(X_i, Y_i) : i = 1,\ldots,n\}$, and the constants C_*, α_0 are defined in Theorem 6.1.

The proof of this theorem is given in the Appendix. It is based on Fano's lemma and our new developments to address key quantities of interest $\|\zeta\|_{\mathcal{H}}$ and σ besides the rate.

We make three remarks on this theorem. First, the conventional asymptotic convergence rate can be recovered from Theorem 6.1. Since $2m > d$ in Assumption 2,

$$\alpha^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1+\frac{\sigma}{\|\zeta\|_{\mathcal{H}}}\right)^{-\frac{2d}{2m+d}} = o(1), \quad \text{as } n \to \infty.$$

Thus Theorem 6.1 yields that $\|\widehat{\zeta}_{n\lambda} - \zeta\|^2_{L_2(\Pi)} = O_{\mathbb{P}}\{n^{-2m/(2m+d)}\}$ as $n \to \infty$. This rate is well known (Cox, 1984; Wahba, 1990).

Second, Theorems 6.1 and 6.2 together immediate imply that the minimax optimal risk for estimating $\zeta \in \mathcal{H}$ in the finite-sample regime is

$$\|\widetilde{\zeta}_n - \zeta\|^2_{L_2(\Pi)} = C_* \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1+\frac{\sigma}{\|\zeta\|_{\mathcal{H}}}\right)^{-\frac{2d}{2m+d}}\right]^2 \cdot \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m}{2m+d}} \left(\|\zeta\|_{\mathcal{H}}+\sigma\right)^2 \left(1+\frac{\sigma}{\|\zeta\|_{\mathcal{H}}}\right)^{-\frac{2d}{2m+d}}, \tag{6.6}$$

where the constant C_* is defined in Theorem 6.1. This non-asymptotic minimax risk indicates the dependence on the signal-to-noise ratio $\|\zeta\|_{\mathcal{H}}/\sigma$ and the magnitude of signal and noise $\|\zeta\|_{\mathcal{H}} + \sigma$.

Third, we point out a tradeoff between approximation errors and prediction errors of the regularized estimator (6.5) in the finite-sample regime. A smaller λ compared to

the appropriately chosen λ in Theorem 6.1, corresponds to a smaller approximation error $\frac{1}{n}\sum_{i=1}^{n}[Y_i - \widehat{\zeta}_{n\lambda}(X_i)]^2$ (Wahba, 1990). On the other hand, a smaller λ yields a large risk for prediction $\|\widehat{\zeta}_{n\lambda} - \zeta\|^2_{L_2(\Pi)}$ as implied by the proof of Theorem 6.1.

6.3.2 Optimal Calibration and Prediction

We apply the non-asymptotic minimax theory in Section 6.3.1 to derive an equivalent form for the optimal calibration in (6.4).

Proposition 6.3. *Under Assumptions 1 and 2, the optimal calibration defined in (6.4) is equivalent to finding the minimizer of model discrepancy under the RKHS norm:*

$$\theta^{\textit{opt-pred}} = \arg\min_{\theta\in\Theta} \left\{ \|\zeta(\cdot) - \eta(\cdot,\theta)\|_{\mathcal{H}} \right\}.$$

Based on this proposition, we derive an algorithm to estimate $\theta^{\text{opt-pred}}$ from a given dataset in Section 6.4. The proof of Proposition 6.3 is given in the Appendix. Here are some explanations. If the model discrepancy $\zeta(\cdot) - \eta(\cdot,\theta)$ has a small RKHS norm, the discrepancy is a simple function in the RKHS. Since the number of data points from (6.1) is *limited*, a simpler function should have a more accurate estimator. As discussed in Section 6.2, a more accurate discrepancy estimator gives a smaller prediction error for the physical system. Proposition 6.3 justifies the use of RKHS norm to measure the model discrepancy with theoretical guarantees. This procedure is different from using L_2-norm (Tuo and Wu, 2015) and empirical l_2-norm (Wong et al., 2017).

We now discuss the non-asymptotic minimax risk for predicting $\zeta(\cdot)$ based on $\theta^{\text{opt-pred}}$. Recall that the model discrepancy $\delta(\cdot,\theta) = \zeta(\cdot) - \eta(\cdot,\theta)$. We define regularized model discrepancy estimators in the RKHS as

$$\widehat{\delta}_{n\lambda}(\cdot,\theta) = \arg\min_{h\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n}[Y_i - \eta(X_i,\theta) - h(X_i)]^2 + \lambda\|h\|^2_{\mathcal{H}} \right\}, \tag{6.7}$$

where $\lambda > 0$ is the smoothing parameter. As a corollary of Theorem 6.1, we present an

upper bound of predictive mean squared errors when using computer model calibrated by $\theta^{\text{opt-pred}}$ and regularized discrepancy estimators.

Corollary 6.4. *Under the regression model (6.1) where $\zeta \in \mathcal{H}$ and Assumptions 1 and 2 hold, then for any $n \geq 1$, with probability at least 99.99%,*

$$\begin{aligned}
&\left\| \left[\eta(\cdot, \theta^{\textit{opt-pred}}) + \widehat{\delta}_{n\lambda}(\cdot, \theta^{\textit{opt-pred}}) \right] - \zeta(\cdot) \right\|_{L_2(\Pi)}^2 \\
&\leq C_* \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{\min_{\theta\in\Theta} \|\delta(\cdot,\theta)\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right]^2 \\
&\quad \cdot \alpha_0^{\frac{4m}{2m+d}} n^{-\frac{2m}{2m+d}} \left(\min_{\theta\in\Theta} \|\delta(\cdot,\theta)\|_{\mathcal{H}} + \sigma \right)^2 \left(1 + \frac{\sigma}{\min_{\theta\in\Theta} \|\delta(\cdot,\theta)\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}}.
\end{aligned}$$

Here, $\widehat{\delta}_{n\lambda}$ is defined by (6.7), and λ is chosen appropriately, and the constants C_*, α_0 are defined in Theorem 6.1.

The formula for an appropriately chosen λ in Corollary 6.4 is provided in Appendix. In practice, λ can be estimated using the GCV in Section 6.4. As a corollary of Theorem 6.2, the following result gives a ower bound of predictive mean squared errors when using the computer model calibrated by $\theta^{\text{opt-pred}}$ and any estimators of model discrepancy.

Corollary 6.5. *Under the regression model (6.1) where $\zeta \in \mathcal{H}$ and Assumptions 1 and 2 hold, there exist constants $\sigma_0', n_0' \in (0, \infty)$ such that for any $\sigma \geq \sigma_0'$ and $n \geq n_0'$,*

$$\begin{aligned}
&\infty_{\widetilde{\delta}_n} \sup_{\zeta\in\mathcal{H}} \mathbb{P} \left\{ \left\| \left[\eta(\cdot, \theta^{\textit{opt-pred}}) + \widetilde{\delta}_n(\cdot) \right] - \zeta(\cdot) \right\|_{L_2(\Pi)}^2 \right. \\
&\geq C_* \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{\min_{\theta\in\Theta} \|\delta(\cdot,\theta)\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right]^2 \alpha_0^{\frac{2m-d}{2m+d}} \\
&\left. \cdot n^{-\frac{2m}{2m+d}} \left(\min_{\theta\in\Theta} \|\delta(\cdot,\theta)\|_{\mathcal{H}} + \sigma \right)^2 \left(1 + \frac{\sigma}{\min_{\theta\in\Theta} \|\delta(\cdot,\theta)\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right\} > 0.
\end{aligned}$$

Here, the infimum is taken over all estimators $\widetilde{\delta}_n$ that are measurable functions of data $\{(X_i, Y_i - \eta(X_i, \theta^{\textit{opt-pred}})) : i = 1, \ldots, n\}$, and the constants C_*, α_0 are defined in Theorem 6.1.

We give the proof of Corollaries 6.4 and 6.5 in the Appendix. These corollaries together imply that the non-asymptotic minimax optimal risk for predicting $\zeta \in \mathcal{H}$ when using the computer model calibrated by $\theta^{\text{opt-pred}}$ and an estimator of model discrepancy is

$$\begin{aligned}
&\left\| \left[\eta(\cdot, \theta^{\text{opt-pred}}) + \widetilde{\delta}_n(\cdot) \right] - \zeta(\cdot) \right\|^2_{L_2(\Pi)} \\
&= C_* \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{\min_{\theta\in\Theta} \|\delta(\cdot,\theta)\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right]^2 \alpha_0^{\frac{4m}{2m+d}} \\
&\quad \cdot n^{-\frac{2m}{2m+d}} \left(\min_{\theta\in\Theta} \|\delta(\cdot,\theta)\|_{\mathcal{H}} + \sigma \right)^2 \left(1 + \frac{\sigma}{\min_{\theta\in\Theta} \|\delta(\cdot,\theta)\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}}.
\end{aligned}$$

This minimax risk is minimal over $\theta \in \Theta$ and it can be achieved by

$$\zeta_{n\lambda}^{\text{opt-pred}}(\cdot) \stackrel{\text{def}}{=} \eta(\cdot, \theta^{\text{opt-pred}}) + \widehat{\delta}_{n\lambda}(\cdot, \theta^{\text{opt-pred}}), \tag{6.8}$$

where $\widehat{\delta}_{n\lambda}$ is defined by (6.7).

6.3.3 Improved Prediction Using the Computer Model

The minimax optimal predictor (6.8) combines the merits of the parametric computer model and the nonparametric discrepancy estimator. We compare (6.8) with its counterpart of the minimax optimal predictor (6.5) without using the computer model.

Theorem 6.6. *Under the regression model (6.1) where $\zeta \in \mathcal{H}$ and Assumptions 1 and 2 hold, if*

$$\min_{\theta\in\Theta} \|\zeta(\cdot) - \eta(\cdot,\theta)\|_{\mathcal{H}} < \|\zeta\|_{\mathcal{H}}, \tag{6.9}$$

then $\zeta_{n\lambda}^{\text{opt-pred}}(\cdot)$ defined by (6.8) with the aid of the computer model achieves a smaller minimax risk than $\widehat{\zeta}_{n\lambda}(\cdot)$ defined by (6.5) without using the information of the computer model.

The proof of this theorem is given in the Appendix. We make three remarks on this theorem. First, the computer model $\eta(\cdot,\theta)$ is built based on some physics knowledge of the system $\zeta(\cdot)$ and Proposition 6.3 shows that $\eta(\cdot, \theta^{\text{opt-pred}})$ is the best approximation to $\zeta(\cdot)$

within the family $\{\eta(\cdot,\theta), \theta \in \Theta\}$. Although the computer model is imperfect for modeling the physical system, $\eta(\cdot,\theta^{\text{opt-pred}})$ can still capture some major shape of $\zeta(\cdot)$ and consequently $\zeta(\cdot) - \eta(\cdot,\theta^{\text{opt-pred}})$ has less variation or smoother in $\mathcal{H}$ than the original $\zeta(\cdot)$ does. This motivates the assumption of $\min_{\theta\in\Theta} \|\zeta(\cdot) - \eta(\cdot,\theta)\|_{\mathcal{H}} = \|\zeta(\cdot) - \eta(\cdot,\theta^{\text{opt-pred}})\|_{\mathcal{H}} < \|\zeta\|_{\mathcal{H}}$ in (6.9).

Second, Theorem 6.6 indicates that it is statistically more efficient to learn the residual function $\zeta(\cdot) - \eta(\cdot,\theta)$ than to learn the original unreferenced function $\zeta(\cdot)$.

Third, the predictor (6.8) is a parametrically-guided nonparametric predictor, where $\eta(\cdot,\theta)$ can have a parametric form.

6.4 An Algorithm for Optimal Calibration

We propose an algorithm to compute the optimal calibration $\theta^{\text{opt-pred}}$ defined in (6.4). From Proposition 6.3,

$$\theta^{\text{opt-pred}} = \arg\min_{\theta\in\Theta} \left\{ \|\delta(\cdot,\theta)\|_{\mathcal{H}}^2 \right\}, \tag{6.10}$$

where the discrepancy $\delta(\cdot,\theta)$ is subject to the constraint (6.2). By evaluating (6.2) at the training data from model (6.1),

$$Y_i = \zeta(X_i) + \varepsilon_i = \eta(X_i,\theta) + \delta(X_i,\theta) + \varepsilon_i, \ \ \forall\theta \in \Theta, \ i = 1,\ldots,n. \tag{6.11}$$

Using the Lagrange multiplier method for the optimization (6.10) with constraint (6.11), we find $\theta \in \Theta$ and $\delta(\cdot) \in \mathcal{H}$ by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - \eta(X_i,\theta) - \delta(X_i)]^2 + \lambda\|\delta\|_{\mathcal{H}}^2, \quad \lambda > 0. \tag{6.12}$$

Here, the tuning of λ is critical for achieving good predictions for $\zeta(\cdot)$. We provide more discussions in Section.

The optimization problem in (6.12) can be solved iteratively as follows. We introduce some notation. Recall that K is the reproducing kernel of $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$. Let Σ be the

$n \times n$ kernel matrix with ijth entry $K(X_i, X_j)$. Denote by $\overrightarrow{Y} = (Y_1, \ldots, Y_n)^\top$, $\eta(\overrightarrow{X}, \theta) = (\eta(X_1, \theta), \ldots, \eta(X_n, \theta))^\top$, $\delta(\overrightarrow{X}, \theta) = (\delta(X_1, \theta), \ldots, \delta(X_n, \theta))^\top$, and $\zeta(\overrightarrow{X}) = (\zeta(X_1), \ldots, \zeta(X_n))^\top$.

For any fixed $\theta \in \Theta$, the minimizer $\delta(\cdot)$ of (6.12) is the same as the regularized estimator $\widehat{\delta}_{n\lambda}(\cdot, \theta)$ in (6.7). By the representer lemma (Kimeldorf and Wahba, 1971),

$$\widehat{\delta}_{n\lambda}(\cdot, \theta) = \sum_{i=1}^{n} c_i K(X_i, \cdot), \tag{6.13}$$

where the coefficient $c = (c_1, \ldots, c_n)^\top \in \mathbb{R}^n$ is given by

$$c = c(\theta) = (\Sigma + n\lambda \mathbf{I})^{-1}[\overrightarrow{Y} - \eta(\overrightarrow{X}, \theta)]. \tag{6.14}$$

In practice, $\sigma^2 = \mathbb{E}[\varepsilon^2]$ is not generally known and λ can be estimated by GCV (Craven and Wahba, 1978). Let $A(\lambda)$ be the influence matrix satisfying $\widehat{\delta}_{n\lambda}(\overrightarrow{X}, \theta) = A(\lambda)[\overrightarrow{Y} - \eta(\overrightarrow{X}, \theta)]$. The GCV estimate of λ is the minimizer of

$$\text{GCV}(\lambda) = \frac{n^{-1} \| \overrightarrow{Y} - \eta(\overrightarrow{X}, \theta) - \widehat{\delta}_{n\lambda}(\overrightarrow{X}, \theta) \|^2}{[n^{-1} \mathbf{tr}(\mathbf{I} - A(\lambda))]^2}. \tag{6.15}$$

The GCV estimate is consistent for minimizing the mean squared errors in Theorem 6.1 and Corollary 6.4 (see, Li (1986); Wahba (1990)).

For any fixed $\delta(\cdot) = \widehat{\delta}_{n\lambda}(\cdot, \theta)$ from (6.13), the minimizer θ of (6.12) is equivalent to

$$\arg\min_{\theta \in \Theta} \left\{ (\overrightarrow{Y} - \eta(\overrightarrow{X}, \theta))^\top (\Sigma + n\lambda \mathbf{I})^{-1} (\overrightarrow{Y} - \eta(\overrightarrow{X}, \theta)) \right\}. \tag{6.16}$$

Since the objective function in (6.16) is a weighted version of the empirical l_2-norm, (6.16) gives a different calibration result than the least square calibration method (6.17).

Putting the above building blocks together, our algorithm for optimizing (6.12) iterates between (6.13) and (6.16). The algorithm shares a similar spirit as the coordinate descent method (Wright, 2015). The algorithm can start with the calibration parameters from the least square calibration method. Applying later iterations of the algorithm continuously

improves the initial values for prediction. This procedure is summarized in Algorithm 1 with a prespecified value $\tau > 0$, e.g., $\tau = 10^{-3}$. Our experience indicates that a small number of iterations is sufficient to obtain good performance of the algorithm.

Algorithm 1 Optimal calibration for prediction

1: **Input:** Noisy data $\{(X_i, Y_i) : i = 1, \ldots, n\}$ and computer model $\eta(\cdot, \theta)$
2: **Initialize:** Solve the least square calibration (6.17). Let $\theta^{(0)} = \widehat{\theta}_n^{l_2}$
3: **for** $k = 1, 2, \ldots$ until $\|\theta^{(k)} - \theta^{(k-1)}\|_{l_2} \leq \tau$ **do**
4: Solve for $\widehat{\delta}_{n\lambda}(\cdot, \theta^{(k-1)})$ in (6.13) and tune λ by GCV
5: With the selected λ, update θ by (6.16) and obtain $\theta^{(k)}$
6: **end for**
7: Let $\widehat{\theta}_n^{\text{opt-pred}} = \theta^{(k)}$. Solve for $\widehat{\delta}_{n\lambda}(\cdot, \widehat{\theta}_n^{\text{opt-pred}})$ in (6.13) and tune λ by GCV
8: **Output:** Calibration parameter $\widehat{\theta}_n^{\text{opt-pred}}$ and optimal predictor $\eta(\cdot, \widehat{\theta}_n^{\text{opt-pred}}) + \widehat{\delta}_{n\lambda}(\cdot, \widehat{\theta}_n^{\text{opt-pred}})$

Proposition 6.7. *The estimated optimal calibration $\widehat{\theta}_n^{\text{opt-pred}}$ from Algorithm 1 is consistent: $\widehat{\theta}_n^{\text{opt-pred}} \to_{\mathbb{P}} \theta^{\text{opt-pred}}$.*

Numerical examples in Section 6.5 show that $\widehat{\theta}_n^{\text{opt-pred}}$ outperform existing frequentist and Bayesian calibrations in term of prediction with finite samples. We provide theoretical comparisons in Sections 6.4.1 and 6.4.2.

6.4.1 Comparison to Existing Frequentist Calibration Methods

We compare the calibration $\widehat{\theta}_n^{\text{opt-pred}}$ obtained by Algorithm 1 with two other frequentist calibration methods: the L_2-calibration and the least square calibration. The L_2-calibration method in Tuo and Wu (2015) is defined as follows:

$$\widehat{\theta}_n^{L_2} = \arg\min_{\theta \in \Theta} \left\{ \|\widehat{\zeta}_{n\lambda}(\cdot) - \eta(\cdot, \theta)\|_{L_2(\Pi)} \right\},$$

where $\widehat{\zeta}_{n\lambda}(\cdot)$ is defined by (6.5). The least square calibration method in Wong et al. (2017) minimizes the model discrepancy equipped with the empirical l_2-norm:

$$\widehat{\theta}_n^{l_2} = \arg\min_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} [Y_i - \eta(X_i, \theta)]^2 \right\}. \tag{6.17}$$

Wong et al. (2017) estimates the model discrepancy after calibrating $\theta = \widehat{\theta}_n^{l_2}$:

$$\widehat{\delta}_{n\lambda}(\cdot, \widehat{\theta}_n^{l_2}) = \arg\min_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} [Y_i - \eta(X_i, \widehat{\theta}_n^{l_2}) - h(X_i)] + \lambda \|h\|_{\mathcal{H}}^2 \right\},$$

and a predictor for $\zeta(\cdot)$ is given by $\eta(\cdot, \widehat{\theta}_n^{l_2}) + \widehat{\delta}_{n\lambda}(\cdot, \widehat{\theta}_n^{l_2})$.

Remark 6.8. *The differences among $\widehat{\theta}^{\text{opt-pred}}$, $\widehat{\theta}_n^{L_2}$, and $\widehat{\theta}_n^{l_2}$ are as follows.*

- *Calibration. $\widehat{\theta}_n^{\text{opt-pred}} \to_{\mathbb{P}} \theta^{\text{opt-pred}} = \arg\min_{\theta \in \Theta} \{\|\zeta(\cdot) - \eta(\cdot, \theta)\|_{\mathcal{H}}\}$ and $\widehat{\theta}_n^{L_2}, \widehat{\theta}_n^{l_2} \to_{\mathbb{P}} \theta^{L_2} \stackrel{\text{def}}{=} \arg\min_{\theta \in \Theta} \{\|\zeta(\cdot) - \eta(\cdot, \theta)\|_{L_2(\Pi)}\}$. The calibration $\widehat{\theta}_n^{\text{opt-pred}}$ is different from $\widehat{\theta}_n^{L_2}$ and $\widehat{\theta}_n^{l_2}$.*
- *Prediction. For a finite sample, the predictors based on $\theta^{\text{opt-pred}}$ achieve a smaller minimax predictive mean squared error compared to predictors based on θ^{L_2}.*

We provide a proof of Remark 6.8 in the Appendix.

6.4.2 Connection to Existing Bayesian Calibration Methods

The connection between frequentist and Bayesian calibrations was first established by Tuo and Wu (2016, 2018). Specifically, Tuo and Wu (2018) derives that under Gaussian process priors, the maximum a posteriori (MAP) estimate of θ and $\delta(\cdot)$ in Bayesian calibration of Kennedy and O'Hagan (2001) agrees with the minimizer of the objective function (6.12) if $\lambda = \sigma^2/n\beta$. Similar results for the posterior mean of θ and $\delta(\cdot)$ have appeared in the literature (e.g., Kimeldorf and Wahba (1971); Wahba (1990); Santner et al. (2003)), where details are provided in the Appendix for completeness. The original work of Tuo and Wu (2016) shows that the MLE of θ is the same as (6.10) if physical data are noiseless ($\sigma = 0$).

We add a few new remarks on the differences between our method and the Bayesian calibration of Kennedy and O'Hagan (2001). First, $\lambda = \sigma^2/n\beta$ is unknown in practice. Kennedy and O'Hagan (2001) proposes to place a noninformative prior on β and use Markov chain Monte Carlo sampler (Geman and Geman, 1984) to draw β, σ, θ, and $\delta(\cdot, \theta)$. However, this approach cannot guarantee the orthogonality in the sense that $\langle \zeta(\cdot) - \eta(\cdot, \theta), \partial\eta(\cdot, \theta)/\partial\theta \rangle_{\mathcal{H}} \neq 0$

even as n gets very large (Plumlee, 2017). Our $\theta^{\text{opt-pred}}$ satisfies the orthogonality by Proposition 6.3 and the KKT condition.

Second, there is a lack of identifiability in Bayesian calibration of Kennedy and O'Hagan (2001); see, e.g., Gramacy et al. (2015). We note in Kennedy and O'Hagan (2001) that each random sampling of β and σ gives different calibration θ, which corresponds to different predictions (see, Section 6.2). On the other hand, the orthogonality property of $\theta^{\text{opt-pred}}$ ensures to avoid the identifiability issue. Given the consistency result of Proposition 6.7, our calibration $\widehat{\theta}^{\text{opt-pred}}$ has smaller prediction errors and variances than Bayesian calibration of Kennedy and O'Hagan (2001). This observation is corroborated by the theory in Section 6.3 and the numerical examples in Section 6.5.

Third, the physical data are generally noisy. The original result for $\sigma = 0$ in Tuo and Wu (2016) cannot be generalized to $\sigma \neq 0$. Our contributions include to provide justifications that $\theta^{\text{opt-pred}}$ gives the minimax optimal predictions in the finite-sample regime if $\sigma \neq 0$ (see, Corollaries 6.4 and 6.5).

The Bayesian calibration method is more time consuming to compute than frequentist calibration methods such as ours. Different from Kennedy and O'Hagan (2001) and our $\theta^{\text{opt-pred}}$, Plumlee (2017) suggests constructing new priors for $\delta(\cdot, \theta)$ in order to satisfy the orthogonality. Plumlee (2017) requires computing the gradient of (estimated) computer model while the proposed $\theta^{\text{opt-pred}}$ does not.

6.5 Simulation and Real Examples

We illustrate the prediction accuracy of the proposed calibration method using several examples. Our simulation study consists of Examples 6.9–6.11, where the tuning parameters for regularized estimators are selected by the GCV. The prediction accuracy is measured by the predictive mean squared error estimated by a Monte Carlo sample of $1,000,000$ test data, where the designs are drawn from the same distribution as training data. A real data example is given in Example 6.12.

Example 6.9. *Consider a physical system* $\zeta(x) = \exp(\pi x/5)\sin 2\pi x, x \in [0,1]$. *The physical data are generated by (6.1) with* $X_i \sim$ *Unif*$([0,1]), \varepsilon_i \sim \mathcal{N}(0,\sigma^2)$ *for* $i = 1,\ldots,50$. *Four different noise variances are investigated:* $\sigma^2 = 0.1, 0.25, 0.5, 1$. *Suppose that the computer model is*

$$\eta(x,\theta) = \zeta(x) - \sqrt{\theta^2-\theta+1}(\sin 2\pi\theta x + \cos 2\pi\theta x) \text{ for } \theta \in [-1,1].$$

Since $\theta^2 - \theta + 1 \geq 3/4$ *for any* $-1 \leq \theta \leq 1$, *the model discrepancy between* $\eta(\cdot,\theta)$ *and* $\zeta(\cdot)$ *always exists no matter how* θ *is chosen. We use the Matérn kernel* $K(x_1,x_2) = (1+|x_1 - x_2|/\psi)\exp\{-|x_1-x_2|/\psi\}$, *where the scale parameter* ψ *is chosen by the five-fold cross-validation minimization (see, e.g., Efron and Tibshirani (1993)). Figure 6.1 plots the squared model discrepancy with different norms:* $\|\zeta(\cdot) - \eta(\cdot,\theta)\|^2_{L_2(\Pi)}$ *and* $\|\zeta(\cdot) - \eta(\cdot,\theta)\|^2_{\mathcal{H}}$. *The corresponding minimizers are different given as* $\theta^{L_2} \approx -0.1780$ *and* $\theta^{\text{opt-pred}} \approx 0.3740$ *(a local minimizer of* $\|\zeta(\cdot) - \eta(\cdot,\theta)\|^2_{\mathcal{H}}$ *in* $[-0.4, 0]$ *is* $\theta^{\text{opt-pred}} \approx -0.1230$*), which illustrates the first part of Remark 6.8.*

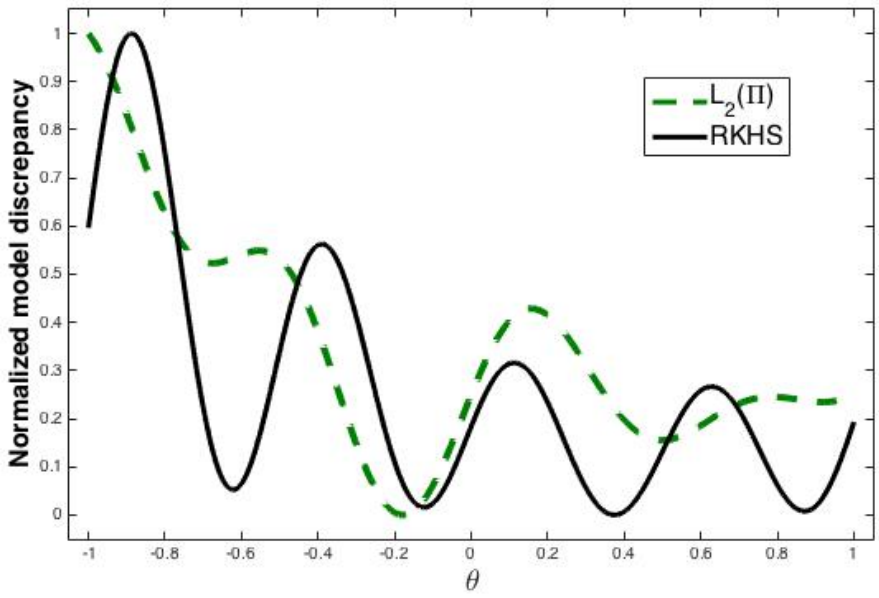

Figure 6.1: Normalized model discrepancy equipped with $L_2(\Pi)$-norm and RKHS-norm in Example 6.9.

We compare the prediction accuracy of four frequentist calibration methods:

1) *The computer model with L_2-calibration (abbreviated as "No Bias Corr."), which is the $\eta(\cdot, \widehat{\theta}_n^{L_2})$ in Section 6.4.1 without the model discrepancy correction;*

2) *The nonparametric predictor (abbreviated as "N.P."), which is the $\widehat{\zeta}_{n\lambda}(\cdot)$ obtained by (6.5) in Section 6.3.1;*

3) *The predictor in Wong et al. (2017) (abbreviated as "LS. Cal."), which is the $\eta(\cdot, \widehat{\theta}_n^{l_2}) + \widehat{\delta}_{n\lambda}(\cdot, \widehat{\theta}_n^{l_2})$ in Section 6.4.1 based on the least square calibration;*

4) *Our predictor (abbreviated as "Opt. Cal."), which is the $\zeta_{n\lambda}^{\text{opt-pred}}$ in Section 6.3.2 based on the proposed calibration and computed by Algorithm 1 in Section 6.4.*

For each chosen noise variance, we replicate the data generation, calibration and prediction procedures 1,000 times and average the results for each method across the replicates. The resulting average predictive mean squared errors and its associated standard errors are given in Table 6.1. The computer model with L_2-calibration (i.e., "No Bias Corr.") gives the largest predictive mean squared error, which shows that an estimator for the model discrepancy is necessary. "No Bias Corr." has the smallest standard error which is not surprising because the parametric estimation of θ here has a faster rate of convergences than nonparametric estimations required for other three methods. Table 6.1 indicates that our method "Opt. Cal." and the predictor in Wong et al. (2017) "LS. Cal." outperform the nonparametric predictor "N.P.". This advantage illustrates the improved predictions by computer models as indicated in Theorem 6.6. Furthermore, "Opt. Cal." gives smaller prediction errors than "LS. Cal.", which agrees with the second part of Remark 6.8. Overall, for a finite sample size $n = 50$, the proposed method "Opt. Cal." outperforms the other frequentist predictors in the settings studied.

Example 6.10. *Consider a two-dimensional physical system*

$$\begin{aligned}\zeta(x_1, x_2) = &\frac{2}{3}\exp(x_1 + 0.2) - x_2 \sin 0.4 + 0.4 \\ &+ \exp(-x_1)\left(x_1 + \frac{1}{2}\right)\left(x_2^2 + x_2 + 1\right), \quad (x_1, x_2) \in [0, 1]^2.\end{aligned}$$

Table 6.1: Comparison of predictive mean squared errors for Example 6.9. PMSE = predictive mean squared error, SE = standard error

	$\sigma^2 = 0.1$		$\sigma^2 = 0.25$	
	Average of PMSE	SE of PMSE	Average of PMSE	SE of PMSE
No Bias Corr.	0.3492	0.0174	0.3601	0.0270
N.P.	0.1701	0.0594	0.2327	0.0680
LS. Cal.	0.1152	0.0604	0.1693	0.0580
Opt. Cal.	**0.0922**	0.0515	**0.1434**	0.0552
	$\sigma^2 = 0.5$		$\sigma^2 = 1$	
	Average of PMSE	SE of PMSE	Average of PMSE	SE of PMSE
No Bias Corr.	0.3744	0.0407	0.3985	0.0666
N.P.	0.2810	0.0930	0.3417	0.1367
LS. Cal.	0.1924	0.0683	0.2074	0.0853
Opt. Cal.	**0.1684**	0.0656	**0.1838**	0.0788

Suppose that the computer model is

$$\eta(x_1, x_2; \theta_1, \theta_2) = \frac{2}{3}\exp(x_1 + \theta_1) - x_2 \sin\theta_2 + \theta_2, \quad (\theta_1, \theta_2) \in [0, 1]^2.$$

The model discrepancy exists between $\eta(\cdot; \theta_1, \theta_2)$ and $\zeta(\cdot)$. Assume the physical data are generated by (6.1) with a uniform design on $[0, 1]^2$ and $n = 50$. We consider four levels of $\sigma^2 = 0.03, 0.05, 0.07, 0.1$, and use the Matérn kernel $K(x_1, x_2) = (1 + \|x_1 - x_2\|/\psi)\exp\{-\|x_1 - x_2\|/\psi\}$ with ψ chosen by the five-fold cross-validation minimization. For each level of σ^2, we replicate the data generation, calibration and prediction procedures 1,000 times for all the methods and average the results for each method across the replicates. Table 6.2 summarizes the results. It is clear that the proposed method "Opt. Cal." outperforms the other three frequentist calibration methods in terms of the predictive mean squared error.

Example 6.11. *We now compare the proposed calibration method with some Bayesian calibration*

Table 6.2: Comparison of predictive mean squared errors for Example 6.10. PMSE = predictive mean squared error, SE = standard error

	$\sigma^2 = 0.03$		$\sigma^2 = 0.05$	
	Average of PMSE	SE of PMSE	Average of PMSE	SE of PMSE
No Bias Corr.	0.1690	0.0027	0.1691	0.0027
N.P.	0.1155	0.0512	0.1823	0.0850
LS. Cal.	0.0611	0.0198	0.0743	0.0212
Opt. Cal.	**0.0564**	0.0187	**0.0691**	0.0205
	$\sigma^2 = 0.07$		$\sigma^2 = 0.1$	
	Average of PMSE	SE of PMSE	Average of PMSE	SE of PMSE
No Bias Corr.	0.1692	0.0028	0.1694	0.0028
N.P.	0.2425	0.1159	0.3327	0.1453
LS. Cal.	0.0825	0.0224	0.0906	0.0235
Opt. Cal.	**0.0776**	0.0220	**0.0863**	0.0234

method. Consider a falling ball example in Plumlee (2017) where the physical system is

$$\zeta(x) = 8 + \frac{5}{2}\log\left(\frac{50}{49} - \frac{50}{49}\tanh\left(\tanh^{-1}(\sqrt{0.02}) + \sqrt{2}x\right)^2\right),\ x \in [0,1]$$

and the computer model derived from Newton's second law is $\eta(x; v_0, g) = 8 + v_0 x - g x^2/2$. Here, calibration parameters (v_0, g) are the vertical velocity and the acceleration rate, respectively. The model discrepancy exists between $\zeta(\cdot)$ and $\eta(\cdot; v_0, g)$. Suppose that the physical data are generated by (6.1) with a uniform design on $[0,1]$ and $n = 30$. We compare the proposed method "Opt. Cal." with two Bayesian methods in terms of prediction accuracy:

1) *The Bayesian method of Kennedy and O'Hagan (2001) (abbreviated as "KO");*

2) *The Bayesian predictor using an orthogonal Gaussian process in L_2-norm as the prior (abbreviated as "OGP") proposed by Plumlee (2017).*

The Matérn kernel $K(x_1, x_2) = (1 + |x_1 - x_2|/\psi)\exp\{-|x_1 - x_2|/\psi\}$ with parameter $\psi = 1$ is used as the reproducing kernel for "Opt. Cal." and the prior covariance function for both "KO"

and "OGP". Four levels of $\sigma^2 = 0.0025, 0.01, 0.0625, 0.25$ *are considered. For each level of* σ^2*, we replicate the data generation, calibration and prediction procedures 1,000 times. Table 6.3 summarizes the prediction results. Here "KO" has large prediction errors and large posterior variances compared with "OGP" and "Opt. Cal.". This is because of the identifiability issue inherent to "KO" as discussed in Section 6.4.2. "OGP" provides stable and accurate predictions and our method "Opt. Cal." gives even smaller prediction errors.*

Table 6.3: Comparison of predictive mean squared errors for Example 6.11. PMSE = predictive mean squared error, SE = standard error

	$\sigma^2 = 0.0025$		$\sigma^2 = 0.01$	
	Average of PMSE	SE of PMSE	Average of PMSE	SE of PMSE
KO	0.5413	2.3944	5.8596	29.0264
OGP	0.0147	0.0132	0.0230	0.0249
Opt. Cal.	**0.0091**	0.0084	**0.0166**	0.0168
	$\sigma^2 = 0.0625$		$\sigma^2 = 0.25$	
	Average of PMSE	SE of PMSE	Average of PMSE	SE of PMSE
KO	1.4644	4.3802	18.5433	97.3962
OGP	0.0500	0.0478	0.0952	0.0995
Opt. Cal.	**0.0318**	0.0338	**0.0620**	0.0677

Example 6.12 (Real data example)**.** *We analyze a real dataset from a single voltage clamp experiment on sodium ion channels of cardiac cell membranes. This dataset consists of* 19 *outputs and is from Plumlee (2017). The response variable is the normalized current for maintaining a fixed membrane potential of* $-35mV$ *and the input variable is the logarithm of time. Suppose the computer model for this experiment is the Markov model for sodium ion channels given by* $\eta(x, \theta) = e_1^\top \exp(\exp(x)A(\theta))e_4$*, where* $\theta = (\theta_1, \theta_2, \theta_3)^\top \in \mathbb{R}^3$*,* $e_1 = (1, 0, 0, 0)^\top$*,* $e_4 = (0, 0, 0, 1)^\top$*, and*

$$A(\theta) = \begin{pmatrix} -\theta_2 - \theta_3 & \theta_1 & 0 & 0 \\ \theta_2 & -\theta_1 - \theta_2 & \theta_1 & 0 \\ 0 & \theta_2 & -\theta_1 - \theta_2 & \theta_1 \\ 0 & 0 & \theta_2 & -\theta_1 \end{pmatrix}$$

For this example, we compare the frequentist methods in Example 6.9 and 6.10, and Bayesian methods in Example 6.11. The Matérn kernel $K(x_1, x_2) = (1+|x_1-x_2|/\psi)\exp\{-|x_1-x_2|/\psi\}$ *with* $\psi = 1$ *is used for all methods and the Metropolis-Hastings algorithm is applied to sample from the posterior of* θ *for Bayesian methods. In each experiment, we perform five-fold cross-validation minimization where the data is randomly partitioned into five roughly equal-sized parts: four parts are for training and the rest part is for testing. The cross-validation process is repeated five times, with each of the five parts is used once for testing. Then, the five predictive mean squared errors are averaged to give a single predictive mean squared error. We replicate the data generation, calibration and prediction procedure 100 times and average the results.*

Table 6.4 summarizes the prediction results, with the abbreviations of the methods given in Examples 6.9 and 6.11. "No Bias Corr." gives the largest predictive mean squared error among the four frequentist methods, indicating the existence of model discrepancy. Here all the frequentist methods outperform the two Bayesian methods. "N.P." outperforms "LS. Cal.". This indicates that if the calibration parameter is not chosen well, the use of computer model does not improve prediction. Overall, the proposed method 'Opt. Cal." gives the smallest prediction error among all the methods applied to this example.

Table 6.4: Comparison of predictive mean squared errors for Example 6.12. PMSE = predictive mean squared error, SE = standard error

	Average of PMSE	SE of PMSE
KO	0.0045	0.0131
OGP	9.4387×10^{-4}	0.0022
No Bias Corr.	4.2823×10^{-4}	6.2333×10^{-4}
N.P.	2.5916×10^{-4}	2.2522×10^{-4}
LS. Cal.	3.0521×10^{-4}	6.3797×10^{-4}
Opt. Cal.	$\mathbf{1.6323 \times 10^{-4}}$	2.1335×10^{-4}

Chapter 7

Discussions and Future Works

In this thesis, we study statistical methods for complex problems in five different settings. We discuss potential future work of the content of this thesis in a few important directions.

Chapter 2 have obtained new minimax optimal rates for nonparametric estimation when data from first-order partial derivatives are available. These results deal with function estimation and partial derivative estimation in functional ANOVA models. Statistical modeling of derivative model is an increasing important problem in engineering, economics and other fields. Our theoretical results provide justification why incorporation of partial derivatives can improve convergence rates in estimation. It would be of interest to incorporate derivative data to another type of functional ANOVA model in Stone (1994); Huang (1998). In particular, Stone (1994) studies sums of tensor products of polynomial splines (as opposed to the smoothing approach in ours) to estimate components of a functional ANOVA model and Huang (1998) investigates the projection estimate in fitting a functional ANOVA model. If the order of interactions $r = 1$ in (2.2), the results in Section 2.3–2.5 still hold for the functional ANOVA model in Stone (1994); Huang (1998). If $1 < r \leq d$ in (2.2), it requires more work to extend our results to cover the minimax rates of convergences with derivatives for the functional ANOVA model in Stone (1994); Huang (1998). Given the interplay between data collection and data modeling in applications like computer experiments, it would be interesting to connect our developed convergence results with the underlying structure of a

chosen design used for data generation.

Chapter 3 studies a framework to integrate longitudinal features from the structural MR images for AD prediction based on varying coefficient models. We propose a novel variable selection method by combining smoothing splines and Lasso, which enables simultaneous selection and estimation and is adaptive to heterogeneous longitudinal data. To illustrate the effectiveness of the proposed method, we conduct experiments with the ADNI dataset and show that the proposed method outperforms the state-of-the-art longitudinal analysis methods. Our work is the first in the literature to model nonlinear progressions of longitudinal features and propose a novel effective variable selection method for the high-dimensional setting. This method shows superior performance in real data AD prediction. It is promising and easy to implement the proposed method in other longitudinal data analysis examples. There are many interesting future directions. For example, we only use MR images for AD prediction in this paper. It is of interest to apply the proposed method to integrate multi-modal data including MRI, PET, and functional MRI. We expect the integration of multi-modal information would further improve the accuracy of the AD prediction.

Chapter 4 proposes a novel class of penalties for regression problems. The desirable theoretical properties of TWIN derive from its unique shape, which acts to inflate coefficient estimates in a certain range, thus alleviating issues in selection arising from shrinkage pseudo-noise. Probabilistic bounds for selection consistency were established under a challenging linear sparsity regime with random Gaussian designs. Minimax optimality was also established under the same data-generating regimes. Empirically, TWIN shows good performance even under scenarios with strong correlations in the design, suggesting that TWIN's theoretical properties may be extendible to more realistic data-generating scenarios. Motivated by this, we expect that exploration of TWIN's theoretical behavior under designs with significant correlation may be fruitful. In this work we provided asymptotically-motivated choices for the tuning parameters, however, the development of comprehensive strategies for simultaneous selection of τ and λ based on finite sample analysis is another interesting avenue of future research.

Chapter 5 investigates the relationship between the sharpness of the minima that SGD converges to with the ratio of the step size and the batch size. Using the SDE as the approximation of SGD, we explain part of the hypothesis proposed by Keskar et al. (2016) that large-batch methods tend to converge to sharp minimizers of the training function using the escaping time theorem in the finite-time regime. We prove that for the isotropic case the probability density function of SGD will converge to the stationary solution for any initial data regardless of the time varying step size and batch size. We give the convergence rate, which indicates that with a larger ratio of the learning rate and the batch size, the probability will converge faster to the stationary solution. Asymptotically the probability of converging to the global minimum is independent of the batch size and learning rate, but it only depends on the sharpness of the minimum. We verify these theoretical findings with numerical experiments. There are many directions for further study such as how the ratio of the step size and batch size influence the generalization error. In our experiment, it indicates that with a larger learning rate to batch size ratio the generalization error is worse. Further theoretical analysis is desired. Another interesting topic is to study the stationary solution and the evolution the probability density function of SGD when the variance matrix is anisotropic, which remain open questions.

Chapter 6 provides a new look at the model calibration problem in computer models. This viewpoint simultaneously considers two facts regarding how computer models are used in practice: computer models are inadequate for physical systems, regardless how the calibration parameters are tuned; and only a finite number of data points are available from the physical experiment to calibrate a related computer model. We establish a non-asymptotic minimax theory and derive an optimal prediction-oriented calibration method. Through several examples, the proposed calibration method has some advantages in prediction when compared with some existing calibration methods. We have developed an algorithm to carry out the proposed calibration method and built a link between our method and the Bayesian calibration method. Beyond the calibration of computer models, our method can be applied to calibrate unknown parameters for general misspecified models

in statistics and engineering. In many applications, bounded linear functional information such as derivative data are observed or can be easily calculated together with the function observations. It would be of interest to include all these data in our proposed calibration method. Furthermore, it is likely to extend our results to non-i.i.d. distributed designs; for example, general triangular arrays of non-random designs. This paper does not address these important questions, and we leave them open for future research.

Appendix A

Appendix For: Minimax Optimal Rates of Estimation in Functional ANOVA Models with Derivatives

This section consists of six parts. In Section A.1, we give a brief review on RKHS for the SS-ANOVA model and on the Fréchet derivative. In Section A.2, we give the proofs for results with deterministic designs of Section 2.3. In Section A.3, we show the proofs for results with random designs of Section 2.4. In Section A.4, we prove the results of estimating partial derivatives of Section 2.5. In Section A.5, we present key lemmas used in the proofs. All auxiliary technical lemmas are deferred to Section A.6.

A.1 Review of RKHS and Fréchet Derivative

A.1.1 RKHS for the SS-ANOVA Model

The SS-ANOVA model (2.2) truncates the sequence up to r interactions. Without loss of generality, we still denote the corresponding function space in (2.3) by $\mathcal{H}$, which is the direct sum of some set of the orthogonal subspaces in the decomposition $\otimes_{j=1}^{d}\mathcal{H}_1$. Denote by

$\|\cdot\|_{\otimes_{j=1}^{d}\mathcal{H}_1}$ the norm on $\otimes_{j=1}^{d}\mathcal{H}_1$, which is induced by the component norms $\|\cdot\|_{\mathcal{H}_1}$. Define $\|\cdot\|_{\mathcal{H}}$ as the norm on $\mathcal{H}$ by restricting $\|\cdot\|_{\otimes_{j=1}^{d}\mathcal{H}_1}$ to $\mathcal{H}$. Then $\mathcal{H}$ is a RKHS equipped with $\|\cdot\|_{\mathcal{H}}$. The quadratic penalty $J(\cdot)$ in (2.10) is defined as a squared semi-norm on $\mathcal{H}$ induced by a univariate penalty in $\mathcal{H}_1$. For example, when $\mathcal{H}_1 = \mathcal{W}_2^m(\mathcal{X}_1)$, it is common to choose $J(\cdot)$ for penalizing only the smooth components of a function. In this case, an explicit form is given in Wahba (1990).

Now we introduce some additional notation. Define a family of the multi-index $\overrightarrow{\boldsymbol{\nu}}$ by

$$\begin{aligned}\mathbf{V} = \{\overrightarrow{\boldsymbol{\nu}} = (\nu_1, \ldots, \nu_d)^\top \in \mathbb{N}^d,\\ \text{where at most } r \geq 1 \text{ of } \nu_k\text{s are not equal to } 1\},\end{aligned} \tag{A.1}$$

which will be used later since f_0 in the SS-ANOVA model (2.2) is truncated up to r interactions.

A.1.2 Fréchet Derivative of Operator

Let X and Y be the normed linear spaces. The Fréchet derivative of an operator $F : X \mapsto Y$ is a bounded linear operator $DF(a) : X \mapsto Y$ with

$$\lim_{h\to 0, h\in X} \frac{\|F(a+h) - F(a) - DF(a)h\|_Y}{\|h\|_X} = 0.$$

For illustration, if $F(a+h)-F(a) = Lh+R(a,h)$ with a linear operator L and $\|R(a,h)\|_Y/\|h\|_X \to 0$ as $h \to 0$, then by the above definition, $L = DF(a)$ is the Fréchet derivative of $F(\cdot)$. The reader is referred to elementary functional analysis textbooks such as Cartan (1971) for a thorough investigation on Fréchet derivative.

Lemma A.1. *Denote the loss function in (2.10) by $l_n(f)$. With the norm $\|\cdot\|_R$ in (A.19), the first*

order Fréchet derivative of the functional $l_n(\cdot)$ for any $f, g \in \mathcal{H}$ is

$$Dl_n(f)g = \frac{2}{n(p+1)} \left[\frac{1}{\sigma_0^2} \sum_{i=1}^{n} \{f(\mathbf{t}_i^{(0)}) - y_i^{(0)}\} g(\mathbf{t}_i^{(0)}) + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \left\{ \frac{\partial f(\mathbf{t}_i^{(j)})}{\partial t_j} - y_i^{(j)} \right\} \frac{\partial g(\mathbf{t}_i^{(j)})}{\partial t_j} \right].$$

The second order Fréchet derivative of $l_n(\cdot)$ for any $f, g, h \in \mathcal{H}$ is

$$D^2 l_n(f)gh = \frac{2}{n(p+1)} \left[\frac{1}{\sigma_0^2} \sum_{i=1}^{n} g(\mathbf{t}_i^{(0)}) h(\mathbf{t}_i^{(0)}) + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \frac{\partial g(\mathbf{t}_i^{(j)})}{\partial t_j} \frac{\partial h(\mathbf{t}_i^{(j)})}{\partial t_j} \right].$$

Proof. By direct calculations, we have

$$l_n(f+g) - l_n(f) = \frac{2}{n(p+1)} \left[\frac{1}{\sigma_0^2} \sum_{i=1}^{n} \{f(\mathbf{t}_i^{(0)}) - y_i^{(0)}\} g(\mathbf{t}_i^{(0)}) + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \left\{ \frac{\partial f(\mathbf{t}_i^{(j)})}{\partial t_j} - y_i^{(j)} \right\} \frac{\partial g(\mathbf{t}_i^{(j)})}{\partial t_j} \right] + R_n(f, g),$$

where

$$\begin{aligned} R_n(f,g) &= \frac{1}{n(p+1)} \left[\frac{1}{\sigma_0^2} \sum_{i=1}^{n} g^2(\mathbf{t}_i^{(0)}) + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \left\{ \frac{\partial g(\mathbf{t}_i^{(j)})}{\partial t_j} \right\}^2 \right] \\ &= \|g\|_0^2 + O(n^{-1/2}), \end{aligned}$$

and the $\|\cdot\|_0$ norm is given in (A.20). Note that $|R_n(f,g)|/\|g\|_R \to 0$ as $\|g\|_R \to 0$ and $n^{1/2}\|g\|_R \to \infty$. This proves the first part of the lemma. For the second order Fréchet derivative, note that

$$\begin{aligned} &Dl_n(f+h)g - Dl_n(f)g \\ &\quad = \frac{2}{n(p+1)} \left[\frac{1}{\sigma_0^2} \sum_{i=1}^{n} g(\mathbf{t}_i^{(0)}) h(\mathbf{t}_i^{(0)}) + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \frac{\partial g(\mathbf{t}_i^{(j)})}{\partial t_j} \frac{\partial h(\mathbf{t}_i^{(j)})}{\partial t_j} \right], \end{aligned}$$

which is linear in h. By definition of Fréchet derivatives, we conclude the form of $D^2 l_n(f)gh$ in the lemma. □

We remark that following a similar derivation in the above proof, we can obtain the first and the second order Fréchet derivatives of the functional $l_\infty(\cdot)$ in (A.26) and (A.28), respectively.

A.2 Proofs for Section 2.3: Deterministic Designs

For brevity, we consider the regular lattice $l_1 = \cdots = l_d = l$ and $n = l^d$. Other regular lattices can be showed similarly. Write

$$\psi_1(t) = 1, \quad \psi_{2\nu}(t) = \sqrt{2}\cos 2\pi\nu t, \quad \psi_{2\nu+1}(t) = \sqrt{2}\sin 2\pi\nu t, \tag{A.2}$$

for $\nu \geq 1$. Since f_0 has periodic boundaries on $\mathcal{X}_1^d$, $\{\psi_\nu(t)\}_{\nu\geq 1}$ forms an orthonormal system in $L_2(\mathcal{X}_1)$ and an eigenfunction system for K. For a d-dimensional vector $\overrightarrow{\boldsymbol{\nu}} = (\nu_1, \ldots, \nu_d) \in \mathbb{N}^d$, write

$$\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t}) = \psi_{\nu_1}(t_1)\cdots\psi_{\nu_d}(t_d) \quad \text{and} \quad \lambda_{\overrightarrow{\boldsymbol{\nu}}} = \lambda_{\nu_1}\lambda_{\nu_2}\cdots\lambda_{\nu_d}, \tag{A.3}$$

where λ_{ν_k}s and $\psi_{\nu_k}(t_k)$s are defined according to the Mercer's theorem,$k = 1, \ldots, d$. Then, any function $f(\cdot)$ in $\mathcal{H}$ admits the Fourier expansion $f(\mathbf{t}) = \sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbb{N}^d}\theta_{\overrightarrow{\boldsymbol{\nu}}}\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})$, where $\theta_{\overrightarrow{\boldsymbol{\nu}}} = \langle f(\mathbf{t}), \psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})\rangle_{L_2}$, and $J(f) = \sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbb{N}^d}\lambda_{\overrightarrow{\boldsymbol{\nu}}}^{-1}\theta_{\overrightarrow{\boldsymbol{\nu}}}^2$. We also write $f_0(\mathbf{t}) = \sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbb{N}^d}\theta_{\overrightarrow{\boldsymbol{\nu}}}^0\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})$.

By Page 23 of Wahba (1990), it is known that

$$l^{-1}\sum_{i=1}^{l}\psi_\mu(i/l)\psi_\nu(i/l) = \begin{cases} 1, & \text{if } \mu = \nu = 1, \ldots, l, \\ 0, & \text{if } \mu \neq \nu, \mu, \nu = 1, \ldots, l. \end{cases}$$

Define

$$\overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\nu}}} = (\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t}_1), \ldots, \psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t}_n))^\top,$$

where $\{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$ are the regular lattice design points. Thus, we have

$$\langle \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\nu}}}, \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\mu}}} \rangle_n = \begin{cases} 1, & \text{if } \nu_k = \mu_k = 1, \ldots, l; k = 1, \ldots, d, \\ 0, & \text{if there exists some } k \text{ such that } \nu_k \neq \mu_k, \end{cases}$$

where $\langle \cdot, \cdot \rangle_n$ is the empirical inner product in $\mathbb{R}^n$. This implies that $\{ \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\nu}}} \mid \nu_k = 1, \ldots, l; k = 1, \ldots, d\}$ form an orthogonal basis in $\mathbb{R}^n$ with respect to the empirical norm $\| \cdot \|_n$. Denote the observed data vectors by $\mathbf{y}^{(0)} = (y_1^{(0)}, \ldots, y_n^{(0)})^\top$ and $\mathbf{y}^{(j)} = (y_1^{(j)}, \ldots, y_n^{(j)})^\top$, and write

$$\begin{cases} z^{(0)}_{\overrightarrow{\boldsymbol{\nu}}} & = \langle \mathbf{y}^{(0)}, \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\nu}}} \rangle_n, \\ z^{(j)}_{\nu_1, \ldots, 2\nu_k - 1, \ldots, \nu_d} & = (2\pi)^{-1} \langle \mathbf{y}^{(j)}, \overrightarrow{\psi}_{\nu_1, \ldots, 2\nu_k, \ldots, \nu_d} \rangle_n, \\ z^{(j)}_{\nu_1, \ldots, 2\nu_k, \ldots, \nu_d} & = -(2\pi)^{-1} \langle \mathbf{y}^{(j)}, \overrightarrow{\psi}_{\nu_1, \ldots, 2\nu_k - 1, \ldots, \nu_d} \rangle_n, \end{cases} \tag{A.4}$$

for $\nu_k = 1, \ldots, l$ and $k = 1, \ldots, d$. Then, $z^{(0)}_{\overrightarrow{\boldsymbol{\nu}}} = \tilde{\theta}^0_{\overrightarrow{\boldsymbol{\nu}}} + \delta^{(0)}_{\overrightarrow{\boldsymbol{\nu}}}$ and $z^{(j)}_{\overrightarrow{\boldsymbol{\nu}}} = \nu_j \tilde{\theta}^0_{\overrightarrow{\boldsymbol{\nu}}} + \delta^{(j)}_{\overrightarrow{\boldsymbol{\nu}}}$, where $\tilde{\theta}^0_{\overrightarrow{\boldsymbol{\nu}}} = \theta^0_{\overrightarrow{\boldsymbol{\nu}}} + \sum_{\mu_k \geq l+1, k=1,\ldots,d} \theta^0_{\overrightarrow{\boldsymbol{\mu}}} \langle \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\nu}}}, \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\mu}}} \rangle_n$. The errors $\delta^{(0)}_{\overrightarrow{\boldsymbol{\nu}}}$ satisfy

$$\begin{aligned} \mathbb{E}[\delta^{(0)}_{\overrightarrow{\boldsymbol{\nu}}}] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\epsilon_i^{(0)}] \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\nu}}}(i) \leq \frac{1}{n} \sqrt{\sum_{i=1}^n \{\mathbb{E}[\epsilon_i^{(0)}]\}^2} \sqrt{\sum_{i=1}^n \overrightarrow{\psi}^2_{\overrightarrow{\boldsymbol{\nu}}}(i)} = o(n^{-1/2}), \\ \mathrm{Var}[\delta^{(0)}_{\overrightarrow{\boldsymbol{\nu}}}] &= \frac{1}{n^2} \sum_{i=1}^n \mathrm{Var}[\epsilon_i^{(0)}] \overrightarrow{\psi}^2_{\overrightarrow{\boldsymbol{\nu}}}(i) + \frac{1}{n^2} \sum_{i \neq i'} \mathrm{Cov}[\epsilon_i^{(0)}, \epsilon_{i'}^{(0)}] \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\nu}}}(i) \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\nu}}}(i') \\ &\leq \frac{\sigma_0^2}{n} \cdot \frac{1}{n} \sum_{i=1}^n \overrightarrow{\psi}^2_{\overrightarrow{\boldsymbol{\nu}}}(i) + \frac{2}{n^2} \sum_{i \neq i'} \mathrm{Cov}[\epsilon_i^{(0)}, \epsilon_{i'}^{(0)}] \\ &= O(n^{-1}) + \frac{2}{n^2} \sum_{i \neq i'} o(|i - i'|^{-\Upsilon}) = O(n^{-1}) + o(n^{-1}) = O(n^{-1}). \end{aligned}$$

Similarly for any j, $\delta^{(j)}_{\overrightarrow{\boldsymbol{\nu}}}$s have mean $o(n^{-1/2})$ and covariances $O(n^{-1})$.

A.2.1 Proof of Minimax Lower Bound: Theorem 2.1

We now prove the lower bound for estimating functions under the regular lattice. By the data transformation (A.4), it suffices to show the optimal rate in a special case

$$\begin{cases} z_{\overrightarrow{\boldsymbol{\nu}}}^{(0)} & = \theta_{\overrightarrow{\boldsymbol{\nu}}}^{0} + \delta_{\overrightarrow{\boldsymbol{\nu}}}^{(0)}, \\ z_{\overrightarrow{\boldsymbol{\nu}}}^{(j)} & = \nu_j \theta_{\overrightarrow{\boldsymbol{\nu}}}^{0} + \delta_{\overrightarrow{\boldsymbol{\nu}}}^{(j)}, \quad \text{for } 1 \le j \le p, \end{cases} \tag{A.5}$$

where $\delta_{\overrightarrow{\boldsymbol{\nu}}}^{(j)} \sim \mathcal{N}(0, \sigma_j^2/n)$ are independent. For any $\overrightarrow{\boldsymbol{\nu}} \in \mathbb{N}^d$, if we have the prior that $|\tilde{\theta}_{\overrightarrow{\boldsymbol{\nu}}}^{0}| \le \pi_{\overrightarrow{\boldsymbol{\nu}}}$, then the minimax linear estimator is

$$\widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}}^{L} = \frac{\sigma_0^{-2} z_{\overrightarrow{\boldsymbol{\nu}}}^{(0)} + \sum_{j=1}^{p} \sigma_j^{-2} \nu_j z_{\overrightarrow{\boldsymbol{\nu}}}^{(j)}}{n^{-1}\pi_{\overrightarrow{\boldsymbol{\nu}}}^{-2} + \sigma_0^{-2} + \sum_{j=1}^{p} \sigma_j^{-2} \nu_j^2},$$

and the minimax linear risk is

$$n^{-1} \left[n^{-1}\pi_{\overrightarrow{\boldsymbol{\nu}}}^{-2} + \sigma_0^{-2} + \sum_{j=1}^{p} \sigma_j^{-2} \nu_j^2 \right]^{-1}.$$

By Lemma 6 and Theorem 7 in Donoho et al. (1990), if σ_j^2s are known, the minimax risk of estimating $\theta_{\overrightarrow{\boldsymbol{\nu}}}^{0}$ under the model (A.5) is larger than 80% of the minimax linear risk of the hardest rectangle subproblem, and the latter linear risk is

$$R^L = n^{-1} \max_{\sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} (1+\lambda_{\overrightarrow{\boldsymbol{\nu}}}) \pi_{\overrightarrow{\boldsymbol{\nu}}}^2 = 1} \sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} \left[n^{-1}\pi_{\overrightarrow{\boldsymbol{\nu}}}^{-2} + \sigma_0^{-2} + \sum_{j=1}^{p} \sigma_j^{-2} \nu_j^2 \right]^{-1}, \tag{A.6}$$

where $\lambda_{\overrightarrow{\boldsymbol{\nu}}}$ is the product of eigenvalues in (A.3) and recall that the set V is defined in (A.1).

We use the Lagrange multiplier method to find $\pi_{\overrightarrow{\boldsymbol{\nu}}}^2$ for solving (A.6). Let a be the scalar multiplier and define

$$L(\pi_{\overrightarrow{\boldsymbol{\nu}}}^2, a) = \sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} \left[n^{-1}\pi_{\overrightarrow{\boldsymbol{\nu}}}^{-2} + \sigma_0^{-2} + \sum_{j=1}^{p} \sigma_j^{-2} \nu_j^2 \right]^{-1} - a(1+\lambda_{\overrightarrow{\boldsymbol{\nu}}})\pi_{\overrightarrow{\boldsymbol{\nu}}}^2.$$

Taking partial derivative with respect to $\pi^2_{\overrightarrow{\boldsymbol{\nu}}}$ gives

$$\frac{\partial L}{\partial \pi^2_{\overrightarrow{\boldsymbol{\nu}}}} = n^{-1}\left[n^{-1} + \left(\sigma_0^{-2} + \sum_{j=1}^{p}\sigma_j^{-2}\nu_j^2\right)\pi^2_{\overrightarrow{\boldsymbol{\nu}}}\right]^{-2} - a(1+\lambda_{\overrightarrow{\boldsymbol{\nu}}}) = 0.$$

This implies

$$\widehat{\pi}^2_{\overrightarrow{\boldsymbol{\nu}}} = \left(\sigma_0^{-2} + \sum_{j=1}^{p}\sigma_j^{-2}\nu_j^2\right)^{-1}\left[b(1+\lambda_{\overrightarrow{\boldsymbol{\nu}}})^{-1/2} - n^{-1}\right]_+,$$

where $b = (na)^{-1/2}$. On one hand, plugging the above formula into the constraint $\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V}}(1+\lambda_{\overrightarrow{\boldsymbol{\nu}}})\pi^2_{\overrightarrow{\boldsymbol{\nu}}} = 1$ gives

$$\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V}}\prod_{k=1}^{d}\nu_k^{2m}\left(\sigma_0^{-2} + \sum_{j=1}^{p}\sigma_j^{-2}\nu_j^2\right)^{-1}\left[b\prod_{k=1}^{d}\nu_k^{-m} - n^{-1}\right]_+ \asymp 1.$$

By restricting $\prod_{k=1}^{d}\nu_k \le (nb)^{1/m}$, this becomes

$$\begin{aligned}\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V},\prod_{k=1}^{d}\nu_k\le(nb)^{1/m}} &\left(\sigma_0^{-2} + \sum_{j=1}^{p}\sigma_j^{-2}\nu_j^2\right)^{-1} \\ &\times\left(b\prod_{k=1}^{d}\nu_k^{m} - n^{-1}\prod_{k=1}^{d}\nu_k^{2m}\right) \asymp 1.\end{aligned} \tag{A.7}$$

On the other hand, the linear risk in (A.6) can be written as

$$\begin{aligned}R^L \asymp n^{-1}\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V},\prod_{k=1}^{d}\nu_k\le(nb)^{1/m}} &\left(1 - \frac{1}{nb}\prod_{k=1}^{d}\nu_k^{m}\right) \\ &\times\left(\sigma_0^{-2} + \sum_{j=1}^{p}\sigma_j^{-2}\nu_j^2\right)^{-1}.\end{aligned} \tag{A.8}$$

We discuss for R^L in the above (A.8) under the condition (A.7) for three cases with $0 \le p \le d-r$, $d-r < p < d$ and $p = d$.

If $0 \le p \le d-r$, since $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$, there are at most r of $\nu_1, \ldots, \nu_d$ not equal to 1, which

implies that the number of combinations of non-1 indices being summed in (A.7) is no greater than $C_d^1 + C_d^2 + \cdots + C_d^r < \infty$. Due to the term $(\sigma_0^{-2} + \sum_{j=1}^p \sigma_j^{-2}\nu_j^2)^{-1}$, the largest terms of the summation (A.7) over $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$ correspond to the combinations of indices where as fewer $\nu_1, \ldots, \nu_p$ being summed as possible, for example, $v_k \equiv 1$ for $k \leq p$ and $k > p + r$, and $(\nu_{p+1}, \ldots, \nu_{p+r}) \in \mathbb{N}^r$ are non-1. Thus, (A.7) is equivalent to

$$\sum_{\prod_{k=1}^r \nu_{p+k} \leq (nb)^{1/m}} \left(b \prod_{k=1}^r \nu_{p+k}^m - n^{-1} \prod_{k=1}^r \nu_{p+k}^{2m} \right) \asymp 1.$$

Using the integral approximation, we have

$$\int_{\prod_{k=1}^r x_{p+k} \leq (nb)^{1/m}, x_{p+k} \geq 1} \left(b \prod_{k=1}^r x_{p+k}^m - \frac{1}{n} \prod_{k=1}^r x_{p+k}^{2m} \right) dx_{p+1} \cdots dx_{p+r} \asymp 1.$$

By letting $z_j = \prod_{1 \leq k \leq j} x_{p+k}$, $j = 1, 2, \ldots, r$, we have

$$\int_1^{(nb)^{1/m}} \left[\int_1^{z_r} \cdots \int_1^{z_2} \left(b z_r^m - \frac{1}{n} z_r^{2m} \right) z_1^{-1} \cdots z_{r-1}^{-1} dz_1 \cdots dz_{r-1} \right] dz_r \asymp 1,$$

where the left-hand side term is the order of $n^{(m+1)/m} b^{(2m+1)/m} [\log(nb)]^{r-1}$ and hence

$$b \asymp n^{-(m+1)/(2m+1)} (\log n)^{-m(r-1)/(2m+1)}. \tag{A.9}$$

The linear risk in (A.8) becomes

$$\begin{aligned} R^L &\asymp n^{-1} \int_{\prod_{k=1}^r x_{p+k} \leq (nb)^{1/m}, x_{p+k} \geq 1} \left(1 - \frac{1}{nb} \prod_{k=1}^r x_{p+k}^m \right) \\ &\asymp [\log(nb)]^{r-1} n^{-1+1/m} b^{1/m} \asymp [n (\log n)^{1-r}]^{-2m/(2m+1)}, \end{aligned}$$

where the last step is by (A.9).

If $d - r < p < d$, as discussed in the previous case, the number of combinations of non-1 indices being summed is finite, and the largest terms of the summation (A.7) over $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$ correspond to the combinations of indices where as fewer than $\nu_1, \ldots, \nu_p$ being summed as

possible, for example, $v_k \equiv 1$ for $k \leq d-r$, and $(\nu_{d-r+1}, \ldots, \nu_d) \in \mathbb{N}^r$ are non-1. Thus, (A.7) is equivalent to

$$\sum_{\prod_{k=1}^r \nu_{d-r+k} \leq (nb)^{1/m}} \left(b \prod_{k=1}^r \nu_{d-r+k}^m - n^{-1} \prod_{k=1}^r \nu_{d-r+k}^{2m} \right) \times \left(1 + \sum_{j=d-r+1}^p \nu_j^2 \right)^{-1} \asymp 1.$$

Using the integral approximation, we have

$$\int_{\prod_{k=1}^r x_{d-r+k} \leq (nb)^{1/m}, x_{d-r+k} \geq 1} \left(b \prod_{k=1}^r x_{d-r+k}^m - n^{-1} \prod_{k=1}^r x_{d-r+k}^{2m} \right) \times \left(1 + \sum_{j=d-r+1}^p x_j^2 \right)^{-1} dx_{d-r+1} \cdots dx_d \asymp 1.$$

By letting $z_j = x_{p+1} x_{p+2} \cdots x_j$, $j = p+1, \ldots, d$, we get

$$\begin{aligned} 1 &\asymp \int_{x_{d-r+1} \cdots x_p z_d \leq (nb)^{1/m}} \left[\int_1^{z_d} \cdots \int_1^{z_{p+2}} \left(b x_{d-r+1}^m \cdots x_p^m z_d^m - \frac{1}{n} x_{d-r+1}^{2m} \cdots x_p^{2m} z_d^{2m} \right) z_{p+1}^{-1} \cdots z_{d-1}^{-1} \right. \\ &\quad \left. \times \left(1 + x_{d-r+1}^2 + \cdots + x_p^2 \right)^{-1} dz_{p+1} \cdots dz_{d-1} \right] dx_{d-r+1} \cdots dx_p dz_d \\ &= \int_{x_{d-r+1} \cdots x_p z_d \leq (nb)^{1/m}} b x_{d-r+1}^m \cdots x_p^m z_d^m \left(1 - \frac{1}{nb} x_{d-r+1}^m \cdots x_p^m z_d^m \right) \\ &\quad \times (\log z_d)^{d-p-1} \left(1 + x_{d-r+1}^2 + \cdots + x_p^2 \right)^{-1} dx_{d-r+1} \cdots dx_p dz_d \\ &\asymp [\log(nb)]^{d-p-1} n^{1+1/m} b^{2+1/m}, \end{aligned}$$

where the last step is by Lemma A.18 in Section A.6. Hence,

$$b \asymp n^{-(m+1)/(2m+1)} (\log n)^{-m(d-p-1)/(2m+1)}. \tag{A.10}$$

The linear risk in (A.8) becomes

$$\begin{aligned} R^L &\asymp n^{-1} \int_{\prod_{k=d-r+1}^{d} x_k \le (nb)^{1/m}, x_k \ge 1} \left(1 - \frac{1}{nb} x_{d-r+1}^m \cdots x_d^m\right) \\ &\qquad \cdot (1 + x_{d-r+1}^2 + \cdots + x_p^2)^{-1} dx_{d-r+1} \cdots dx_d \\ &\asymp n^{-1} \int_{x_{d-r+1} \cdots x_p z_d \le (nb)^{1/m}} \left(1 - \frac{1}{nb} x_{d-r+1}^m \cdots x_p^m z_d^m\right) (\log z_d)^{d-p-1} \\ &\qquad \cdot (1 + x_{d-r+1}^2 + \cdots + x_p^2)^{-1} dx_{d-r+1} \cdots dx_p dz_d \\ &\asymp [\log(nb)]^{d-p-1} n^{-1+1/m} b^{1/m}, \end{aligned}$$

where the second step uses the same change of variables by letting $z_j = x_{p+1} x_{p+2} \cdots x_j$, $j = p+1, \ldots, d$, and the last step is by Lemma A.18 in Section A.6. By (A.10), we have

$$R^L \asymp [n(\log n)^{1+p-d}]^{-2m/(2m+1)}.$$

If $p = d$, as discussed in the previous two cases, the number of combinations of non-1 indices being summed is finite, and the largest terms of the summation (A.7) over $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$ correspond to any combinations of r non-1 indices, for example, $\nu_k \equiv 1$ for $k \ge r+1$, and $(\nu_1, \ldots, \nu_r) \in \mathbb{N}^r$. Thus, (A.7) is equivalent to

$$\sum_{\prod_{k=1}^{r} \nu_k \le (nb)^{1/m}} \left(b \prod_{k=1}^{r} \nu_k^m - n^{-1} \prod_{k=1}^{r} \nu_k^{2m}\right) \left(1 + \sum_{k=1}^{r} \nu_k^2\right)^{-1} \asymp 1.$$

Using the integral approximation, we have

$$\begin{aligned} 1 &\asymp \int_{\prod_{k=1}^{r} x_k \le (nb)^{1/m}, x_k \ge 1} \left(b \prod_{k=1}^{r} x_k^m - n^{-1} \prod_{k=1}^{r} x_k^{2m}\right) \left(1 + \sum_{k=1}^{r} x_k^2\right)^{-1} dx_1 \cdots dx_r \\ &\asymp \int_{\prod_{k=1}^{r} x_k \le (nb)^{1/m}, x_k \ge 1} b \prod_{k=1}^{r} x_k^m \left(1 + \sum_{k=1}^{r} x_k^2\right)^{-1} dx_1 \cdots dx_r \end{aligned}$$

By letting $\beta = m > 1$ and $\alpha = 2$ in Lemma A.19 in Section A.6, we have for any $r \geq 1$,

$$b \asymp n^{-(mr+r-2)/(2mr+r-2)}. \tag{A.11}$$

The linear risk in (A.8) becomes

$$\begin{aligned} R^L &\asymp n^{-1} \int_{\prod_{k=1}^r x_k \leq (nb)^{1/m}, x_k \geq 1} \left(1 - \frac{1}{nb} x_1^m \cdots x_r^m\right) \\ &\qquad \cdot (1 + x_1^2 + \cdots + x_r^2)^{-1} dx_1 \cdots dx_r \\ &\asymp n^{-1} \int_{\prod_{k=1}^r x_k \leq (nb)^{1/m}, x_k \geq 1} (1 + x_1^2 + \cdots + x_r^2)^{-1} dx_1 \cdots dx_r \\ &\asymp \left[n^{-1}(nb)^{(r-2)/(mr)}\right] 1_{r\geq 3} + \left[n^{-1}\log(nb)\right] 1_{r=2} + \left(n^{-1}\right) 1_{r=1}, \end{aligned}$$

where the last step uses Lemma A.19 in Section A.6 by letting $\beta = 0$ and $\alpha = 2$. By (A.11), we have

$$R^L \asymp \left[n^{-(2mr)/[(2m+1)r-2]}\right] 1_{r\geq 3} + \left[n^{-1}\log(n)\right] 1_{r=2} + n^{-1} 1_{r=1},$$

where the constant factor does not depend on n. This completes the proof.

A.2.2 Proof of Minimax Upper Bound: Theorem 2.2

We now prove the theorem for only $r = d$ and $p = d - 1$. Other cases can be proved similarly with slight changes.

Using the discrete transformed data (A.4), the regularized estimator $\widehat{f}_{n\lambda}$ by (2.10) can be obtained through

$$\begin{aligned} \widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}} = \underset{\tilde{\theta}_{\overrightarrow{\boldsymbol{\nu}}} \in \mathbb{R}}{\arg\min} \Bigg\{ &\frac{1}{n(p+1)} \Bigg[\frac{1}{\sigma_0^2} \sum_{\overrightarrow{\boldsymbol{\nu}} \in V, \|\overrightarrow{\boldsymbol{\nu}}\|_{\min} \leq l} \left(z_{\overrightarrow{\boldsymbol{\nu}}}^{(0)} - \theta_{\overrightarrow{\boldsymbol{\nu}}}\right)^2 \\ &+ \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{\overrightarrow{\boldsymbol{\nu}} \in V, \|\overrightarrow{\boldsymbol{\nu}}\|_{\min} \leq l} \left(z_{\overrightarrow{\boldsymbol{\nu}}}^{(j)} - \nu_j \theta_{\overrightarrow{\boldsymbol{\nu}}}\right)^2 \Bigg] + \lambda \sum_{\overrightarrow{\boldsymbol{\nu}} \in V, \|\overrightarrow{\boldsymbol{\nu}}\|_{\min} \leq l} \lambda_{\overrightarrow{\boldsymbol{\nu}}} \theta_{\overrightarrow{\boldsymbol{\nu}}}^2 \Bigg\} \end{aligned}$$

and $\widehat{f}_{n\lambda}(\mathbf{t}) = \sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V},\|\overrightarrow{\boldsymbol{\nu}}\|_{\min}\le l} \widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}}\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})$, where $\mathbf{V}$ is defined in (A.1). Direct calculations give

$$\widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}} = \frac{\sigma_0^{-2} z_{\overrightarrow{\boldsymbol{\nu}}}^{(0)} + \sum_{j=1}^p \sigma_j^{-2}\nu_j z_{\overrightarrow{\boldsymbol{\nu}}}^{(j)}}{\sigma_0^{-2} + \sum_{j=1}^p \sigma_j^{-2}\nu_j^2 + \lambda\lambda_{\overrightarrow{\boldsymbol{\nu}}}^{-1}}.$$

The deterministic error of $\widehat{f}_{n\lambda}$ can be analyzed by two parts. On one hand, since $f_0 \in \mathcal{H}$ and $\lambda_\nu \asymp \nu^{-2m}$, we know $\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V},\|\overrightarrow{\boldsymbol{\nu}}\|_{\min}\ge l+1}(\theta_{\overrightarrow{\boldsymbol{\nu}}}^0)^2 \asymp n^{-2m}$. This is the truncation error due to $\widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}} = 0$ for $\nu_k \ge l+1$, $1 \le k \le d$. On the other hand, note that $\langle \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\nu}}}, \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\mu}}}\rangle_n^2 \le 1$ and then

$$\left(\sum_{\overrightarrow{\boldsymbol{\mu}}\in\mathbf{V},\|\overrightarrow{\boldsymbol{\mu}}\|_{\min}\ge l+1} \theta_{\overrightarrow{\boldsymbol{\mu}}}^0 \langle \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\nu}}}, \overrightarrow{\psi}_{\overrightarrow{\boldsymbol{\mu}}}\rangle_n\right)^2 \le \sum_{\overrightarrow{\boldsymbol{\mu}}\in\mathbf{V},\|\overrightarrow{\boldsymbol{\mu}}\|_{\min}\ge l+1} (\theta_{\overrightarrow{\boldsymbol{\mu}}}^0)^2 \asymp n^{-2m}.$$

Thus,

$$\begin{aligned}
&\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V},\|\overrightarrow{\boldsymbol{\nu}}\|_{\min}\le l} \left(\mathbb{E}\widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}} - \theta_{\overrightarrow{\boldsymbol{\nu}}}^0\right)^2 \\
&\lesssim \sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V},\|\overrightarrow{\boldsymbol{\nu}}\|_{\min}\le l} \frac{(\lambda\lambda_{\overrightarrow{\boldsymbol{\nu}}}^{-1}\theta_{\overrightarrow{\boldsymbol{\nu}}}^0)^2 + [\mathbb{E}\delta_{\overrightarrow{\boldsymbol{\nu}}}^{(0)}]^2 + \sum_{j=1}^p \nu_j^2[\mathbb{E}\delta_{\overrightarrow{\boldsymbol{\nu}}}^{(j)}]^2}{(\sigma_0^{-2} + \sum_{j=1}^p \sigma_j^{-2}\nu_j^2 + \lambda\lambda_{\overrightarrow{\boldsymbol{\nu}}}^{-1})^2} + n^{-2m+1} \\
&\le \lambda^2 \sup_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V}} \frac{\lambda_{\overrightarrow{\boldsymbol{\nu}}}^{-1}}{\left(\sigma_0^{-2} + \sum_{j=1}^p \sigma_j^{-2}\nu_j^2 + \lambda\lambda_{\overrightarrow{\boldsymbol{\nu}}}^{-1}\right)^2} \sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V}} \lambda_{\overrightarrow{\boldsymbol{\nu}}}^{-1}(\theta_{\overrightarrow{\boldsymbol{\nu}}}^0)^2 \\
&\quad + o(n^{-1}) \sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V},\|\overrightarrow{\boldsymbol{\nu}}\|_{\min}\le l} \frac{1 + \sum_{j=1}^p \nu_j^2}{(1 + \sum_{j=1}^p \nu_j^2 + \lambda\nu_1^{2m}\cdots\nu_d^{2m})^2} + n^{-2m+1} \\
&\asymp \lambda^2 J(f_0) \sup_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V}} \frac{\nu_1^{2m}\cdots\nu_d^{2m}}{(1 + \sum_{j=1}^p \nu_j^2 + \lambda\nu_1^{2m}\cdots\nu_d^{2m})^2} + o\{n^{-1}\lambda^{-1/2m}\} + n^{-2m+1},
\end{aligned}$$

where the last step uses Lemma A.12 in Section A.5.3 with $a = 0$ and $p = d-1$. Define that

$$B_\lambda(\overrightarrow{\boldsymbol{\nu}}) = \frac{\nu_1^{2m}\cdots\nu_d^{2m}}{(1 + \sum_{j=1}^p \nu_j^2 + \lambda\nu_1^{2m}\cdots\nu_d^{2m})^2}.$$

For the $\sup_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V}} B_\lambda(\overrightarrow{\boldsymbol{\nu}})$ term above, suppose that $\prod_{j=1}^d \nu_j^{2m} > 0$ is fixed and denoted by x^{-1}, then $B_\lambda(\overrightarrow{\boldsymbol{\nu}})$ is maximized by letting $\sum_{j=1}^p \nu_j^2$ be as small as possible, where $p = d-1$. This

suggests $\nu_1 = \nu_2 = \cdots = \nu_p = 1$, and

$$\sup_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} B_\lambda(\overrightarrow{\boldsymbol{\nu}}) \asymp \sup_{x>0} \frac{x^{-1}}{(1+\lambda x^{-1})^2} \asymp \lambda^{-1},$$

where the last step is achieved when $x \asymp \lambda$. Combining all parts of bias gives

$$\sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} \left(\mathbb{E}\widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}} - \theta^0_{\overrightarrow{\boldsymbol{\nu}}} \right)^2 = O\left\{ \lambda J(f_0) + n^{-2m+1} \right\} + o\{ n^{-1}\lambda^{-1/2m} \}, \tag{A.12}$$

where the constant factor on the upper bound does not depend on n.

The stochastic error is bounded as follows:

$$\begin{aligned} \sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} \mathbb{E}\left(\widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}} - \mathbb{E}\widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}} \right)^2 &= \sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}, \|\overrightarrow{\boldsymbol{\nu}}\|_{\min} \le l} \frac{n^{-1}(\sigma_0^{-2} + \sum_{j=1}^p \sigma_j^{-2}\nu_j^2)}{(\sigma_0^{-2} + \sum_{j=1}^p \sigma_j^{-2}\nu_j^2 + \lambda\lambda_{\overrightarrow{\boldsymbol{\nu}}}^{-1})^2} \\ &\lesssim \sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}, \|\overrightarrow{\boldsymbol{\nu}}\|_{\min} \le l} \frac{1 + \sum_{j=1}^p \nu_j^2}{n(1 + \sum_{j=1}^p \nu_j^2 + \lambda\nu_1^{2m}\cdots\nu_d^{2m})^2}. \end{aligned}$$

Using Lemma A.12 in Section A.5.3 with $a = 0$ and $p = d-1$, we have

$$\sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} \mathbb{E}\left(\widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}} - \mathbb{E}\widehat{\theta}_{\overrightarrow{\boldsymbol{\nu}}} \right)^2 = O\left\{ n^{-1}\lambda^{-1/2m} \right\}. \tag{A.13}$$

Combining (A.12) and (A.13) and letting $\lambda \asymp n^{-2m/(2m+1)}$ completes the proof.

A.3 Proofs of Results in Section 2.4: Random Designs

A.3.1 Proof of the Minimax Lower Bound: Theorem 2.3

We establish the lower bound for the random design via Fano's lemma. It suffices to consider a special case where noises $\epsilon^{(0)}$ and $\epsilon^{(j)}$s are independent Gaussian with zero mean and unit variance, and $\Pi^{(0)}$ and $\Pi^{(j)}$s are uniform, and $\mathcal{H}_1$ is generated by periodic kernels.

Let N be a natural number whose value will be clear later. We first derive the eigenvalue decay rate for the kernel K_d which generates the RKHS $\mathcal{H}$. For a given $\tau > 0$, the number of

multi-indices $\overrightarrow{\boldsymbol{\nu}} = (\nu_1, \ldots, \nu_r) \in \mathbb{N}^r$ satisfying $\nu_1^{-2m} \cdots \nu_r^{-2m} \geq \tau$ is the same as the number of multi-indices such that $\nu_1 \cdots \nu_r \leq \tau^{-1/(2m)}$, which amounts to

$$\begin{aligned} \sum_{\nu_2 \cdots \nu_r \leq \tau^{-1/(2m)}} \tau^{-1/(2m)} / (\nu_2 \cdots \nu_r) &= \tau^{-1/(2m)} \left(\sum_{\nu \leq \tau^{-1/(2m)}} 1/\nu \right)^{r-1} \\ &\asymp \tau^{-1/(2m)} (\log 1/\tau)^{r-1}. \end{aligned} \tag{A.14}$$

Denote by $\lambda_N(K_d)$ the Nth eigenvalues of K_d. By inverting (A.14), obtain

$$\lambda_N(K_d) \asymp \left[N (\log N)^{1-r} \right]^{-2m}.$$

Hence, the multi-indices $\overrightarrow{\boldsymbol{\nu}} = (\nu_1, \ldots, \nu_r) \in \mathbb{N}^r$ satisfying $\nu_1 \cdots \nu_r \leq N$ correspond to the first

$$c_0 N (\log N)^{r-1}$$

eigenvalues of K_d for some constant c_0. Let $b = \{b_{\overrightarrow{\boldsymbol{\nu}}} : \nu_1 \cdots \nu_r \leq N\} \in \{0,1\}^{c_0 N (\log N)^{r-1}}$ be a length-$\{c_0 N (\log N)^{r-1}\}$ binary sequence, and $\{\tilde{\lambda}_{\overrightarrow{\boldsymbol{\nu}}} : \nu_1 \cdots \nu_r \leq N\}$ be the first $c_0 N (\log N)^{r-1}$ eigenvalues of K_d. Denote by $\{\tilde{\lambda}_{\overrightarrow{\boldsymbol{\nu}} + c_0 N (\log N)^{r-1}} : \nu_1 \cdots \nu_r \leq N\}$ the $\{c_0 N (\log N)^{r-1} + 1\}$th, $\{c_0 N (\log N)^{r-1} + 2\}$th,…, $\{2 c_0 N (\log N)^{r-1}\}$th eigenvalues of K_d.

For brevity, we only prove for the case $p = d$ and $r \geq 3$. The other cases $p = d$, $r \leq 2$ and $0 \leq p < d$ can be showed similarly. We deal with the differences among these cases for deterministic designs in Section A.2.1. Write

$$\begin{aligned} f_b(t_1, \ldots, t_r) = N^{-1/2+1/r} \sum_{\nu_1 \cdots \nu_r \leq N} & b_{\overrightarrow{\boldsymbol{\nu}}} \left(1 + \nu_1^2 + \cdots + \nu_r^2\right)^{-1/2} \\ & \times \tilde{\lambda}^{1/2}_{\overrightarrow{\boldsymbol{\nu}} + c_0 N (\log N)^{r-1}} \psi_{\overrightarrow{\boldsymbol{\nu}} + c_0 N (\log N)^{r-1}}(t_1, \ldots, t_r), \end{aligned}$$

where $\psi_{\overrightarrow{\boldsymbol{\nu}} + c_0 N (\log N)^{r-1}}(t_1, \ldots, t_r)$ are the corresponding eigenfunctions of $\tilde{\lambda}_{\overrightarrow{\boldsymbol{\nu}} + c_0 N (\log N)^{r-1}}$

of K_d. Note that

$$\begin{aligned}\|f_b\|_{\mathcal{H}}^2 &= N^{-1+2/r} \sum_{\nu_1 \cdots \nu_r \le N} b_{\overrightarrow{\boldsymbol{\nu}}}^2 (1+\nu_1^2+\cdots+\nu_r^2)^{-1} \\ &\le N^{-1+2/r} \sum_{\nu_1 \cdots \nu_r \le N} (1+\nu_1^2+\cdots+\nu_r^2)^{-1} \asymp 1,\end{aligned}$$

where the last step by Lemma A.19, and this implies $f_b(\cdot) \in \mathcal{H}$.

By the Varshamov-Gilbert bound, e.g., Tsybakov (2009), there exists a collection of binary sequences $\{b^{(1)}, \ldots, b^{(M)}\} \subset \{0,1\}^{c_0 N (\log N)^{r-1}}$ such that

$$M \ge 2^{c_0 N (\log N)^{r-1}/8}$$

and

$$H(b^{(l)}, b^{(q)}) \ge c_0 N (\log N)^{r-1}/8, \quad \forall 1 \le l < q \le M,$$

where $H(\cdot,\cdot)$ is the Hamming distance. Then, for $b^{(l)}, b^{(q)} \in \{0,1\}^{c_0 N (\log N)^{r-1}}$,

$$\begin{aligned}&\|f_{b^{(l)}} - f_{b^{(q)}}\|_{L_2}^2 \\ &\quad \ge N^{-1+2/r} (2N)^{-2m} \sum_{\nu_1 \cdots \nu_r \le N} (1+\nu_1^2+\cdots+\nu_r^2)^{-1} \left[b_{\overrightarrow{\boldsymbol{\nu}}}^{(l)} - b_{\overrightarrow{\boldsymbol{\nu}}}^{(q)} \right]^2 \\ &\quad \ge N^{-1+2/r} (2N)^{-2m} \sum_{c_1 7N/8 \le \nu_1 \cdots \nu_r \le N} (1+\nu_1^2+\cdots+\nu_r^2)^{-1} \\ &\quad = c_2 N^{-2m}\end{aligned}$$

for some constants c_1 and c_2, where the last step is by Lemma A.19.

On the other hand, for any $b^{(l)} \in \{b^{(1)}, \ldots, b^{(M)}\}$, by Lemma A.19,

$$\begin{aligned}\|f_{b^{(l)}}\|_{L_2}^2 + \sum_{j=1}^p \|\partial f_{b^{(l)}}/\partial t_j\|_{L_2}^2 &\le N^{-1+2/r} \sum_{\nu_1 \cdots \nu_r \le N} \nu_1^{-2m} \cdots \nu_r^{-2m} \left[b_{\overrightarrow{\boldsymbol{\nu}}}^{(l)} \right]^2 \\ \le N^{-1+2/r} \sum_{\nu_1 \cdots \nu_r \le N} \nu_1^{-2m} \cdots \nu_r^{-2m} &= c_3 N^{-2m+2/r} (\log N)^{r-1}\end{aligned}$$

for some constant c_3.

A standard argument gives that the lower bound can be reduced to the error probability in a multi-way hypothesis test (Tsybakov, 2009). Specifically, let Θ be a random variable uniformly distributed on $\{1, \ldots, M\}$. Note that

$$\infty_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{P}\left\{ \|\tilde{f} - f_0\|_{L_2}^2 \geq \frac{1}{4} \min_{b^{(l)} \neq b^{(q)}} \|f_{b^{(l)}} - f_{b^{(q)}}\|_{L_2}^2 \right\} \geq \infty_{\widehat{\Theta}} \mathbb{P}\{\widehat{\Theta} \neq \Theta\}, \tag{A.15}$$

where the infimum on the right-hand side is taken over all decision rules that are measurable functions of the data. By Fano's lemma,

$$\begin{aligned} &\mathbb{P}\left\{ \widehat{\Theta} \neq \Theta | \mathbf{t}_1^{(0)}, \ldots, \mathbf{t}_n^{(0)}; \ldots; \mathbf{t}_1^{(p)}, \ldots, \mathbf{t}_n^{(p)} \right\} \geq 1 - \frac{1}{\log M} \times \\ &\left[1_{\mathbf{t}_1^{(0)}, \ldots, \mathbf{t}_n^{(0)}; \ldots; \mathbf{t}_1^{(p)}, \ldots, \mathbf{t}_n^{(p)}} (y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)}; \Theta) + \log 2 \right], \end{aligned} \tag{A.16}$$

where $1_{\mathbf{t}_1^{(0)}, \ldots, \mathbf{t}_n^{(0)}; \ldots; \mathbf{t}_1^{e_p}, \ldots, \mathbf{t}_n^{e_p}} (y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)})$ is the mutual information between Θ and $\{y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)}\}$ with the design points $\{\mathbf{t}_1^{(0)}, \ldots, \mathbf{t}_n^{(0)}; \ldots; \mathbf{t}_1^{(p)}, \ldots, \mathbf{t}_n^{(p)}\}$ being fixed. We can derive that

$$\begin{aligned} &\mathbb{E}_{\mathbf{t}_1^{(0)}, \ldots, \mathbf{t}_n^{(0)}; \ldots; \mathbf{t}_1^{(p)}, \ldots, \mathbf{t}_n^{(p)}} \\ &\quad \cdot \left[1_{\mathbf{t}_1^{(0)}, \ldots, \mathbf{t}_n^{(0)}; \ldots; \mathbf{t}_1^{(p)}, \ldots, \mathbf{t}_n^{(p)}} \left(y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)}; \Theta \right) \right] \\ &\leq \binom{M}{2}^{-1} \sum_{b^{(l)} \neq b^{(q)}} \mathbb{E}_{\mathbf{t}_1^{(0)}, \ldots, \mathbf{t}_n^{(0)}; \ldots; \mathbf{t}_1^{(p)}, \ldots, \mathbf{t}_n^{(p)}} \mathcal{K}\left(\mathbf{P}_{f_{b^{(l)}}} | \mathbf{P}_{f_{b^{(q)}}} \right) \\ &\leq \frac{n(p+1)}{2} \binom{M}{2}^{-1} \sum_{b^{(l)} \neq b^{(q)}} \mathbb{E}_{\mathbf{t}_1^{(0)}, \ldots, \mathbf{t}_n^{(0)}; \ldots; \mathbf{t}_1^{(p)}, \ldots, \mathbf{t}_n^{(p)}} \|f_{b^{(l)}} - f_{b^{(q)}}\|_{*n}^2, \end{aligned} \tag{A.17}$$

where $\mathcal{K}(\cdot|\cdot)$ is the Kullback-Leibler distance, $\mathbf{P}_f$ is conditional distribution of $y_i^{(0)}$ and $y_i^{(j)}$s given $\{\mathbf{t}_1^{(0)}, \ldots, \mathbf{t}_n^{(0)}; \ldots; \mathbf{t}_1^{(p)}, \ldots, \mathbf{t}_n^{(p)}\}$, and the norm $\|\cdot\|_*$ is defined as

$$\|f\|_{*n}^2 = \frac{1}{n(p+1)} \sum_{i=1}^{n} \left\{ [f(\mathbf{t}_i^{(0)})]^2 + \sum_{j=1}^{p} [\partial f(\mathbf{t}_i^{(j)}) / \partial t_j]^2 \right\}, \quad \forall f : \mathcal{X}_1^r \mapsto \mathbb{R}.$$

Thus,

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{t}_1^{(0)},\ldots,\mathbf{t}_n^{(0)};\ldots;\mathbf{t}_1^{(p)},\ldots,\mathbf{t}_n^{(p)}} \\
&\qquad \cdot \left[1_{\mathbf{t}_1^{(0)},\ldots,\mathbf{t}_n^{(0)};\ldots;\mathbf{t}_1^{(p)},\ldots,\mathbf{t}_n^{(p)}}(y_1^{(0)},\ldots,y_n^{(0)},\ldots,y_1^{(p)},\ldots,y_n^{(p)};\Theta)\right] \\
&\leq \frac{n(p+1)}{2}\binom{M}{2}^{-1} \sum_{b^{(l)}\neq b^{(q)}} \left\{ \|f_{b^{(l)}} - f_{b^{(q)}}\|_{L_2}^2 \right. \\
&\qquad\qquad \left. + \sum_{j=1}^{p} \|\partial f_{b^{(l)}}/\partial t_j - \partial f_{b^{(q)}}/\partial t_j\|_{L_2}^2 \right\} \\
&\leq \frac{n(p+1)}{2} \max_{b^{(l)}\neq b^{(q)}} \left\{ \|f_{b^{(l)}} - f_{b^{(q)}}\|_{L_2}^2 \right. \\
&\qquad\qquad \left. + \sum_{j=1}^{p} \|\partial f_{b^{(l)}}/\partial t_j - \partial f_{b^{(q)}}/\partial t_j\|_{L_2}^2 \right\} \\
&\leq 2n(p+1) \max_{b^{(l)}\in\{b^{(1)},\ldots,b^{(M)}\}} \left\{ \|f_{b^{(l)}}\|_{L_2}^2 + \sum_{j=1}^{p} \|\partial f_{b^{(l)}}/\partial t_j\|_{L_2}^2 \right\} \\
&\leq 2c_3 n(p+1) N^{-2m+2/r} (\log N)^{r-1}.
\end{aligned}
\tag{A.18}
$$

Now, (E.15) yields

$$
\begin{aligned}
&\infty_{\tilde{f}} \sup_{f_0\in\mathcal{H}} \mathbb{P}\left\{ \|\tilde{f} - f_0\|_{L_2}^2 \geq \frac{1}{4} c_2 N^{-2m} \right\} \\
&\geq \infty_{\widehat{\Theta}} \mathbb{P}\{\widehat{\Theta} \neq \Theta\} \\
&\geq 1 - \frac{1}{\log M}\left[\mathbb{E} 1_{\mathbf{t}_1^{(0)},\ldots,\mathbf{t}_n^{(0)};\ldots;\mathbf{t}_1^{(p)},\ldots,\mathbf{t}_n^{(p)}}(y_1^{(0)},\ldots,y_n^{(0)},\ldots,y_1^{(p)},\ldots,y_n^{(p)};\Theta) + \log 2\right] \\
&\geq 1 - \frac{2c_3 n(p+1) N^{-2m+2/r}(\log N)^{r-1} + \log 2}{c_0 (\log 2) N (\log N)^{r-1}/8}.
\end{aligned}
$$

Taking $N = c_4 n^{r/(2mr+r-2)}$ with an appropriate choice of c_4, we have

$$
\limsup_{n\to\infty} \infty_{\tilde{f}} \sup_{f_0\in\mathcal{H}} \mathbb{P}\left\{ \|\tilde{f} - f_0\|_{L_2}^2 \geq C_3 n^{-2mr/(2mr+r-2)} \right\} > 0,
$$

where C_3 does not depend on n. This completes the proof.

A.3.2 Proof of the Minimax Upper Bound: Theorem 2.4

Preliminaries for the proof We define a new norm for any $f \in \mathcal{H}$,

$$\|f\|_R^2 = \frac{1}{p+1}\left[\frac{1}{\sigma_0^2}\int f^2(\mathbf{t})d\Pi^{(0)}(\mathbf{t}) + \sum_{j=1}^{p}\frac{1}{\sigma_j^2}\int\left\{\frac{\partial f(\mathbf{t})}{\partial t_j}\right\}^2 d\Pi^{(j)}(\mathbf{t})\right] + J(f). \tag{A.19}$$

Note that $\|\cdot\|_R$ is a norm since it is a quadratic form and is equal to zero if and only if $f = 0$. Let $\langle\cdot,\cdot\rangle_R$ be the inner product associated with $\|\cdot\|_R$. The following lemma shows that $\|\cdot\|_R$ is well defined in $\mathcal{H}$ and is equivalent to the RKHS norm $\|\cdot\|_{\mathcal{H}}$. In particular, $\|f\|_R < \infty$ if and only if $\|f\|_{\mathcal{H}} < \infty$. The proof of this lemma is given in Section A.5.1.

Lemma A.2. *The norm $\|\cdot\|_R$ is equivalent to $\|\cdot\|_{\mathcal{H}}$ in $\mathcal{H}$.*

We introduce another norm $\|\cdot\|_0$ given by

$$\|f\|_R^2 = \frac{1}{p+1}\left[\frac{1}{\sigma_0^2}\int f^2(\mathbf{t})d\Pi^{(0)}(\mathbf{t}) + \sum_{j=1}^{p}\frac{1}{\sigma_j^2}\int\left\{\frac{\partial f(\mathbf{t})}{\partial t_j}\right\}^2 d\Pi^{(j)}(\mathbf{t})\right]. \tag{A.20}$$

We define a function space F_0 to be the direct sum of some set of the orthogonal subspaces in the decomposition of $\otimes_{j=1}^{d} L_2(\mathcal{X}_1)$ as in (2.3) and equipped with the norm $\|\cdot\|_0$. Let $\langle\cdot,\cdot\rangle_0$ be the inner product associated with $\|\cdot\|_0$ in F_0.

For the above two norms, we introduce some additional notation. Denote the loss function in (2.10) by $l_n(f)$, that is,

$$l_n(f) = \frac{1}{n(p+1)}\left[\frac{1}{\sigma_0^2}\sum_{i=1}^{n}\{f(\mathbf{t}_i^{(0)}) - y_i^{(0)}\}^2 + \sum_{j=1}^{p}\frac{1}{\sigma_j^2}\sum_{i=1}^{n}\left\{\frac{\partial f(\mathbf{t}_i^{(j)})}{\partial t_j} - y_i^{(j)}\right\}^2\right],$$

and write $l_{n\lambda}(f) = l_n(f) + \lambda J(f)$. Then the regularized estimator $\widehat{f}_{n\lambda} = \arg\min_{f\in\mathcal{H}} l_{n\lambda}(f)$. Denote the expected loss by $l_\infty(f) = \mathbb{E}l_n(f)$ and write $l_{\infty\lambda}(f) = l_\infty(f) + \lambda J(f)$. Since $l_{\infty\lambda}(f)$

a positive quadratic form in $f \in \mathcal{H}$, it has a unique minimizer in $\mathcal{H}$ given by

$$\bar{f}_{\infty\lambda} = \arg\min_{f\in\mathcal{H}} l_{\infty\lambda}(f).$$

Thus, we decompose

$$\widehat{f}_{n\lambda} - f_0 = (\widehat{f}_{n\lambda} - \bar{f}_{\infty\lambda}) + (\bar{f}_{\infty\lambda} - f_0),$$

where $(\widehat{f}_{n\lambda} - \bar{f}_{\infty\lambda})$ is referred to the stochastic error and $(\bar{f}_{\infty\lambda} - f_0)$ is referred to the deterministic error. If data $Y^{(0)}$ and $Y^{(j)}$s in (2.1) are observed without random noises as in deterministic computer experiments, then the total error is only the deterministic error with $\widehat{f}_{n\lambda} - f_0 = \bar{f}_{\infty\lambda} - f_0$. For brevity, we omit the subscripts of $\bar{f}_{\infty\lambda}$ and $\widehat{f}_{n\lambda}$ hereafter if no confusion occurs.

Outline of the proof Before proceeding to the proof, we make two remarks on the setup of Theorem 2.4. First, since the distributions $\Pi^{(0)}$ and $\Pi^{(j)}$s are known, by the inverse transform sampling, it suffices to consider the uniform distribution. A detailed discussion on this transform is given in Lemma A.17. Second, it suffices to assume f_0 to have a periodic boundary on $\mathcal{X}_1^d$ in the proof of the theorem. This is because f_0 is a tensor product function and each component function space is supported in a compact domain. Thus, we can smoothly extend f_0 to a larger compact support domain and achieve periodicity on the new boundary, e.g., uniformly zero on the new boundary. These two simplifications make the proof easier to present.

Recall that the trigonometrical basis on $L_2(\mathcal{X}_1)$ is $\psi_1(t) = 1$, $\psi_{2\nu}(t) = \sqrt{2}\cos 2\pi\nu t$ and $\psi_{2\nu+1}(t) = \sqrt{2}\sin 2\pi\nu t$ for $\nu \geq 1$. Write

$$\phi_{\vec{\boldsymbol{\nu}}}(t_1, \ldots, t_d) = \frac{\psi_{\nu_1}(t_1)\cdots\psi_{\nu_d}(t_d)}{\|\psi_{\nu_1}(t_1)\cdots\psi_{\nu_d}(t_d)\|_0}. \tag{A.21}$$

Since f_0 has a periodic boundary on $\mathcal{X}_1^d$ and $\pi^{(j)} \equiv 1$, $\{\phi_{\vec{\boldsymbol{\nu}}}(\mathbf{t}) : \vec{\boldsymbol{\nu}} \in \mathbf{V}\}$, where $\mathbf{V}$ in (A.1) forms an orthogonal basis for $\mathcal{H}$ in $\langle\cdot,\cdot\rangle_R$; an orthogonal system for $L_2(\mathcal{X}_1^d)$; and an orthonormal basis for F_0 in $\langle\cdot,\cdot\rangle_0$, that is $\langle\phi_{\vec{\boldsymbol{\nu}}}(\mathbf{t}), \phi_{\vec{\boldsymbol{\mu}}}(\mathbf{t})\rangle_0 = \delta_{\vec{\boldsymbol{\nu}}\,\vec{\boldsymbol{\mu}}}$, where $\delta_{\vec{\boldsymbol{\nu}}\,\vec{\boldsymbol{\mu}}}$ is Kronecker's

delta. Hence, any $f \in \mathcal{H}$ has the decomposition

$$f(t_1, \ldots, t_d) = \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} f_{\vec{\boldsymbol{\nu}}} \phi_{\vec{\boldsymbol{\nu}}}(t_1, \ldots, t_d), \quad \text{where } f_{\vec{\boldsymbol{\nu}}} = \langle f(\mathbf{t}), \phi_{\vec{\boldsymbol{\nu}}}(\mathbf{t}) \rangle_0. \tag{A.22}$$

We denote a positive scalar series $\{\rho_{\vec{\boldsymbol{\nu}}}\}_{\boldsymbol{\nu} \in \mathbf{V}}$ such that $\langle \phi_{\vec{\boldsymbol{\nu}}}, \phi_{\vec{\boldsymbol{\mu}}} \rangle_R = (1 + \rho_{\vec{\boldsymbol{\nu}}}) \delta_{\vec{\boldsymbol{\nu}}\vec{\boldsymbol{\mu}}}$. Then,

$$J(f) = \langle f, f \rangle_R - \langle f, f \rangle_0 = \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \rho_{\vec{\boldsymbol{\nu}}} f_{\vec{\boldsymbol{\nu}}}^2. \tag{A.23}$$

First, we analyze the deterministic error $(\bar{f} - f_0)$. By (A.22), write $f_0(\mathbf{t}) = \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} f^0_{\vec{\boldsymbol{\nu}}} \phi_{\vec{\boldsymbol{\nu}}}(\mathbf{t})$ and $\bar{f}(\mathbf{t}) = \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \bar{f}_{\vec{\boldsymbol{\nu}}} \phi_{\vec{\boldsymbol{\nu}}}(\mathbf{t})$. Note the bias satisfies $\mathbb{E}[\epsilon_i^{(j)}] = o(n^{-1/2})$, we have $l_\infty(f) = \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} (f_{\vec{\boldsymbol{\nu}}} - f^0_{\vec{\boldsymbol{\nu}}})^2 + o(n^{-1/2}) \sqrt{\sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} (f_{\vec{\boldsymbol{\nu}}} - f^0_{\vec{\boldsymbol{\nu}}})^2} + 1$ and

$$\bar{f}_{\vec{\boldsymbol{\nu}}} = \frac{f^0_{\vec{\boldsymbol{\nu}}}(1 + \kappa_{\vec{\boldsymbol{\nu}}})}{1 + \kappa_{\vec{\boldsymbol{\nu}}} + \lambda \rho_{\vec{\boldsymbol{\nu}}}}, \quad \text{where } \kappa_{\vec{\boldsymbol{\nu}}} = o(1), \ \forall \vec{\boldsymbol{\nu}} \in \mathbf{V}. \tag{A.24}$$

An upper bound of the deterministic error will be given in Lemma A.3.

Second, we analyze the stochastic error $(\widehat{f} - \bar{f})$. The existence of the following Fréchet derivatives is guaranteed by Lemma A.1:

$$\begin{aligned} Dl_n(f)g = \frac{2}{n(p+1)} \Bigg[& \frac{1}{\sigma_0^2} \sum_{i=1}^{n} \{ f(\mathbf{t}_i^{(0)}) - y_i^{(0)} \} g(\mathbf{t}_i^{(0)}) \\ & + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \left\{ \frac{\partial f(\mathbf{t}_i^{(j)})}{\partial t_j} - y_i^{(j)} \right\} \frac{\partial g(\mathbf{t}_i^{(j)})}{\partial t_j} \Bigg], \end{aligned} \tag{A.25}$$

$$\begin{aligned} Dl_\infty(f)g = \frac{2}{p+1} \Bigg[& \frac{1}{\sigma_0^2} \int \left\{ f(\mathbf{t}) - f_0(\mathbf{t}) + o(n^{-1/2}) \right\} g(\mathbf{t}) d\Pi^{(0)}(\mathbf{t}) \\ & + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \int \left\{ \frac{\partial f(\mathbf{t})}{\partial t_j} - \frac{\partial f_0(\mathbf{t})}{\partial t_j} + o(n^{-1/2}) \right\} \frac{\partial g(\mathbf{t})}{\partial t_j} d\Pi^{(j)}(\mathbf{t}) \Bigg], \end{aligned} \tag{A.26}$$

$$D^2 l_n(f)gh = \frac{2}{n(p+1)}\left[\frac{1}{\sigma_0^2}\sum_{i=1}^{n} g(\mathbf{t}_i^{(0)})h(\mathbf{t}_i^{(0)}) + \sum_{j=1}^{p}\frac{1}{\sigma_j^2}\sum_{i=1}^{n}\frac{\partial g(\mathbf{t}_i^{(j)})}{\partial t_j}\frac{\partial h(\mathbf{t}_i^{(j)})}{\partial t_j}\right], \tag{A.27}$$

$$D^2 l_\infty(f)gh = \frac{2}{p+1}\left[\frac{1}{\sigma_0^2}\int g(\mathbf{t})h(\mathbf{t})d\Pi^{(0)}(\mathbf{t}) + \sum_{j=1}^{p}\frac{1}{\sigma_j^2}\int\frac{\partial g(\mathbf{t})}{\partial t_j}\frac{\partial h(\mathbf{t})}{\partial t_j}d\Pi^{(j)}(\mathbf{t})\right] = 2\langle g, h\rangle_0, \tag{A.28}$$

where $Dl_n(f)$, $Dl_\infty(f)$, $D^2 l_n(f)g$, and $D^2 l_\infty(f)g$ are bounded linear operators on $\mathcal{H}$. By Riesz representation theorem, with slight abuse of notation, write

$$Dl_n(f)g = \langle Dl_n(f), g\rangle_R, \quad Dl_\infty(f)g = \langle Dl_\infty(f), g\rangle_R,$$
$$D^2 l_n(f)gh = \langle D^2 l_n(f)g, h\rangle_R, \quad D^2 l_\infty(f)gh = \langle D^2 l_\infty(f)g, h\rangle_R.$$

From Oden and Reddy (2012); Weinberger (1974), there exists a bounded linear operator $U : F_0 \mapsto \mathcal{H}$ such that $U\phi_{\vec{\boldsymbol{\nu}}} = (1+\rho_{\vec{\boldsymbol{\nu}}})^{-1}\phi_{\vec{\boldsymbol{\nu}}}$ and $\langle f, Ug\rangle_R = \langle f, g\rangle_0$ for any $f \in \mathcal{H}$ and $g \in F_0$, and the restriction of U to $\mathcal{H}$ is self-adjoint and positive definite. By (A.28), we further derive

$$D^2 l_{\infty\lambda}(f)\phi_{\vec{\boldsymbol{\nu}}}(\mathbf{t}) = 2(U + \lambda(I-U))\phi_{\vec{\boldsymbol{\nu}}}(\mathbf{t}) = 2(1+\rho_{\vec{\boldsymbol{\nu}}})^{-1}(1+\lambda\rho_{\vec{\boldsymbol{\nu}}})\phi_{\vec{\boldsymbol{\nu}}}(\mathbf{t}).$$

Define that $G_\lambda\phi_{\vec{\boldsymbol{\nu}}} = \frac{1}{2}D^2 l_{\infty\lambda}(\bar{f})\phi_{\vec{\boldsymbol{\nu}}}$. By the Lax-Milgram theorem, $G_\lambda : \mathcal{H} \mapsto \mathcal{H}$ has a bounded inverse G_λ^{-1} on $\mathcal{H}$, and

$$G_\lambda^{-1}\phi_{\vec{\boldsymbol{\nu}}} = (1+\rho_{\vec{\boldsymbol{\nu}}})(1+\lambda\rho_{\vec{\boldsymbol{\nu}}})^{-1}\phi_{\vec{\boldsymbol{\nu}}}. \tag{A.29}$$

Define

$$\tilde{f}^* = \bar{f} - \frac{1}{2}G_\lambda^{-1}Dl_{n\lambda}(\bar{f}).$$

Then the stochastic error can be decomposed as

$$\widehat{f} - \bar{f} = (\tilde{f}^* - \bar{f}) + (\widehat{f} - \tilde{f}^*).$$

The two terms on the right-hand side will be studied separately and their upper bounds will be given in Lemma A.4 and Lemma A.5, respectively.

Finally, we define the following norm to serve as a basis for further development. For $f \in \mathcal{H}$,

$$\|f\|^2_{L_2(a)} = \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \left(1 + \frac{\rho_{\vec{\boldsymbol{\nu}}}}{\|\phi_{\vec{\boldsymbol{\nu}}}\|^2_{L_2}}\right)^a f^2_{\vec{\boldsymbol{\nu}}} \|\phi_{\vec{\boldsymbol{\nu}}}\|^2_{L_2}, \quad \text{for } 0 \le a \le 1, \tag{A.30}$$

where $f_{\vec{\boldsymbol{\nu}}} = \langle f, \phi_{\vec{\boldsymbol{\nu}}} \rangle_0$. By direct calculations, when $a = 0$ this norm coincides with $\|\cdot\|_{L_2}$ on F_0, and when $a = 1$ this norm is equivalent to $\|\cdot\|_R$ on $\mathcal{H}$.

Details of the proof Now we give the details by following the above outline. First, we present an upper bound of the deterministic error $(\bar{f} - f_0)$.

Lemma A.3. *For any $0 \le a \le 1$, the deterministic error satisfies*

$$\|\bar{f} - f_0\|^2_{L_2(a)} = \begin{cases} O\left\{\lambda^{1-a} J(f_0)\right\} & \text{when } 0 \le p < d, \\ O\{\lambda^{\frac{(1-a)mr}{mr-1}} J(f_0)\} & \text{when } p = d. \end{cases}$$

Proof. For any $0 \le a \le 1$, by (A.23) and (A.24), we have

$$\begin{aligned} &\|\bar{f} - f_0\|^2_{L_2(a)} \\ &= \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \left(1 + \frac{\rho_{\vec{\boldsymbol{\nu}}}}{\|\phi_{\vec{\boldsymbol{\nu}}}\|^2_{L_2}}\right)^a \left(\frac{\lambda \rho_{\vec{\boldsymbol{\nu}}}}{1 + \kappa_{\vec{\boldsymbol{\nu}}} + \lambda \rho_{\vec{\boldsymbol{\nu}}}}\right)^2 (f^0_{\vec{\boldsymbol{\nu}}})^2 \|\phi_{\vec{\boldsymbol{\nu}}}\|^2_{L_2} \\ &\lesssim \lambda^2 \sup_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \frac{(1 + \rho_{\vec{\boldsymbol{\nu}}} / \|\phi_{\vec{\boldsymbol{\nu}}}\|^2_{L_2})^a \rho_{\vec{\boldsymbol{\nu}}} \|\phi_{\vec{\boldsymbol{\nu}}}\|^2_{L_2}}{(1 + \lambda \rho_{\vec{\boldsymbol{\nu}}})^2} \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \rho_{\vec{\boldsymbol{\nu}}} (f^0_{\vec{\boldsymbol{\nu}}})^2 \\ &\lesssim \lambda^2 J(f_0) \sup_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \frac{(\prod_{k=1}^d \nu_k^{2m})^{1+a}}{(1 + \sum_{j=1}^p \nu_j^2 + \lambda \prod_{k=1}^d \nu_k^{2m})^2}. \end{aligned} \tag{A.31}$$

Write

$$B_\lambda(\overrightarrow{\boldsymbol{\nu}}) = \frac{(\prod_{k=1}^d \nu_k^{2m})^{1+a}}{(1+\sum_{j=1}^p \nu_j^2 + \lambda \prod_{k=1}^d \nu_k^{2m})^2}, \quad \overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}.$$

We discuss $B_\lambda(\overrightarrow{\boldsymbol{\nu}})$ for $0 \le p \le d-1$ and $p = d$ separately.

For $0 \le p \le d-1$, since $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$, there are at most r of $\nu_1, \ldots, \nu_d$ not equal to 1. Suppose for any $x = \prod_{k=1}^d \nu_k^{-2m} > 0$ fixed. Then $B_\lambda(\overrightarrow{\boldsymbol{\nu}})$ is maximized by letting $\sum_{j=1}^p \nu_j^2$ be as small as possible, which implies $\nu_1 = \nu_2 = \cdots = \nu_p = 1$. Then,

$$\begin{aligned} \sup_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} B_\lambda(\overrightarrow{\boldsymbol{\nu}}) &\asymp \sup_{(\nu_{p+1}, \ldots, \nu_{(p+r)\wedge d})^\top \in \mathbb{N}^{r \wedge (d-p)}} \frac{\prod_{k=p+1}^{(p+r)\wedge d} \nu_k^{2m(1+a)}}{(1+\lambda \prod_{k=p+1}^{(p+r)\wedge d} \nu_k^{2m})^2} \\ &\asymp \sup_{x>0} \frac{x^{-(1+a)}}{(1+\lambda x^{-1})^2} \asymp \lambda^{-(a+1)}, \end{aligned} \tag{A.32}$$

where the last step is achieved when $x \asymp \lambda$.

For $p = d$, since $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$ and by the symmetry of coordinates $v_1, \ldots, v_d$, assume that all indices except $v_1, \ldots, v_r$ being 1. Letting $z = \prod_{j=1}^r \nu_j^{-2m} > 0$, we have

$$\sup_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} B_\lambda(\overrightarrow{\boldsymbol{\nu}}) \asymp \sup_{z>0} \frac{z^{-(1+a)}}{(z^{-1/mr} + \lambda z^{-1})^2} \asymp \lambda^{\frac{2-(1+a)mr}{mr-1}}, \tag{A.33}$$

where the last step is achieved when $z \asymp \lambda^{mr/(mr-1)}$. Combining (A.31), (A.32) and (A.33) we complete the proof. □

Second, we establish an upper bound of $(\tilde{f}^* - \bar{f})$, which is a part of the stochastic error.

Lemma A.4. *When $0 \le p < d$, we have for any $0 \le a < 1 - 1/2m$,*

$$\|\tilde{f}^* - \bar{f}\|_{L_2(a)}^2 = O_{\mathbb{P}}\left\{ n^{-1} \lambda^{-(a+1/2m)} [\log(1/\lambda)]^{(d-p)\wedge r - 1} \right\}.$$

When $p = d$, we have for any $0 \le a \le 1$,

$$\|\tilde{f}^* - \bar{f}\|_{L_2(a)}^2 = \begin{cases} O_{\mathbb{P}}\left\{n^{-1}\lambda^{\frac{mr}{1-mr}\left(a+\frac{r-2}{2mr}\right)}\right\}, \text{ if } r \ge 3; \\ O_{\mathbb{P}}\left\{n^{-1}\log(1/\lambda)\right\}, \text{ if } r = 2, a = 0; \quad O_{\mathbb{P}}\left\{n^{-1}\right\}, \text{ if } r = 2, 0 < a \le 1; \\ O_{\mathbb{P}}\left\{n^{-1}\right\}, \text{ if } r = 1, a < \frac{1}{2m}; \quad O_{\mathbb{P}}\left\{n^{-1}\log(1/\lambda)\right\}, \text{ if } r = 1, a = \frac{1}{2m}; \\ O_{\mathbb{P}}\left\{n^{-1}\lambda^{\frac{1-2ma}{2m-2}}\right\}, \text{ if } r = 1, a > \frac{1}{2m}. \end{cases}$$

Proof. Notice that $Dl_{n,\lambda}(\bar{f}) = Dl_{n,\lambda}(\bar{f}) - Dl_{\infty,\lambda}(\bar{f}) = Dl_n(\bar{f}) - Dl_\infty(\bar{f})$. Hence, for any $g \in \mathcal{H}$,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{2}Dl_{n,\lambda}(\bar{f})g\right]^2 = \mathbb{E}\left[\frac{1}{2}Dl_n(\bar{f})g - \frac{1}{2}Dl_\infty(\bar{f})g\right]^2 \\
&\lesssim \frac{1}{n(p+1)^2}\sum_{j=0}^{p}\mathrm{Var}\left[\frac{1}{\sigma_j^2}\left\{\frac{\partial\bar{f}(\mathbf{t}^{(j)})}{\partial t_j} - Y^{(j)}\right\}\frac{\partial g(\mathbf{t}^{(j)})}{\partial t_j}\right] \\
&+ \sum_{j=0}^{p}\frac{\sigma_j^{-4}}{n^2(p+1)^2}\sum_{i\neq i'}\mathrm{Cov}\left[\left(\frac{\partial\bar{f}(\mathbf{t}_i^{(j)})}{\partial t_j} - y_i^{(j)}\right)\frac{\partial g(\mathbf{t}_i^{(j)})}{\partial t_j}, \left(\frac{\partial\bar{f}(\mathbf{t}_{i'}^{(j)})}{\partial t_j} - y_{i'}^{(j)}\right)\frac{\partial g(\mathbf{t}_{i'}^{(j)})}{\partial t_j}\right] \\
&+ \sum_{j\neq k}\frac{\sigma_j^{-2}\sigma_k^{-2}}{n^2(p+1)^2}\sum_{i,i'=1}^{n}\mathrm{Cov}\left[\left(\frac{\partial\bar{f}(\mathbf{t}_i^{(j)})}{\partial t_j} - y_i^{(j)}\right)\frac{\partial g(\mathbf{t}_i^{(j)})}{\partial t_j}, \left(\frac{\partial\bar{f}(\mathbf{t}_{i'}^{(k)})}{\partial t_k} - y_{i'}^{(j)}\right)\frac{\partial g(\mathbf{t}_{i'}^{(k)})}{\partial t_k}\right] \\
&\lesssim \frac{1}{n(p+1)}\left[\frac{1}{\sigma_0^4}\mathbb{E}\left\{\bar{f}(\mathbf{t}^{(0)}) - f_0(\mathbf{t}^{(0)})\right\}^2\{g(\mathbf{t}^{(0)})\}^2 + \frac{1}{\sigma_0^2}\mathbb{E}\{g(\mathbf{t}^{(0)})\}^2\right. \\
&\left.+ \sum_{j=1}^{p}\frac{1}{\sigma_j^4}\mathbb{E}\left\{\frac{\partial\bar{f}(\mathbf{t}^{(j)})}{\partial t_j} - \frac{\partial f_0(\mathbf{t}^{(j)})}{\partial t_j}\right\}^2\left\{\frac{\partial g(\mathbf{t}^{(j)})}{\partial t_j}\right\}^2 + \sum_{j=1}^{p}\frac{1}{\sigma_j^2}\mathbb{E}\left\{\frac{\partial g(\mathbf{t}^{(j)})}{\partial t_j}\right\}^2\right] \\
&+ o(n^{-1})\frac{1}{(p+1)^2}\sum_{j,k=0}^{p}\mathbb{E}\left[\frac{\partial g(\mathbf{t}^{(j)})}{\partial t_j}\right]\mathbb{E}\left[\frac{\partial g(\mathbf{t}^{(k)})}{\partial t_k}\right] \\
&\lesssim \frac{1}{n(p+1)}\left[\frac{1}{\sigma_0^4}c_K^{2d}\|\bar{f} - f_0\|_R^2\mathbb{E}\left\{g(\mathbf{t}^{(0)})\right\}^2 + \frac{1}{\sigma_0^2}\mathbb{E}\left\{g(\mathbf{t}^{(0)})\right\}^2\right. \\
&\left.+ \sum_{j=1}^{p}\frac{1}{\sigma_j^4}c_K^{2d}\|\bar{f} - f_0\|_R^2\mathbb{E}\left\{\frac{\partial g(\mathbf{t}^{(j)})}{\partial t_j}\right\}^2 + \sum_{j=0}^{p}\frac{1}{\sigma_j^2}\mathbb{E}\left\{\frac{\partial g(\mathbf{t}^{(j)})}{\partial t_j}\right\}^2\right] \lesssim n^{-1}\|g\|_0^2,
\end{aligned} \tag{A.34}$$

where the second step is due to $\sum_{i\neq i'}\mathrm{Cov}[\epsilon_i^{(j)}, \epsilon_{i'}^{(k)}] = \sum_{i\neq i'} o(|i-i'|^{-\Upsilon}) = o(n)$. The third

step above is by Lemma A.2, A.14 and the Cauchy-Schwarz inequality. The last step above is by Lemma A.3 and the definition of the norm $\|\cdot\|_0$. From the definition of G_λ^{-1} in (A.29), we have that $\forall g \in \mathcal{H}$,

$$\left\|G_\lambda^{-1} g\right\|_{L_2(a)}^2 = \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \left(1 + \frac{\rho_{\vec{\boldsymbol{\nu}}}}{\|\phi_{\vec{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a (1 + \lambda \rho_{\vec{\boldsymbol{\nu}}})^{-2} \|\phi_{\vec{\boldsymbol{\nu}}}\|_{L_2}^2 \langle g, \phi_{\vec{\boldsymbol{\nu}}} \rangle_R^2.$$

Then by the definition of $\tilde{f}^*$,

$$\begin{aligned}
&\mathbb{E}\|\tilde{f}^* - \bar{f}\|_{L_2(a)}^2 = \mathbb{E}\left\|\frac{1}{2} G_\lambda^{-1} D l_{n\lambda}(\bar{f})\right\|_{L_2(a)}^2 \\
&= \frac{1}{4}\mathbb{E}\left[\sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \left(1 + \frac{\rho_{\vec{\boldsymbol{\nu}}}}{\|\phi_{\vec{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a (1 + \lambda \rho_{\vec{\boldsymbol{\nu}}})^{-2} \|\phi_{\vec{\boldsymbol{\nu}}}\|_{L_2}^2 \langle D l_{n\lambda}(\bar{f}), \phi_{\vec{\boldsymbol{\nu}}} \rangle_R^2\right] \\
&\leq \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \left(1 + \frac{\rho_{\vec{\boldsymbol{\nu}}}}{\|\phi_{\vec{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a (1 + \lambda \rho_{\vec{\boldsymbol{\nu}}})^{-2} \|\phi_{\vec{\boldsymbol{\nu}}}\|_{L_2}^2 \mathbb{E}\left[\frac{1}{2} D l_{n\lambda}(\bar{f}) \phi_{\vec{\boldsymbol{\nu}}}\right]^2 \\
&\lesssim n^{-1} \sum_{\vec{\boldsymbol{\nu}} \in \mathbf{V}} \left(1 + \frac{\rho_{\vec{\boldsymbol{\nu}}}}{\|\phi_{\vec{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a (1 + \lambda \rho_{\vec{\boldsymbol{\nu}}})^{-2} \|\phi_{\vec{\boldsymbol{\nu}}}\|_{L_2}^2 \|\phi_{\vec{\boldsymbol{\nu}}}\|_0^2 \\
&\asymp n^{-1} N_a(\lambda),
\end{aligned}$$

where the fourth step is by (A.34) and the last step is because of $\|\phi_{\vec{\boldsymbol{\nu}}}\|_0 = 1$, $\|\phi_{\vec{\boldsymbol{\nu}}}\|_{L_2}^2 \asymp (1 + \sum_{j=1}^p \nu_j^2)^{-1}$, $\rho_{\vec{\boldsymbol{\nu}}} \asymp (1 + \sum_{j=1}^p \nu_j^2)^{-1} \prod_{k=1}^d \nu_k^{2m}$, and $N_a(\lambda)$ is defined in Lemma A.12. Hence, by Lemma A.12, we complete the proof. □

Then, we give an upper bound of $(\widehat{f} - \tilde{f}^*)$, which is another part of the stochastic error. Since $l_{n\lambda}(f)$ is a quadratic form of f, the Taylor expansion of $D l_{n\lambda}(\widehat{f}) = 0$ at $\bar{f}$ gives

$$D l_{n\lambda}(\bar{f}) + D^2 l_{n\lambda}(\bar{f})(\widehat{f} - \bar{f}) = 0,$$

and by the definition of $\tilde{f}^*$ and G_λ, we have

$$D l_{n\lambda}(\bar{f}) + D^2 l_{\infty\lambda}(\bar{f})(\tilde{f}^* - \bar{f}) = 0.$$

Thus, $G_\lambda(\widehat{f}-\tilde{f}^*) = \frac{1}{2}D^2 l_\infty(\bar{f})(\widehat{f}-\bar{f}) - \frac{1}{2}D^2 l_n(\bar{f})(\widehat{f}-\bar{f})$, and

$$\widehat{f}-\tilde{f}^* = G_\lambda^{-1}\left[\frac{1}{2}D^2 l_\infty(\bar{f})(\widehat{f}-\bar{f}) - \frac{1}{2}D^2 l_n(\bar{f})(\widehat{f}-\bar{f})\right]. \tag{A.35}$$

Lemma A.5. *If $n^{-1}\lambda^{-(2a+3/2m)}[\log(1/\lambda)]^{r-1} \to 0$ and $1/2m < a < (2m-3)/4m$, we have for any $0 \le c \le a+1/m$,*

$$\|\widehat{f}-\tilde{f}^*\|^2_{L_2(c)} = o_{\mathbb{P}}\left\{\|\tilde{f}^*-\bar{f}\|^2_{L_2(c)}\right\}.$$

Proof. A sufficient condition for this lemma is that for any $1/(2m) < a < (2m-3)/(4m)$ and $0 \le c \le a+1/m$,

$$\begin{aligned}
&\|\widehat{f}-\tilde{f}^*\|^2_{L_2(c)} \\
&= \begin{cases}
O_{\mathbb{P}}\left\{n^{-1}\lambda^{-(c+a+1/2m)}[\log(1/\lambda)]^{r\wedge(d-p)-1}\right\} & \\
\qquad\qquad \cdot\|\widehat{f}-\bar{f}\|^2_{L_2(a+1/m)}, & \text{if } 0 \le p < d, \\
O_{\mathbb{P}}\left\{n^{-1}\lambda^{\frac{mr}{1-mr}\left(a+c+\frac{r-2}{2mr}\right)}\right\}\|\widehat{f}-\bar{f}\|^2_{L_2(a+1/m)}, & \text{if } p = d, r \ge 3, \\
O_{\mathbb{P}}\left\{n^{-1}\right\}\|\widehat{f}-\bar{f}\|_{L_2(a+1/m)}, & \text{if } p = d, r = 2, \\
O_{\mathbb{P}}\left\{n^{-1}\lambda^{\frac{1-2m(a+c)}{2m-2}}\right\}\|\widehat{f}-\bar{f}\|_{L_2(a+1/m)}, & \text{if } p = d, r = 1.
\end{cases}
\end{aligned} \tag{A.36}$$

This is because once (A.36) established, by letting $c = a+1/m$ and under the assumption $n^{-1}\lambda^{-(2a+3/2m)}[\log(1/\lambda)]^{r-1} \to 0$, we have

$$\|\widehat{f}-\tilde{f}^*\|^2_{L_2(a+1/m)} = o_{\mathbb{P}}(1)\|\widehat{f}-\bar{f}\|^2_{L_2(a+1/m)}.$$

By the triangle inequality, we have $\|\tilde{f}^*-\bar{f}\|_{L_2(a+1/m)} \ge \|\widehat{f}-\bar{f}\|_{L_2(a+1/m)} - \|\widehat{f}-\tilde{f}^*\|_{L_2(a+1/m)} = [1-o_{\mathbb{P}}(1)]\|\widehat{f}-\bar{f}\|_{L_2(a+1/m)}$, which implies $\|\widehat{f}-\bar{f}\|^2_{L_2(a+1/m)} = O_{\mathbb{P}}\{\|\tilde{f}^*-\bar{f}\|^2_{L_2(a+1/m)}\}$. Thus, by (A.36) and Lemma A.4, we complete the proof.

We now are in the position to prove (A.36). For any $0 \leq c \leq a + 1/m$, by (A.35), we have

$$
\begin{aligned}
&\|\widehat{f} - \tilde{f}^*\|^2_{L_2(c)} \\
&\leq \sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|^2_{L_2}}\right)^c (1 + \lambda\rho_{\overrightarrow{\boldsymbol{\nu}}})^{-2} \|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|^2_{L_2} \cdot \frac{1}{p+1} \cdot \\
&\left\{ \left[\frac{\sum_{i=1}^n (\widehat{f} - \bar{f})(\mathbf{t}_i^{(0)}) \phi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t}_i^{(0)})}{n\sigma_0^2} - \frac{\int (\widehat{f} - \bar{f})(\mathbf{t}) \phi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t}) d\Pi^{(0)}(\mathbf{t})}{\sigma_0^2} \right]^2 + \right. \\
&\left. \sum_{j=1}^p \left[\frac{\sum_{i=1}^n \frac{\partial(\widehat{f} - \bar{f})}{\partial t_j}(\mathbf{t}_i^{(j)}) \frac{\partial \phi_{\overrightarrow{\boldsymbol{\nu}}}}{\partial t_j}(\mathbf{t}_i^{(j)})}{n\sigma_j^2} - \frac{\int \frac{\partial(\widehat{f} - \bar{f})(\mathbf{t})}{\partial t_j} \frac{\partial \phi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})}{\partial t_j} d\Pi^{(j)}(\mathbf{t})}{\sigma_j^2} \right]^2 \right\}.
\end{aligned} \tag{A.37}
$$

For brevity, we denote $f(\mathbf{t}) = \partial f / \partial t_0$. Let $g_j(\mathbf{t}) = \frac{1}{\sigma_j^2} \frac{\partial(\widehat{f} - \bar{f})}{\partial t_j} \frac{\partial \phi_{\overrightarrow{\boldsymbol{\nu}}}}{\partial t_j}$ and $g_0(\mathbf{t}) = \frac{1}{\sigma_0^2}(\widehat{f} - \bar{f})\phi_{\overrightarrow{\boldsymbol{\nu}}}$. Hence, we can do the expansion on the basis $\{\phi_{\overrightarrow{\boldsymbol{\mu}}}\}_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d}$,

$$
g_j(\mathbf{t}) = \sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} Q^j_{\overrightarrow{\boldsymbol{\mu}}} \phi_{\overrightarrow{\boldsymbol{\mu}}}(\mathbf{t}), \quad \text{where } Q^j_{\overrightarrow{\boldsymbol{\mu}}} = \langle g_j(\mathbf{t}), \phi_{\overrightarrow{\boldsymbol{\mu}}}(\mathbf{t}) \rangle_0. \tag{A.38}
$$

Unlike (A.22) with the multi-index $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$, we require $\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d$ in (A.38) since now $g_j(\mathbf{t})$is a product function. By Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\left[\frac{1}{n\sigma_j^2} \sum_{i=1}^n \frac{\partial(\widehat{f} - \bar{f})}{\partial t_j}(\mathbf{t}_i^{(j)}) \frac{\partial \phi_{\overrightarrow{\boldsymbol{\nu}}}}{\partial t_j}(\mathbf{t}_i^{(j)}) - \frac{1}{\sigma_j^2} \int \frac{\partial(\widehat{f} - \bar{f})(\mathbf{t})}{\partial t_j} \frac{\partial \phi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})}{\partial t_j} \right]^2 \\
&= \left[\sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} Q^j_{\overrightarrow{\boldsymbol{\mu}}} \left(\frac{1}{n} \sum_{i=1}^n \phi_{\overrightarrow{\boldsymbol{\mu}}}(\mathbf{t}_i^{(j)}) - \int \phi_{\overrightarrow{\boldsymbol{\mu}}}(\mathbf{t}) \right) \right]^2 \\
&\leq \left[\sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} (Q^j_{\overrightarrow{\boldsymbol{\mu}}})^2 \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\mu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|^2_{L_2}} \right)^a \|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|^2_{L_2} \right] \\
&\cdot \left[\sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\mu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|^2_{L_2}} \right)^{-a} \|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|^{-2}_{L_2} \left(\frac{1}{n} \sum_{i=1}^n \phi_{\overrightarrow{\boldsymbol{\mu}}}(\mathbf{t}_i^{(j)}) - \int \phi_{\overrightarrow{\boldsymbol{\mu}}}(\mathbf{t}) \right)^2 \right].
\end{aligned} \tag{A.39}
$$

By Lemma A.16, if $a > 1/2m$, then the sum of the first part in the right-hand side of (A.39)

over $j = 0, 1, \ldots, p$ is bounded by

$$\begin{aligned}
&\sum_{j=0}^{p} \sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\mu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|_{L_2}^2}\right)^a \|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|_{L_2}^2 \left\langle \frac{\partial(\widehat{f} - \bar{f})}{\partial t_j} \frac{\partial \phi_{\overrightarrow{\boldsymbol{\nu}}}}{\partial t_j}, \phi_{\overrightarrow{\boldsymbol{\mu}}} \right\rangle_0^2 \\
&\lesssim \|\widehat{f} - \bar{f}\|_{L_2(a+1/m)}^2 \sum_{j=0}^{p} \sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\mu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|_{L_2}^2}\right)^a \|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|_{L_2}^2 \left\langle \frac{\partial \phi_{\overrightarrow{\boldsymbol{\nu}}}}{\partial t_j}, \phi_{\overrightarrow{\boldsymbol{\mu}}} \right\rangle_0^2 \\
&\lesssim \|\widehat{f} - \bar{f}\|_{L_2(a+1/m)}^2 \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a \|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2 \left(1 + \sum_{j=1}^{p} \nu_j^2\right) \\
&\asymp \|\widehat{f} - \bar{f}\|_{L_2(a+1/m)}^2 \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a .
\end{aligned} \tag{A.40}$$

The second part in the right-hand side of (A.39) can be bounded by

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\mu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|_{L_2}^2}\right)^{-a} \|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|_{L_2}^{-2} \left(\frac{1}{n}\sum_{i=1}^{n} \phi_{\overrightarrow{\boldsymbol{\mu}}}(\mathbf{t}_i^{(j)}) - \int \phi_{\overrightarrow{\boldsymbol{\mu}}}(\mathbf{t})\right)^2\right] \\
&\quad \le n^{-1} \sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\mu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|_{L_2}^2}\right)^{-a} \|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|_{L_2}^{-2} \int \phi_{\overrightarrow{\boldsymbol{\mu}}}^2(\mathbf{t}) \\
&\quad \asymp n^{-1} \sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\mu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|_{L_2}^2}\right)^{-a} \lesssim n^{-1} \sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} \mu_1^{-2ma} \cdots \mu_d^{-2ma} \\
&\quad \le n^{-1} \left(\sum_{\mu_1=1}^{\infty} \mu_1^{-2ma}\right)^d \asymp n^{-1},
\end{aligned} \tag{A.41}$$

where the third step uses $\rho_{\overrightarrow{\boldsymbol{\mu}}}/\|\phi_{\overrightarrow{\boldsymbol{\mu}}}\|_{L_2}^2 \asymp \mu_1^{2m} \cdots \mu_d^{2m}$, and the fourth step holds for $a > 1/2m$. Combing (A.40) and (A.41), we have that for $a > 1/2m$,

$$\begin{aligned}
&\sum_{j=0}^{p} \mathbb{E}\left[\sum_{\overrightarrow{\boldsymbol{\mu}} \in \mathbb{N}^d} Q_{\overrightarrow{\boldsymbol{\mu}}}^j \left(\frac{1}{n}\sum_{i=1}^{n} \phi_{\overrightarrow{\boldsymbol{\mu}}}(\mathbf{t}_i^{(j)}) - \int \phi_{\overrightarrow{\boldsymbol{\mu}}}(\mathbf{t})\right)\right]^2 \\
&\lesssim \frac{1}{n}\|\widehat{f} - \bar{f}\|_{L_2(a+1/m)}^2 \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a .
\end{aligned} \tag{A.42}$$

Put all together Therefore, if $1/2m < a < (2m-3)/4m$ and $0 \le c \le a + 1/m$, (A.37) and (A.42) imply that

$$\mathbb{E}\|\widehat{f} - \tilde{f}^*\|^2_{L_2(c)} \lesssim n^{-1}\|\widehat{f} - \bar{f}\|^2_{L_2(a+1/m)} N_{a+c}(\lambda).$$

By Lemma A.12 we complete the proof for (A.36) and this lemma. □

Finally, we combine Lemma A.3, Lemma A.4 and Lemma A.5 to obtain the following proposition.

Proposition A.6. *Under the conditions of Theorem 2.3 and assuming the distributions $\Pi^{(0)}$ and $\Pi^{(j)}$s are known. If $1/2m < a < (2m-3)/4m$, $m > 2$, and $n^{-1}\lambda^{-(2a+3/2m)}[\log(1/\lambda)]^{r-1} \to 0$, then for any $c \in [0, a+1/m]$, the $\widehat{f}$ given by (2.10) satisfies, when $0 \le p < d$,*

$$\|\widehat{f} - f_0\|^2_{L_2(c)} = O\{\lambda^{1-c} J(f_0)\} + O_{\mathbb{P}}\left\{n^{-1}\lambda^{-(c+1/2m)}[\log(1/\lambda)]^{r\wedge(d-p)-1}\right\},$$

and when $p = d$,

$$\|\widehat{f} - f_0\|^2_{L_2(c)} = \begin{cases} O\left\{\lambda^{\frac{(1-c)mr}{mr-1}} J(f_0)\right\} + O_{\mathbb{P}}\left\{n^{-1}\lambda^{\frac{mr}{1-mr}\left(c+\frac{r-2}{2mr}\right)}\right\} & \text{if } r \ge 3, \\ O\left\{\lambda^{\frac{2m}{2m-1}} J(f_0)\right\} + O_{\mathbb{P}}\left\{n^{-1}\log(1/\lambda)\right\} \text{ if } r = 2, c = 0, \\ O\left\{\lambda^{\frac{2(1-c)m}{2m-1}} J(f_0)\right\} + O_{\mathbb{P}}\left\{n^{-1}\lambda^{\frac{2mc}{1-2m}}\right\} \text{ if } r = 2, c > 0, \\ O\left\{\lambda^{\frac{(1-c)m}{m-1}} J(f_0)\right\} + O_{\mathbb{P}}\left\{n^{-1}\right\} \text{ if } r = 1, c < \frac{1}{2m}, \\ O\left\{\lambda^{\frac{2m-1}{2(m-1)}} J(f_0)\right\} + O_{\mathbb{P}}\left\{n^{-1}\log(1/\lambda)\right\} \text{ if } r = 1, c = \frac{1}{2m}, \\ O\left\{\lambda^{\frac{(1-c)m}{m-1}} J(f_0)\right\} + O_{\mathbb{P}}\left\{n^{-1}\lambda^{\frac{1-2mc}{2m-2}}\right\} \text{ if } r = 1, c > \frac{1}{2m}. \end{cases}$$

Many results on the regularized estimator $\widehat{f}$ can be derived from Proposition A.6 including Theorem 2.4. In fact, for $p = d$ and $r \ge 3$, by letting $\lambda \asymp n^{-\frac{2mr-2}{(2m+1)r-2}}$, $a = 1/2m + \epsilon$ for some $\epsilon > 0$ and $c = 0$, we have the condition $n^{-1}\lambda^{-(2a+3/2m)}[\log(1/\lambda)]^{r-1} \to 0$ is equivalent

to

$$-1+\frac{5(mr-1)}{2m^2r+mr-2m}<0, \tag{A.43}$$

and $m>2$ is sufficient for (A.43). Thus, the conditions for Proposition A.6 are satisfied. Similarly, we can verify that when $p=d$ and $r=2$, $\lambda \asymp [n(\log n)]^{-(2m-1)/2m}$ satisfies the conditions for Proposition A.6. When $p=d$ and $r=1$, $\lambda \lesssim n^{-(m-1)/m}$ satisfies the conditions for the above proposition. When $0\le p\le d-r$, $\lambda \asymp [n(\log n)^{1-r}]^{-2m/(2m+1)}$ satisfies the conditions for the above Proposition, as well as when $d-r<p<d$ by letting $\lambda \asymp [n(\log n)^{1+p-d}]^{-2m/(2m+1)}$. This completes the proof for Theorem 2.4.

A.3.3 Proof of Corollary 2.5

This corollary can be directly derived from Proposition A.6 in the main text. Observe that

$$\int_{\mathcal{X}_1^d}\left[\frac{\partial^d \widehat{f}_{n\lambda}(\mathbf{t})}{\partial t_1\cdots\partial t_d}-\frac{\partial^d f_0(\mathbf{t})}{\partial t_1\cdots\partial t_d}\right]^2 d\mathbf{t} \asymp \|\widehat{f}_{n\lambda}-f_0\|_{L_2(1/m)}.$$

If $d-r<p<d$, we let $c=a=1/m$ and $\lambda \asymp [n(\log n)^{1+p-d}]^{-2m/(2m+1)}$ in Proposition A.6, then the condition $n^{-1}\lambda^{-(2a+3/2m)}[\log(1/\lambda)]^{r-1}\to 0$ is equivalent to

$$-1+7/(2m+1)<0, \tag{A.44}$$

and $m>3$ is sufficient for (A.44). Thus the condition for Proposition A.6 are satisfied, and Proposition A.6 yields the rate of convergence for $\|\widehat{f}_{n\lambda}-f_0\|_{L_2(1/m)}$ is

$$O_{\mathbb{P}}\left([n(\log n)^{1+p-d}]^{-2(m-1)/(2m+1)}\right).$$

Similarly, if $0\le p\le d-r$, we let $\lambda \asymp [n(\log n)^{1-r}]^{-2m/(2m+1)}$; if $p=d$ and $r\ge 3$, let $\lambda \asymp n^{-2(mr-1)/(2mr+r-2)}$; if $p=d$ and $r=2$, let $\lambda \asymp n^{-(2m-1)/2m}$; if $p=d$ and $r=1$, let $\lambda \asymp n^{-(2m-2)/(2m-1)}$, then the conditions for Proposition A.6 will be satisfied. This completes the proof.

A.4 Proofs of Results in Section 2.5: Estimating Partial Derivatives

We now turn to prove the results for estimating partial derivatives under the random design.

A.4.1 Proof of Minimax Lower Bound: Theorem 2.6

The minimax lower bound will be established by using Fano's lemma but the proof is different from Section A.3.1 in construction details. It suffices to consider a special case that noises $\epsilon^{(0)}$ and $\epsilon^{(j)}$s are Gaussian with $\sigma_0 = 1$ and $\sigma_j = 1$, and $\Pi^{(0)}$ and $\Pi^{(j)}$s are uniform, and $\mathcal{H}_1$ is generated by periodic kernels. For simplicity, we still use the notation introduced in Section A.3.1. In the rest of this section, without less of generality, we consider estimating $\partial f_0/\partial t_1(\cdot)$ with $p \geq 1$.

First, the number of multi-indices $\overrightarrow{\boldsymbol{\nu}} = (\nu_1, \ldots, \nu_r) \in \mathbb{N}^r$ satisfying

$$\nu_1^{(m-1)/m}\nu_2 \cdots \nu_r \leq N$$

is $c_0' N^{m/(m-1)}$, where c_0' is some constant. Define a length-$\{c_0' N^{m/(m-1)}\}$ binary sequence as

$$b = \{b_{\overrightarrow{\boldsymbol{\nu}}} : \nu_1^{(m-1)/m}\nu_2 \cdots \nu_r \leq N\} \in \{0,1\}^{c_0' N^{m/(m-1)}}.$$

We write

$$\begin{aligned} h_b(t_1, \ldots, t_r) = N^{-m/2(m-1)} \sum_{\nu_1^{(m-1)/m}\nu_2 \cdots \nu_r \leq N} & b_{\overrightarrow{\boldsymbol{\nu}}} \left(1 + \nu_1^2 + \cdots + \nu_r^2\right)^{-1/2} \\ & \times \left[\nu_1^{(m-1)/m}\nu_2 \cdots \nu_r + N\right]^{-m} \psi_{\nu_1}(t_1)\psi_{\nu_2}(t_2) \cdots \psi_{\nu_r}(t_r). \end{aligned}$$

where $\psi_{\nu_k}(t_j)$s are the trigonometric basis in (A.2). Note that

$$\begin{aligned}\|h_b\|_{\mathcal{H}}^2 &\lesssim N^{-m/(m-1)} \sum_{\nu_1^{(m-1)/m}\nu_2\cdots\nu_r\le N} b_{\overrightarrow{\boldsymbol{\nu}}}^2 \nu_1^2 \left(1+\nu_1^2+\cdots+\nu_r^2\right)^{-1} \\ &\le N^{-m/(m-1)} \sum_{\nu_1^{(m-1)/m}\nu_2\cdots\nu_r\le N} \nu_1^2 \left(1+\nu_1^2+\cdots+\nu_r^2\right)^{-1} \asymp 1,\end{aligned}$$

where the last step is by Lemma A.21 in Section A.6. Hence, $h_b(\cdot) \in \mathcal{H}$.

Then, using the Varshamov-Gilbert bound, there exists a collection of binary sequences $\{b^{(1)},\ldots,b^{(M)}\} \subset \{0,1\}^{c_0' N^{m/(m-1)}}$ such that

$$M \ge 2^{c_0' N^{m/(m-1)}/8}$$

and

$$H(b^{(l)}, b^{(q)}) \ge c_0' N^{m/(m-1)}/8, \quad \forall 1 \le l < q \le M.$$

For $b^{(l)}, b^{(q)} \in \{0,1\}^{c_0' N^{m/(m-1)}}$, we have

$$\begin{aligned}&\left\| \frac{\partial h_{b^{(l)}}}{\partial t_1} - \frac{\partial h_{b^{(q)}}}{\partial t_1} \right\|_{L_2}^2 \\ &\ge c' N^{-m/(m-1)} (2N)^{-2m} \sum_{\nu_1^{(m-1)/m}\nu_2\cdots\nu_r\le N} \nu_1^2 (1+\nu_1^2+\cdots+\nu_r^2)^{-1} \left[b_{\overrightarrow{\boldsymbol{\nu}}}^{(l)} - b_{\overrightarrow{\boldsymbol{\nu}}}^{(q)} \right]^2 \\ &\ge c' N^{-m/(m-1)} (2N)^{-2m} \sum_{c_1' 7N/8 \le \nu_1^{(m-1)/m}\nu_2\cdots\nu_r\le N} \nu_1^2 (1+\nu_1^2+\cdots+\nu_r^2)^{-1} \\ &= c_2' N^{-2m}\end{aligned}$$

for some constant c', c_1' and c_2', where the last step is by Lemma A.21 in Section A.6. On the

other hand, for any $b^{(l)} \in \{b^{(1)}, \ldots, b^{(M)}\}$,

$$\begin{aligned}
&\|h_{b^{(l)}}\|_{L_2}^2 + \sum_{j=1}^{p} \|\partial h_{b^{(l)}}/\partial t_j\|_{L_2}^2 \\
&\leq N^{-m/(m-1)} N^{-2m} \sum_{\nu_1^{(m-1)/m}\nu_2\cdots\nu_r \leq N} \left[b_{\vec{\boldsymbol{\nu}}}^{(l)}\right]^2 \leq c_3' N^{-2m}
\end{aligned}$$

with some constant c_3', where the last step is a corollary of Lemma A.21.

Last, by the same argument as (A.15), (E.15), (A.17) and (A.18) in the main text, we obtain

$$\begin{aligned}
\infty_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{P}\left\{ \left\| \tilde{f}(\mathbf{t}) - \frac{\partial f_0(\mathbf{t})}{\partial t_1} \right\|_{L_2}^2 \geq \frac{1}{4} c_2' N^{-2m} \right\} \\
\geq 1 - \frac{2c_3' n(p+1) N^{-2m} + \log 2}{c_0'(\log 2) N^{m/(m-1)}/8}.
\end{aligned}$$

Taking $N = c_4' n^{(m-1)/(2m^2-m)}$ with an appropriately chosen c_4', we have

$$\limsup_{n\to\infty} \infty_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{P}\left\{ \left\| \tilde{f}(\mathbf{t}) - \frac{\partial f_0(\mathbf{t})}{\partial t_1} \right\|_{L_2}^2 \geq C_4 n^{-2(m-1)/(2m-1)} \right\} > 0,$$

where the constant factor C_4 does not depend on n. This completes the proof.

A.4.2 Proof of Minimax Upper Bound: Theorem 2.7

We continue to use the notation and definitions such as the minimizer $\bar{f}$, the Fréchet derivatives $Dl_n(f)g$, $Dl_\infty(f)g$, $D^2 l_n(f)gh$, $D^2 l_\infty(f)gh$, the operator G_λ^{-1} and most importantly $\tilde{f}^*$ in Section A.3.2. Unlike Section A.3.2, here we do not require $\Pi^{(j)}$s are known nor f_0 has periodic boundaries on $\mathcal{X}_1^d$ by some transformation. For brevity, we consider the random errors to be centered and independent in this proof while the general error structure (2.4) can be similarly studied as Section A.3.2.

By the assumption that $\Pi^{(j)}$s are bounded away from 0 and infinity, we have for any $1 \leq j \leq p$,

$$\int_{\mathcal{X}_1^d} \left[\frac{\partial \widehat{f}_{n\lambda}(\mathbf{t})}{\partial t_j} - \frac{\partial f_0(\mathbf{t})}{\partial t_j} \right]^2 d\mathbf{t} \lesssim \|\widehat{f} - f_0\|_0^2.$$

Hence, the following lemma is sufficient for proving Theorem 2.7.

Lemma A.7. *Under the conditions of Theorem 2.6, then $\widehat{f}_{n\lambda}$ given by (2.10) satisfies*

$$\lim_{C_4'\to\infty} \limsup_{n\to\infty} \sup_{f_0\in\mathcal{H}} \mathbb{P}\left\{\|\widehat{f} - f_0\|_0^2 > C_4' n^{-2(m-1)/(2m-1)}\right\} = 0,$$

if the tuning parameter λ is chosen by $\lambda \asymp n^{-2(m-1)/(2m-1)}$.

A lemma for the proof In $\mathcal{H}$, the quadratic form $\langle f, f\rangle_0$ is completely continuous with respect to $\langle f, f\rangle_R$. By the theory in Section 3.3 of Weinberger Weinberger (1974), there exists an eigen-decomposition for the generalized Rayleigh quotient $\langle f, f\rangle_0/\langle f, f\rangle_R$ in $\mathcal{H}$, where we denote the eigenvalues are $\{(1+\gamma_\nu)^{-1}\}_{\nu\geq 1}$ and the corresponding eigenfunctions are $\{(1+\gamma_\nu)^{-1/2}\xi_\nu\}_{\nu\geq 1}$. Thus, $\langle \xi_\nu, \xi_\mu\rangle_R = (1+\gamma_\nu)\delta_{\nu\mu}$ and $\langle \xi_\nu, \xi_\mu\rangle_0 = \delta_{\nu\mu}$, where $\delta_{\nu\mu}$ is Kronecker's delta. The following lemma gives the decay rate of γ_ν and its proof is given in Section A.5.2.

Lemma A.8. *By the well-ordering principle, the elements in the set*

$$\left\{\left(1+\sum_{j=1}^{p}\nu_j^2\right)\prod_{k=1}^{d}\nu_k^{-2m} : \overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}\right\}$$

can be ordered from large to small, where V is defined in (A.1). Denote by $\{\gamma_\nu'\}_{\nu\geq 1}$ the ordered sequence. Then $\gamma_\nu \asymp (\gamma_\nu')^{-1}$.

The proof of this lemma is delegated to Section A.5.2. The lemma bridges the gap between the proof needed for Lemma A.7 and the proof for Theorem 2.4 shown in Section A.3.2 since the eigenvalues $\rho_{\overrightarrow{\boldsymbol{\nu}}}$ in Section A.3.2 satisfies $\rho_{\overrightarrow{\boldsymbol{\nu}}} \asymp (1+\sum_{j=1}^{p}\nu_j^2)^{-1}\prod_{k=1}^{d}\nu_k^{2m}$. Hence in later analysis, we can exchange the use of $\{\gamma_\nu, \nu\in\mathbb{N}\}$ and $\{\rho_{\overrightarrow{\boldsymbol{\nu}}} : \overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}\}$ in some asymptotic calculation settings.

For any function $f \in \mathcal{H}$, it can be decomposed as

$$f(t_1, \ldots, t_d) = \sum_{\nu \in \mathbb{N}} f_\nu \xi_\nu(t_1, \ldots, t_d), \quad \text{where } f_\nu = \langle f(\mathbf{t}), \xi_\nu(\mathbf{t}) \rangle_0,$$

and $J(f) = \langle f, f \rangle_R - \langle f, f \rangle_0 = \sum_{\nu \in \mathbb{N}} \gamma_\nu f_\nu^2$.

First, we present an upper bound of the deterministic error $(\bar{f} - f_0)$.

Lemma A.9. *The deterministic error satisfies*

$$\|\bar{f} - f_0\|_0^2 = O\left\{\lambda J(f_0)\right\}.$$

Proof. For any $0 \leq a \leq 1$,

$$\begin{aligned}
\|\bar{f} - f_0\|_0^2 &= \sum_{\nu=1}^{\infty} \left(\frac{\lambda \gamma_\nu}{1 + \lambda \gamma_\nu} \right)^2 (f_\nu^0)^2 \\
&\leq \lambda^2 \sup_{\nu \in \mathbb{N}} \frac{\gamma_\nu}{(1 + \lambda \gamma_\nu)^2} \sum_{\nu=1}^{\infty} \gamma_\nu (f_\nu^0)^2 \\
&\leq \lambda^2 J(f_0) \sup_{x > 0} \frac{x^{-1}}{(1 + \lambda x^{-1})^2} \\
&\asymp \lambda^2 J(f_0) \lambda^{-1} = \lambda J(f_0),
\end{aligned}$$

where the fourth step is achieved when $x \asymp \lambda$. □

Second, we show an upper bound of $(\tilde{f}^* - \bar{f})$, which accounts for a part of the stochastic error.

Lemma A.10. *For $1 \leq p \leq d$, then if $m > 5/4$, we have*

$$\|\tilde{f}^* - \bar{f}\|_0^2 = O_{\mathbb{P}}\left\{ n^{-1} \lambda^{-1/(2m-2)} \right\}.$$

Proof. As shown in (A.34), $\mathbb{E}[\frac{1}{2} D l_{n,\lambda}(\bar{f}) g]^2 = O\{n^{-1} \|g\|_0^2\}$. By the definition of G_λ^{-1} in (A.29),

$$\|G_\lambda^{-1} g\|_0^2 = \sum_{\nu=1}^{\infty} (1 + \lambda \gamma_\nu)^{-2} \langle g, \xi_\nu \rangle_R^2, \quad \forall g \in \mathcal{H}.$$

Thus,

$$\begin{aligned}\mathbb{E}\|\tilde{f}^* - \bar{f}\|_0^2 &= \frac{1}{4}\mathbb{E}\left[\sum_{\nu=1}^{\infty}(1+\lambda\gamma_\nu)^{-2}\langle Dl_{n\lambda}(\bar{f}), \xi_\nu\rangle_R^2\right] \\ &\leq \sum_{\nu=1}^{\infty}(1+\lambda\gamma_\nu)^{-2}\mathbb{E}\left[\frac{1}{2}Dl_{n\lambda}(\bar{f})\xi_\nu\right]^2 \\ &\lesssim n^{-1}\sum_{\nu=1}^{\infty}(1+\lambda\gamma_\nu)^{-2} \\ &\asymp n^{-1}M_0(\lambda),\end{aligned}$$

where the last step is because of Lemma A.8, and $M_a(\lambda)$ for $0 \leq a \leq 1$ is defined in Lemma A.13 of Section A.5.4. Hence, we complete the proof by using Lemma A.13. □

Then, we give an upper bound of $(\widehat{f} - \tilde{f}^*)$, which accounts for another part of the stochastic error.

Lemma A.11. *If* $n^{-1}\lambda^{-[a+ma/(m-1)+3/2m]}\left[\log(1/\lambda)\right]^{r-1} \to 0$ *and* $1/2m < a < (2m-3)/2m$, *we have*

$$\|\widehat{f} - \tilde{f}^*\|_0^2 = o_{\mathbb{P}}\left\{n^{-1}\lambda^{-1/(2m-2)}\right\}.$$

Proof. Observe that

$$
\begin{aligned}
&\mathbb{E}\|\widehat{f}-\tilde{f}\|_0^2 \asymp \mathbb{E}\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V}}(1+\lambda\gamma_{\overrightarrow{\boldsymbol{\nu}}})^{-2}\left[\frac{1}{2}D^2 l_\infty(\bar{f})(\widehat{f}-\bar{f})\phi_{\overrightarrow{\boldsymbol{\nu}}}-\frac{1}{2}D^2 l_n(\bar{f})(\widehat{f}-\bar{f})\phi_{\overrightarrow{\boldsymbol{\nu}}}\right]^2\\
&\le \mathbb{E}\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V}}(1+\lambda\gamma_{\overrightarrow{\boldsymbol{\nu}}})^{-2}\\
&\times\frac{1}{p+1}\left\{\left[\frac{1}{n\sigma_0^2}\sum_{i=1}^{n}(\widehat{f}-\bar{f})(\mathbf{t}_i^{(0)})\phi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t}_i^{(0)})-\frac{1}{\sigma_0^2}\int(\widehat{f}-\bar{f})(\mathbf{t})\phi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})\Pi^{(0)}(\mathbf{t})\right]^2\right.\\
&\left.+\sum_{j=1}^{p}\left[\frac{1}{n\sigma_j^2}\sum_{i=1}^{n}\frac{\partial(\widehat{f}-\bar{f})}{\partial t_j}(\mathbf{t}_i^{(0)})\frac{\partial\phi_{\overrightarrow{\boldsymbol{\nu}}}}{\partial t_j}(\mathbf{t}_i^{(0)})-\frac{1}{\sigma_j^2}\int\frac{\partial(\widehat{f}-\bar{f})(\mathbf{t})}{\partial t_j}\frac{\partial\phi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})}{\partial t_j}\Pi^{(0)}(\mathbf{t})\right]^2\right\}\\
&\lesssim n^{-1}\|\widehat{f}-\bar{f}\|_{L_2(a+1/m)}^2\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V}}\left(1+\frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a(1+\lambda\rho_{\overrightarrow{\boldsymbol{\nu}}})^{-2}\\
&= n^{-1}\|\widehat{f}-\bar{f}\|_{L_2(a+1/m)}^2 M_a(\lambda)\\
&\le\left\{n^{-1}\lambda^{-[a+3/2m+ma/(m-1)]}[\log(1/\lambda)]^{r-1}\right\}n^{-1}\lambda^{-1/(2m-2)},
\end{aligned}
$$

where the first step exchange the use of $\{\gamma_\nu,\nu\in\mathbb{N}\}$ and $\{\rho_{\overrightarrow{\boldsymbol{\nu}}}:\overrightarrow{\boldsymbol{\nu}}\in\mathbf{V}\}$, the third step is by (A.42), and the last step is Lemma A.4, Lemma A.5 and Lemma A.13 in Section A.5.4. The above inequality holds for any $1/2m<a<(2m-3)/2m$. This completes the proof. □

Last, we combine Lemma A.9, Lemma A.10 and Lemma A.11. By letting $\lambda\asymp n^{-2(m-1)/(2m-1)}$ and $a=1/2m+\epsilon$ for some $\epsilon>0$, then

$$
n^{-1}\lambda^{-(a+3/2m+ma/(m-1))}[\log(1/\lambda)]^{r-1}\to 0
$$

holds as long as $m>2$. Therefore, we conclude that for any $1\le p\le d$ and $m>2$,

$$
\begin{aligned}
\|\widehat{f}-f_0\|_0^2 &= O\left\{\lambda J(f_0)\right\}+O_{\mathbb{P}}\left\{n^{-1}\lambda^{-1/(2m-2)}\right\}+o_{\mathbb{P}}\left\{n^{-1}\lambda^{-1/(2m-2)}\right\}\\
&= O_{\mathbb{P}}\left\{n^{-2(m-1)/(2m-1)}\right\}.
\end{aligned}
$$

This completes the proof for Lemma A.7 and the proof for Theorem 2.7 .

A.5 Key Lemmas

Now we prove and show some keys lemmas used for the proofs in Section A.3, Section A.2 and Section A.4. We remind the reader that the proofs in this section rely on some lemmas to be stated later in Section A.6.

A.5.1 Proof of Lemma A.2

The norm $\|\cdot\|_R$ is equivalent to $\|\cdot\|_{\mathcal{H}}$ in $\mathcal{H}$.

Proof. Observe that for any $g \in \mathcal{H}$, by the assumption that $\Pi^{(0)}$ and $\Pi^{(j)}$s are bounded away from 0 and infinity, we have

$$\begin{aligned}
&\frac{1}{p+1}\left[\frac{1}{\sigma_0^2}\int g^2(\mathbf{t})\Pi^{(0)}(\mathbf{t}) + \sum_{j=1}^{p}\frac{1}{\sigma_j^2}\int\left\{\frac{\partial g(\mathbf{t})}{\partial t_j}\right\}^2\Pi^{(j)}(\mathbf{t})\right] \\
&\leq c_1\left[\int g^2(\mathbf{t}) + \sum_{j=1}^{p}\int\left\{\frac{\partial g(\mathbf{t})}{\partial t_j}\right\}^2\right] \leq c_2\cdot c_K^{2d}\|g\|_{\mathcal{H}}^2,
\end{aligned}$$

for some constant c_1 and c_2, where the last step is by Lemma A.14. Hence

$$\|g\|_R^2 \leq (c_2 c_K^{2d}+1)\|g\|_{\mathcal{H}}^2. \tag{A.45}$$

One the other hand, for any $g \in \mathcal{H}$ we can do the orthogonal decomposition $g = g^0 + g^1$ where $\langle g^0, g^1\rangle_{\mathcal{H}} = 0$, g^0 is in the null space of $J(\cdot)$ and g^1 is in the orthogonal space of the null space of $J(\cdot)$ in $\mathcal{H}$. Since the null space of $J(\cdot)$ has a finite basis which forms a positive definite kernel matrix, we assume the minimal eigenvalue of the kernel matrix is $\mu'_{\min} > 0$. Then there exists a constant $c_3 > 0$ such that

$$\|g^0\|_R^2 \geq c_3\|g^0\|_{L_2}^2 \geq c_3\mu'_{\min}\|g^0\|_{\mathcal{H}}^2. \tag{A.46}$$

For g^1, we have $\|g^1\|_R^2 \geq J(g^1) = \|g^1\|_{\mathcal{H}}^2$. Thus, for any $g \in \mathcal{H}$,

$$\begin{aligned}\|g\|_R^2 &\geq c_3 \int \left(g^0 + g^1\right)^2 + \|g^1\|_{\mathcal{H}}^2 \\ &\geq c_3 \left\{ \|g^0\|_{L_2}^2 + \frac{1+c_3}{c_3}\|g^1\|_{L_2}^2 - 2\|g^0\|_{L_2}\|g^1\|_{L_2} \right\} \\ &\geq \frac{c_3}{1+c_3}\|g^0\|_{L_2}^2,\end{aligned}$$

where the second inequality is by $\|g^1\|_{\mathcal{H}}^2 \geq \|g^1\|_{L_2}^2$. Then by (A.46), we obtain $\|g\|_R^2 \geq (1+c_3)^{-1}c_3\mu'_{\min}\|g^0\|_{\mathcal{H}}^2$. Together with $\|g\|_R^2 \geq J(g^1) = \|g^1\|_{\mathcal{H}}^2$, we have

$$\|g\|_R^2 \geq \left(1 + \frac{1+c_3}{c_3\mu'_{\min}}\right)^{-1} \|g\|_{\mathcal{H}}^2. \tag{A.47}$$

Combining (A.45) and (A.47) completes the proof. □

A.5.2 Proof of Lemma A.8

Proof. When $d = 1$, this problem is solved in Cox (1988). Their method is finding an orthonormal basis in $L_2(\mathcal{X}_1)$ to simultaneously diagonalize $\langle f, f\rangle_0$ and $\langle f, f\rangle_R$, and then obtain the decay rate of γ_ν. However, their method cannot be applied to our case when $2 \leq p \leq d$. Alternatively, we use the Courant-Fischer-Weyl min-max principle to prove the lemma.

Note that for any $f \in \mathcal{H}$, the norm $\|f\|_0^2$ is equivalent to

$$\int f^2 + \sum_{j=1}^{p} \int \left(\frac{\partial f(\mathbf{t})}{\partial t_j}\right)^2.$$

From Lemma A.2, the norm $\|\cdot\|_R^2$ is equivalent to $\|\cdot\|_{\mathcal{H}}^2$. Now by applying the mapping principle [see, e.g., Theorem 3.8.1 in Weinberger (1974)], we may replace $\langle f, f\rangle_0$ by $\int f^2 + \sum_{j=1}^{p}\int(\partial f/\partial t_j)^2$ and $\langle f, f\rangle_R$ by $\|f\|_{\mathcal{H}}^2$, and the resulting eigenvalues $\{\gamma''_\nu\}_{\nu\geq 1}$ of $\{\int f^2 + \sum_{j=1}^{p}\int(\partial f/\partial t_j)^2\}/\|f\|_{\mathcal{H}}^2$ satisfy

$$\gamma''_\nu \asymp (1+\gamma_\nu)^{-1}. \tag{A.48}$$

Thus, we only need to study $\{\gamma''_\nu\}_{\nu\geq 1}$. Since $f \in \mathcal{H}$ has the tensor product structure, we denote by $\lambda_{\vec{\boldsymbol{\nu}}}[\{\int f^2 + \sum_{j=1}^p \int (\partial f/\partial t_j)^2\}/\langle f, f\rangle_{\mathcal{H}}]$ the $\vec{\boldsymbol{\nu}}$th eigenvalue of the generalized Rayleigh quotient, where $\vec{\boldsymbol{\nu}} \in \mathbf{V}$ and V is defined in (A.1).

Second, by the assumption that $\lambda_\nu \asymp \nu^{-2m}$, $\mathcal{H}_1$ is equivalent to a Sobolev space $\mathcal{W}_2^m(\mathcal{X}_1)$ and the trigonometric functions $\{\psi_\nu\}_{\nu\geq 1}$ in (A.2) form an eigenfunction basis of $\mathcal{H}_1$ up to a m-dimensional linear space of polynomials of order less than m. See, for example, Wahba Wahba (1990). Denote the latter linear space of polynomials by $\mathcal{G}$. Denote by $\mathcal{F}_\mu$ and $\mathcal{F}_\mu^\perp$ the linear spaces spanned by $\{\psi_\nu : 1 \leq \nu \leq \mu\}$ and $\{\psi_\nu : \nu \geq \mu + 1\}$, respectively. For any $\vec{\boldsymbol{\nu}} = (\nu_1, \nu_2, \ldots, \nu_d) \in \mathbf{V}$, by the Courant-Fischer-Weyl min-max principle,

$$\begin{aligned}
&\lambda_{(\nu_1-m)\vee 0,(\nu_2-m)\vee 0,\ldots,(\nu_d-m)\vee 0} \left[\left\{\int f^2 + \sum_{j=1}^p \int \left(\frac{\partial f}{\partial t_j}\right)^2\right\} \Big/ \langle f, f\rangle_{\mathcal{H}}\right] \\
&\geq \min_{f\in\mathcal{H}\cap\otimes_{k=1}^d\{\mathcal{F}_{\nu_k}\cap\mathcal{G}^\perp\}} \left[\left\{\int f^2 + \sum_{j=1}^p \int \left(\frac{\partial f}{\partial t_j}\right)^2\right\} \Big/ \langle f, f\rangle_{\mathcal{H}}\right] \\
&\geq c_1 \left(1 + \sum_{j=1}^p \nu_j^2\right) \prod_{k=1}^d \nu_k^{-2m}
\end{aligned}$$

for some constant $c_1 > 0$, where the last inequality is by the fact that $d\psi_{2\nu-1}(t)/dt = 2\Pi\nu\psi_{2\nu}(t)$ and $d\psi_{2\nu}(t)/dt = -2\Pi\nu\psi_{2\nu-1}(t)$. On the other hand,

$$\begin{aligned}
&\lambda_{\nu_1+m,\nu_2+m,\ldots,\nu_d+m} \left[\left\{\int f^2 + \sum_{j=1}^p \int \left(\frac{\partial f}{\partial t_j}\right)^2\right\} \Big/ \langle f, f\rangle_{\mathcal{H}}\right] \\
&\leq \max_{f\in\mathcal{H}\cap\otimes^d\{\mathcal{F}_{k-1}^\perp\cap\mathcal{G}^\perp\}} \left[\left\{\int f^2 + \sum_{j=1}^p \int \left(\frac{\partial f}{\partial t_j}\right)^2\right\} \Big/ \langle f, f\rangle_{\mathcal{H}}\right] \\
&\leq c_2 \left(1 + \sum_{j=1}^p \nu_j^2\right) \prod_{k=1}^d \nu_k^{-2m}
\end{aligned}$$

for some constant $c_2 > 0$. Thus, for any $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$,

$$\lambda_{\overrightarrow{\boldsymbol{\nu}}} \left[\left\{ \int f^2 + \sum_{j=1}^{p} \int \left(\frac{\partial f}{\partial t_j} \right)^2 \right\} \Big/ \langle f, f \rangle_{\mathcal{H}} \right] \asymp \left(1 + \sum_{j=1}^{p} \nu_j^2 \right) \prod_{k=1}^{d} \nu_k^{-2m}.$$

This implies $\gamma'_\nu = \gamma''_\nu$, where γ'_ν is defined in Lemma A.8. Together with (A.48), we complete the proof. □

A.5.3 Definition of $N_a(\lambda)$ and Its Upper Bound

Lemma A.12. *Recall that V as a family of multi-index $\overrightarrow{\boldsymbol{\nu}}$ is defined in (A.1). We let*

$$N_a(\lambda) = \sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} \frac{\left(\prod_{k=1}^{d} \nu_k^{2m} \right)^a \left(1 + \sum_{j=1}^{p} \nu_j^2 \right)}{\left(1 + \sum_{j=1}^{p} \nu_j^2 + \lambda \prod_{k=1}^{d} \nu_k^{2m} \right)^2}. \tag{A.49}$$

Then, when $0 \le p < d$, we have for any $0 \le a < 1 - 1/2m$,

$$N_a(\lambda) = O\left\{ \lambda^{-a-1/2m} \left[\log(1/\lambda) \right]^{(d-p)\wedge r - 1} \right\},$$

and when $p = d$, we have for any $0 \le a \le 1$,

$$N_a(\lambda) = \begin{cases} O\left\{ \lambda^{\frac{mr}{1-mr}\left(a + \frac{r-2}{2mr}\right)} \right\}, \text{ if } r \ge 3; \\ O\left\{ \log(1/\lambda) \right\}, \text{ if } r = 2, a = 0; \quad O\left\{ 1 \right\}, \text{ if } r = 2, 0 < a \le 1; \\ O\left\{ 1 \right\}, \text{ if } r = 1, a < \frac{1}{2m}; \quad O\left\{ \log(1/\lambda) \right\}, \text{ if } r = 1, a = \frac{1}{2m}; \\ O\left\{ \lambda^{\frac{1-2ma}{2m-2}} \right\}, \text{ if } r = 1, a > \frac{1}{2m}. \end{cases}$$

Proof. We will discuss three separate cases for $0 \le p \le d - r$, $d - r < p < d$ and $p = d$.

First, consider $0 \le p \le d - r$. Since $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$, there are at most r of $\nu_1, \ldots, \nu_d$ not equal to 1, which implies that the number of combinations of non-1 indices being summed in (A.49) is no greater than $C_d^1 + C_d^2 + \cdots + C_d^r < \infty$. Due to the appearance of $(1 + \sum_{j=1}^{p} \nu_j^2)$ in the denominator of (A.49), the largest terms of the summation (A.49) over $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$ correspond

to the combinations of r indices where as few $\nu_1, \ldots, \nu_p$ being summed as possible, which is the indices $\overrightarrow{\boldsymbol{\nu}} = (\nu_{k_1}, \nu_{k_2}, \ldots, \nu_{k_r})^\top \in \mathbb{N}^r$ with $k_1, k_2, \ldots, k_r > p$. Thus, by the integral approximation,

$$
\begin{aligned}
& N_a(\lambda) \\
& \asymp \sum_{\nu_{p+1}=1}^{\infty} \cdots \sum_{\nu_{p+r-1}=1}^{\infty} \sum_{\nu_{p+r}=1}^{\infty} \frac{\prod_{k=p+1}^{p+r} \nu_k^{2ma}}{\left(1 + \lambda \prod_{k=p+1}^{p+r} \nu_k^{2m}\right)^2} \\
& \asymp \int_1^{\infty} \int_1^{\infty} \cdots \int_1^{\infty} \left(1 + \lambda x_{p+1}^b \cdots x_{p+r-1}^b x_{p+r}^b\right)^{-2} dx_{p+1} \cdots dx_{p+r-1} dx_{p+r},
\end{aligned}
$$

where $b = 2m/(2ma+1)$. Let $z_k = x_{p+1} x_{p+2} \cdots x_k$ for $k = p+1, \ldots, p+r$. By using the change of variables to replace $(x_{p+1}, \ldots, x_{p+r})$ by $(z_{p+1}, \ldots, z_{p+r})$ and z_{p+r} by $x = \lambda^{1/b} z_{p+r}$,

$$
\begin{aligned}
& N_a(\lambda) \\
& \asymp \int_1^{\infty} \int_1^{z_{p+r}} \cdots \int_1^{z_{p+2}} \left(1 + \lambda z_{p+r}^b\right)^{-2} z_{p+1}^{-1} \cdots z_{p+r-1}^{-1} dz_{p+1} \cdots dz_{p+r-1} dz_{p+r} \\
& \asymp \int_1^{\infty} (1 + \lambda z_{p+r}^b)^{-2} (\log z_{p+r})^{r-1} dz_{p+r} \\
& \asymp \lambda^{-1/b} \int_{\lambda^{1/b}}^{\infty} (1 + x^b)^{-2} \left(\log x - b^{-1} \log \lambda\right)^{r-1} dx \\
& \asymp \lambda^{-a-1/2m} \left[\log(1/\lambda)\right]^{r-1},
\end{aligned}
$$

where the last step follows from the fact that $2b > 1$ for any $0 \leq a < (2m-1)/(2m)$.

Second, we consider $d - r < p < d$. As discussed in the previous case, the number of combinations of non-1 indices being summed is finite, and the largest terms of the summation (A.49) over $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$ correspond to the indices $\overrightarrow{\boldsymbol{\nu}} = (\nu_{k_1}, \ldots, \nu_{k_{r+p-d}}, \nu_{p+1}, \ldots, \nu_d)^\top \in \mathbb{N}^r$,

where the indices $k_1, \ldots, k_{r+p-d} \le p$. Thus, by the integral approximation,

$$
\begin{aligned}
& N_a(\lambda) \\
& \asymp \sum_{v_{d-r+1}=1}^{\infty} \cdots \sum_{v_d=1}^{\infty} \frac{\prod_{k=d-r+1}^{d} \nu_k^{2ma} \left(1 + \sum_{k=d-r+1}^{p} \nu_k^2\right)}{\left(1 + \sum_{k=d-r+1}^{p} \nu_k^2 + \lambda \prod_{k=d-r+1}^{d} \nu_k^{2m}\right)^2} \\
& \asymp \int_1^{\infty} \cdots \int_1^{\infty} \frac{1 + x_{d-r+1}^{b/m} + \cdots + x_p^{b/m}}{\left(1 + x_{d-r+1}^{b/m} + \cdots + x_p^{b/m} + \lambda x_{d-r+1}^{b} \cdots x_d^{b}\right)^2} dx_{d-r+1} \cdots dx_d,
\end{aligned}
$$

where $b = 2m/(2ma+1)$. Set $z_k = x_{p+1} x_{p+2} \cdots x_k$ for $k = p+1, \ldots, d$. By using the change the variables to replace $(x_{p+1}, \ldots, x_d)$ by $(z_{p+1}, \ldots, z_d)$, and z_d by $x = \lambda^{1/b} z_d$, and x by $u = x_{d-r+1} \cdots x_p \cdot x$. We have

$$
\begin{aligned}
N_a(\lambda) \asymp & \int_1^{\infty} \cdots \int_1^{\infty} \left[\int_1^{\infty} \int_1^{z_d} \cdots \int_1^{z_{p+2}} \right. \\
& \qquad x_{d-r+1}^{b/m} \left(1 + x_{d-r+1}^{b/m} + \cdots x_p^{b/m} + \lambda x_{d-r+1}^{b} \cdots x_p^{b} z_d^{b}\right)^{-2} \\
& \qquad \left. \cdot z_{p+1}^{-1} \cdots z_{d-1}^{-1} dz_{p+1} \cdots dz_{d-1} dz_d \right] dx_{d-r+1} \cdots dx_p \\
\asymp \lambda^{-1/b} & \int_1^{\infty} \cdots \int_1^{\infty} \left[\int_{\lambda^{1/b}}^{\infty} \right. \\
& \qquad x_{d-r+1}^{b/m} (1 + x_{d-r+1}^{b/m} + \cdots x_p^{b/m} + x_{d-r+1}^{b} \cdots x_p^{b} x^{b})^{-2} \\
& \qquad \left. \cdot \left(\log x - b^{-1} \log \lambda\right)^{d-p-1} dx \right] dx_{d-r+1} \cdots dx_p \\
\lesssim \lambda^{-1/b} & \int_{\lambda^{1/b}}^{\infty} \left[\int_1^{\infty} \cdots \int_1^{\infty} \right. \\
& \qquad x_{d-r+1}^{b/m} \left(1 + x_{d-r+1}^{b/m} + \cdots + x_p^{b/m} + u^{b}\right)^{-2} x_{d-r+1}^{-1} \cdots x_p^{-1} \\
& \left. \cdot \left(\log u - \log x_{d-r+1} - \cdots - \log x_p - b^{-1} \log \lambda\right)^{d-p-1} dx_{d-r+1} \cdots dx_p \right] du.
\end{aligned}
$$

By Lemma A.15, then for any $0 < \tau < 1$,

$$
\begin{aligned}
& \left(1 + x_{d-r+1}^{b/m} + x_{d-r+2}^{b/m} + \cdots + x_p^{b/m} + u^{b}\right)^{-2} \\
& \qquad \lesssim \left(1 + x_{d-r+2}^{b/m} + \cdots + x_p^{b/m} + u^{b}\right)^{-1+\tau} \cdot \left(x_{d-r+1}^{b/m}\right)^{-(1+\tau)}.
\end{aligned}
$$

Together with the fact $\int_1^\infty t^{-1-\tau}(\log t)^k dt < \infty$ for any $k < \infty$, we have

$$N_a(\lambda) \lesssim \lambda^{-1/b} \int_{\lambda^{1/b}}^\infty \Bigg[\int_1^\infty \cdots \int_1^\infty \Big(1 + x_{d-r+2}^{b/m} + \cdots + x_p^{b/m} + u^b\Big)^{-1+\tau} x_{d-r+2}^{-1} \cdots x_p^{-1} \cdot \left(\log u - \log x_{d-r+2} - \cdots - \log x_p - b^{-1} \log \lambda\right)^{d-p-1} dx_{d-r+2} \cdots dx_p \Bigg] du.$$

Continuing this procedure gives

$$N_a(\lambda) \lesssim \lambda^{-1/b} \int_{\lambda^{1/b}}^\infty \Big(1 + u^b\Big)^{-(1-\tau)^{p-d+r}} \left(\log u - b^{-1} \log \lambda\right)^{d-p-1} du.$$

Since for any $\epsilon > 0$ and $d - r < p < d$, we know if $\tau < \epsilon/d$,

$$(1-\tau)^{p-d+r} \geq 1 - \tau(p-d+r) \geq 1 - \tau(d-1) > 1 - \epsilon.$$

Hence, for any $0 \leq a < (2m-1)/(2m)$, there exists τ such that $(1-\tau)^{p-d+r} > a + 1/(2m) = 1/b$. Therefore,

$$N_a(\lambda) \lesssim \lambda^{-1/b} \left[\log(1/\lambda)\right]^{d-p-1} = \lambda^{-a-1/2m} \left[\log(1/\lambda)\right]^{d-p-1}.$$

Finally, we consider $p = d$. As argued in the previous two cases, the number of combinations of non-1 indices being summed is finite. Now since $p = d$, by the symmetry of indices, the largest terms of the summation (A.49) over $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$ correspond to any combinations of r

non-1 indices, for example, the first r indices. Thus, by the integral approximation,

$$\begin{aligned} & N_a(\lambda) \\ & \asymp \sum_{\nu_1=1}^{\infty} \cdots \sum_{\nu_{r-1}=1}^{\infty} \sum_{\nu_r=1}^{\infty} \frac{\prod_{k=1}^{r} \nu_k^{2ma} \left(1+\sum_{k=1}^{r} \nu_k^2\right)}{\left(1+\sum_{k=1}^{r} \nu_k^2 + \lambda \prod_{k=1}^{r} \nu_k^{2m}\right)^2} \\ & \asymp \int_1^{\infty} \int_1^{\infty} \cdots \int_1^{\infty} \frac{1+x_1^{b/m}+\cdots+x_{r-1}^{b/m}+x_r^{b/m}}{\left(1+x_1^{b/m}+\cdots+x_r^{b/m}+\lambda x_1^b \cdots x_{r-1}^b x_r^b\right)^2} \\ & \qquad\qquad dx_1 \cdots dx_{r-1} dx_r \end{aligned}$$

where $b = 2m/(2ma+1)$. Observe that if $x_1 \cdots x_{r-1} x_r \lesssim \lambda^{mr/[b(1-mr)]}$, then

$$\lambda x_1^b \cdots x_{r-1}^b x_r^b \lesssim x_1^{b/m} + \cdots + x_{r-1}^{b/m} + x_r^{b/m}.$$

By Lemma A.19 with $\beta = 0$ and $\alpha = b/m \leq 2$, we have

$$\begin{aligned} N_a(\lambda) & \asymp \int_{x_1 \cdots x_{r-1} x_r \lesssim \lambda^{mr/[b(1-mr)]}} \left(1+x_1^{b/m}+\cdots+x_{r-1}^{b/m}+x_r^{b/m}\right)^{-1} dx_1 \cdots dx_{r-1} dx_r \\ & \asymp \begin{cases} \lambda^{\frac{mr}{1-mr}\left(a+\frac{r-2}{2mr}\right)}, \text{ if } r \geq 3; \\ \log(1/\lambda), \text{ if } r=2, a=0; \quad \lambda^{\frac{2ma}{1-2m}}, \text{ if } r=2, 0<a\leq 1; \\ 1, \text{ if } r=1, a<\frac{1}{2m}; \quad \log(1/\lambda), \text{ if } r=1, a=\frac{1}{2m}; \\ \lambda^{\frac{1-2ma}{2m-2}}, \text{ if } r=1, a>\frac{1}{2m}. \end{cases} \end{aligned} \tag{A.50}$$

On the other hand, if $\lambda^{mr/[b(1-mr)]}(x_1 \cdots x_{r-1} x_r)^{-1} = o(1)$, without less of generality, we

assume $x_r = \min\{x_1, \cdots, x_r\}$. Let $z = \lambda^{1/b} x_1 \cdots x_{r-1} x_r$. By changing x_r to z, we have

$$
\begin{aligned}
N_a(\lambda) &\asymp \int_{\lambda^{mr/[b(1-mr)]}(x_1\cdots x_{r-1}x_r)^{-1}=o(1)} \\
&\qquad \left(1 + x_1^{b/m} + \cdots + x_r^{b/m} + \lambda x_1^b \cdots x_{r-1}^b x_r^b\right)^{-1} dx_1 \cdots dx_{r-1} dx_r \\
&\lesssim \lambda^{-1/b} \int_{\lambda^{1/[b(1-mr)]} z^{-1}=o(1), \lambda^{-(r-1)/(br)} z^{(r-1)/r} \le x_1 \cdots x_{r-1} \le \lambda^{-1/b} z} \\
&\qquad \left(1 + x_1^{b/m} + \cdots + x_{r-1}^{b/m} + z^b\right)^{-1} x_1^{-1} \cdots x_{r-1}^{-1} dx_1 \cdots dx_{r-1} dz \\
&\lesssim \lambda^{-1/b} \int_{\lambda^{1/[b(1-mr)]} z^{-1}=o(1)} \left[\int_{\lambda^{-(r-1)/(br)} z^{(r-1)/r} \le x_1 \cdots x_{r-1} \le \lambda^{-1/b} z} \right. \\
&\qquad \left. \left(x_1^{b/m} + \cdots + x_{r-1}^{b/m}\right)^{-\tau} x_1^{-1} \cdots x_{r-1}^{-1} dx_1 \cdots dx_{r-1} \right] z^{b(-1+\tau)} dz \\
&\lesssim \lambda^{-1/b} \int_{\lambda^{1/[b(1-mr)]} z^{-1}=o(1)} \lambda^{\tau/(mr)} z^{-\tau b/(mr)} \cdot z^{b(-1+\tau)} dz \\
&= o\left[\lambda^{\frac{mr}{1-mr}\left(a + \frac{r-2}{2mr}\right)}\right],
\end{aligned}
\tag{A.51}
$$

where the third step follows from the Lemma A.20 in Section A.6 for $\beta = -1$ and $\alpha = \tau b/m$. Combining (A.50) and (A.51), we complete the proof for $p = d$ and this lemma. □

A.5.4 Definition of $M_a(\lambda)$ and Its Upper Bound

Lemma A.13. *Recall that V as a family of multi-index $\overrightarrow{\boldsymbol{\nu}}$ is defined in (A.1). We let*

$$
M_a(\lambda) = \sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}} \frac{\left(\prod_{k=1}^d \nu_k^{2m}\right)^a}{\left[1 + \lambda \prod_{k=1}^d \nu_k^{2m} (1 + \sum_{j=1}^p \nu_j^2)^{-1}\right]^2}.
$$

When $m > 5/(4-2a)$, we have for any $1 \le p \le d$ and $0 \le a \le 1$,

$$
M_a(\lambda) = O\left\{\lambda^{-(2ma+1)/(2m-2)}\right\}.
$$

Proof. We first show for any $1 \le s \le r$,

$$\begin{aligned}\sum_{\nu_1=1}^{\infty}\cdots\sum_{\nu_r=1}^{\infty}\frac{\prod_{k=1}^{r}\nu_k^{2ma}}{\left[1+\lambda\prod_{k=1}^{r}\nu_k^{2m}(1+\sum_{j=1}^{s}\nu_j^2)^{-1}\right]^2}\\ \asymp\sum_{\nu_1=1}^{\infty}\cdots\sum_{\nu_r=1}^{\infty}\frac{\prod_{k=1}^{r}\nu_k^{2ma}}{\left[1+\lambda\prod_{k=1}^{r}\nu_k^{2m}(1+\nu_s^2)^{-1}\right]^2}.\end{aligned} \tag{A.52}$$

Note that in (A.52), the left-hand side is greater than the right-hand side up to some constant. On the contrary, observe that

$$\begin{aligned}&\sum_{\nu_1=1}^{\infty}\cdots\sum_{\nu_r=1}^{\infty}\frac{\prod_{k=1}^{r}\nu_k^{2ma}}{\left[1+\lambda\prod_{k=1}^{r}\nu_k^{2m}(1+\sum_{j=1}^{s}\nu_j^2)^{-1}\right]^2}\\ &\quad\asymp\sum_{\nu_1=1}^{\infty}\cdots\sum_{\nu_r=1}^{\infty}\sum_{i=1}^{s}\frac{(1+\nu_i^2)^2\prod_{k=1}^{r}\nu_k^{2ma}}{\left(1+\sum_{j=1}^{s}\nu_j^2+\lambda\prod_{k=1}^{r}\nu_k^{2m}\right)^2}\\ &\quad\asymp\sum_{\nu_1=1}^{\infty}\cdots\sum_{\nu_r=1}^{\infty}\frac{(1+\nu_s^2)^2\prod_{k=1}^{r}\nu_k^{2ma}}{\left(1+\sum_{j=1}^{s}\nu_j^2+\lambda\prod_{k=1}^{r}\nu_k^{2m}\right)^2}\\ &\quad\le\sum_{\nu_1=1}^{\infty}\cdots\sum_{\nu_r=1}^{\infty}\frac{\prod_{k=1}^{r}\nu_k^{2ma}}{\left[1+\lambda\prod_{k=1}^{r}\nu_k^{2m}(1+\nu_s^2)^{-1}\right]^2}.\end{aligned}$$

This proves (A.52). Moreover, note that

$$\begin{aligned}\sum_{\nu_1=1}^{\infty}\cdots\sum_{\nu_r=1}^{\infty}\frac{\prod_{k=1}^{r}\nu_k^{2ma}}{\left[1+\lambda\prod_{k=1}^{r}\nu_k^{2m}(1+\nu_s^2)^{-1}\right]^2}\\ \ge\sum_{\nu_1=1}^{\infty}\cdots\sum_{\nu_r=1}^{\infty}\frac{\prod_{k=1}^{r}\nu_k^{2ma}}{\left(1+\lambda\prod_{k=1}^{r}\nu_k^{2m}\right)^2}.\end{aligned} \tag{A.53}$$

Now return to the proof of the lemma. Since $\overrightarrow{\boldsymbol{\nu}} \in \mathbf{V}$ and $1 \le p \le d$, by (A.52), (A.53) and the integral approximation, we have

$$\begin{aligned}M_a(\lambda)&\asymp\sum_{\nu_1=1}^{\infty}\cdots\sum_{\nu_r=1}^{\infty}\frac{\prod_{k=1}^{r}\nu_k^{2ma}}{\left[1+\lambda\prod_{k=1}^{r}\nu_k^{2m}(1+\nu_r^2)^{-1}\right]^2}\\ &\asymp\int_1^{\infty}\int_1^{\infty}\cdots\int_1^{\infty}\left[1+\lambda x_1^b\cdots x_{r-1}^b x_r^{b(m-1)/m}\right]^{-2}dx_1\cdots dx_{r-1}dx_r,\end{aligned}$$

where $b = 2m/(2ma+1)$. Let $z = \lambda^{m/[b(m-1)]} x_1^{m/(m-1)} \cdots x_{r-1}^{m/(m-1)} x_r$ and change x_r to z. Then,

$$
\begin{aligned}
& M_a(\lambda) \\
& \asymp \lambda^{-m/[b(m-1)]} \int_{\lambda^{-m/[b(m-1)]}}^{\infty} \int_1^{\infty} \cdots \int_1^{\infty} \\
& \qquad \left[1 + z^{b(m-1)/m}\right]^{-2} x_1^{-m/(m-1)} \cdots x_{d-1}^{-m/(m-1)} dx_1 \cdots dx_{d-1} dz \\
& \asymp \lambda^{-m/[b(m-1)]} \int_{\lambda^{-m/[b(m-1)]}}^{\infty} \left[1 + z^{b(m-1)/m}\right]^{-2} dz, \\
& \leq \lambda^{-m/[b(m-1)]} \int_0^{\infty} \left[1 + z^{b(m-1)/m}\right]^{-2} dz \\
& = O\left\{\lambda^{-(2ma+1)/(2m-2)}\right\},
\end{aligned}
$$

where the second step is because $m/(m-1) > 1$ and the last step holds for any $m > 5/(4-2a)$. □

A.5.5 Boundedness of Functions in the RKHS $\mathcal{H}$

Lemma A.14. *For any $g \in \mathcal{H}$, there exists a constant c_K which is independent of g such that*

$$\sup_{\mathbf{t} \in \mathcal{X}_1^d} |g(\mathbf{t})| \leq c_K^d \|g\|_{\mathcal{H}},$$

and

$$\sup_{\mathbf{t} \in \mathcal{X}_1^d} |\partial g / \partial t_j(\mathbf{t})| \leq c_K^d \|g\|_{\mathcal{H}}, \quad \forall 1 \leq j \leq d.$$

Proof. Since we assume that K is continuous in the compact domain $\mathcal{X}_1$ and satisfies (2.9), there exists some constant c_K such that

$$\sup_{t \in \mathcal{X}_1} |K(t,t)| \leq c_K \quad \text{and} \quad \sup_{t \in \mathcal{X}_1} \left| \frac{\partial^2 K(t,t)}{\partial t \partial t'} \right| \leq c_K.$$

This implies for any $\mathbf{t} \in \mathcal{X}_1^d$,

$$\left\| \frac{\partial K_d(\mathbf{t}, \cdot)}{\partial t_j} \right\|_{\mathcal{H}}^2 = \left| \frac{\partial^2 K(t_j, t_j)}{\partial t_j \partial t_j'} \right| \prod_{l \neq j} |K(t_l, t_l)| \leq c_K^d.$$

Thus, for any $g \in \mathcal{H}$, by the Cauchy-Schwarz inequality,

$$\sup_{\mathbf{t} \in \mathcal{X}_1^d} \left| \frac{\partial g(\mathbf{t})}{\partial t_j} \right| \leq \sup_{\mathbf{t} \in \mathcal{X}_1^d} \left\| \frac{\partial K_d(\mathbf{t}, \cdot)}{\partial t_j} \right\|_{\mathcal{H}} \|g\|_{\mathcal{H}} \leq c_K^d \|g\|_{\mathcal{H}}, \quad \forall 1 \leq j \leq d.$$

Similarly, we can show that $\sup_{\mathbf{t}} |g(\mathbf{t})| \leq c_K^d \|g\|_{\mathcal{H}}$.

□

A.6 Auxiliary Technical Lemmas

Lemma A.15 (A variant of Young's inequality)**.** *For any $a, b \geq 0$ and $0 < \tau < 1$, we have*

$$(a+b)^{-2} \leq \frac{(1-\tau)^{1-\tau}(1+\tau)^{1+\tau}}{4} a^{-(1+\tau)} b^{-(1-\tau)}. \tag{A.54}$$

When τ is small, the coefficient $(1-\tau)^{1-\tau}(1+\tau)^{1+\tau}/4$ is close to $1/4$.

Proof. To prove (A.54), it is sufficient to show

$$a + b \geq 2(1-\tau)^{-(1-\tau)/2}(1+\tau)^{-(1+\tau)/2} a^{(1+\tau)/2} b^{(1-\tau)/2}.$$

Letting $p = 2/(1+\tau)$, $a' = a^{1/p}$, $b' = [b/(p-1)]^{(p-1)/p}$, the above formula is equivalent to

$$\frac{a'}{p} + \frac{(b')^{p/(p-1)}}{p/(p-1)} \geq a'b',$$

which holds by Young's inequality. This completes the proof. □

Lemma A.16 (Bounding the norm of product of functions)**.** *For any $f, g \in \otimes^d \mathcal{H}_1$, $a > 1/2m$,*

and $1 \le p \le d$*, we have that*

$$\sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a \|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2 \left\langle \frac{\partial f(\mathbf{t})}{\partial t_j} \frac{\partial g(\mathbf{t})}{\partial t_j}, \phi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t}) \right\rangle_0^2$$
$$\lesssim \|f\|_{L_2(a+1/m)}^2 \left[\sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a \|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2 \left\langle \frac{\partial g(\mathbf{t})}{\partial t_j}, \phi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t}) \right\rangle_0^2 \right].$$

Proof. Recall that $\{\psi_\nu(t)\}_{\nu \ge 1}$ is the trigonometrical basis on $L_2(\mathcal{X}_1)$ and $\phi_{\overrightarrow{\boldsymbol{\nu}}}(\cdot)$ is defined in (A.21). Write $\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t}) = \psi_{\nu_1}(t_1)\psi_{\nu_2}(t_2)\cdots\psi_{\nu_d}(t_d)$. Note that

$$\sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a \|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2 \langle f, \phi_{\overrightarrow{\boldsymbol{\nu}}} \rangle_0^2 = \sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a \left(\int_{\mathcal{X}_1^d} f \psi_{\overrightarrow{\boldsymbol{\nu}}}\right)^2.$$

By Theorem A.2.2 and Corollary A.2.1 in Lin (1998), if $a > 1/2m$, then for any $f, g \in \otimes^d \mathcal{H}_1$,

$$\sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a \left(\int_{\mathcal{X}_1^d} f g \psi_{\overrightarrow{\boldsymbol{\nu}}}\right)^2$$
$$\lesssim \left[\sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a \left(\int_{\mathcal{X}_1^d} f \psi_{\overrightarrow{\boldsymbol{\nu}}}\right)^2 \right] \left[\sum_{\overrightarrow{\boldsymbol{\nu}} \in \mathbb{N}^d} \left(1 + \frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a \left(\int_{\mathcal{X}_1^d} g \psi_{\overrightarrow{\boldsymbol{\nu}}}\right)^2 \right].$$

Thus,

$$
\begin{aligned}
&\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbb{N}^d}\left(1+\frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2\left\langle\frac{\partial f(\mathbf{t})}{\partial t_j}\frac{\partial g(\mathbf{t})}{\partial t_j},\phi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})\right\rangle_0^2\\
&=\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbb{N}^d}\left(1+\frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a\left(\int_{\mathcal{X}_1^d}\frac{\partial f(\mathbf{t})}{\partial t_j}\frac{\partial g(\mathbf{t})}{\partial t_j}\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})\right)^2\\
&\lesssim\left[\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbb{N}^d}\nu_j^2\left(1+\prod_{k=1}^d\nu_k^{2m}\right)^a\left(\int_{\mathcal{X}_1^d}f(\mathbf{t})\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})\right)^2\right]\\
&\qquad\qquad\times\left[\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbb{N}^d}\left(1+\frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a\left(\int_{\mathcal{X}_1^d}\frac{\partial g(\mathbf{t})}{\partial t_j}\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})\right)^2\right]\\
&\leq\left\{\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbb{N}^d}\left[1+\prod_{k=1}^d\nu_k^{2m}\right]^{a+\frac{1}{m}}\left(\int_{\mathcal{X}_1^d}f(\mathbf{t})\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})\right)^2\right\}\\
&\qquad\qquad\times\left[\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbb{N}^d}\left(1+\frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a\left(\int_{\mathcal{X}_1^d}\frac{\partial g(\mathbf{t})}{\partial t_j}\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})\right)^2\right]\\
&\asymp\|f\|_{L_2(a+1/m)}^2\left[\sum_{\overrightarrow{\boldsymbol{\nu}}\in\mathbb{N}^d}\left(1+\frac{\rho_{\overrightarrow{\boldsymbol{\nu}}}}{\|\phi_{\overrightarrow{\boldsymbol{\nu}}}\|_{L_2}^2}\right)^a\left(\int_{\mathcal{X}_1^d}\frac{\partial g(\mathbf{t})}{\partial t_j}\psi_{\overrightarrow{\boldsymbol{\nu}}}(\mathbf{t})\right)^2\right].
\end{aligned}
$$

This completes the proof. □

Lemma A.17 (Inverse transformation)**.** *Assume that design points* $\mathbf{t}^{(j)}$*s have known distribution* $\Pi^{(j)}$*s which are supported on* $\mathcal{X}_1^d$*. Then, there exists a linear transformation to data* $(\mathbf{t}^{(j)},Y^{(j)})$ *such that transformed design points* $\mathbf{x}^{(j)}$*s are independently uniformly distributed on* $\mathcal{X}_1^d$ *and the transformed responses* $Z^{(j)}$*s are the* j*th first-order partial derivative data of some function.*

Proof. As remarked after (2.12), the design under our consideration has the following structure: different types design points can be grouped to some sets, where within the sets different types design points are drawn identically and across the sets the design points are drawn independently. We give the proof for two cases as follows for illustration.

First, we consider that function observations and partial derivatives data share a common design, i.e., $\mathbf{t}_i^{(j)}=\mathbf{t}_i^{(k)},\forall 1\leq i\leq n,0\leq j<k\leq p$. Write $\mathbf{t}^{(j)}=(t_1^{(j)},\ldots,t_d^{(j)})\in\mathcal{X}_1^d$. We

allow covariates of $\mathbf{t}^{(j)}$ can be correlated, that is the density of $\mathbf{t}^{(j)}$ is decomposed as:

$$d\Pi^{(j)}(t_1,\ldots,t_d) = d\Pi_d^{(j)}(t_d)d\Pi_{d-1}^{(j)}(t_{d-1}|t_d)\cdots d\Pi_1^{(j)}(t_1|t_d,t_{d-1},\ldots,t_2).$$

Now let

$$x_d^{(j)} = \Pi_d^{(j)}(t_d^{(j)}), \quad x_{d-1}^{(j)} = \Pi_{d-1}^{(j)}(t_{d-1}^{(j)}|t_d^{(j)}),\ldots,$$
$$x_1^{(j)} = \Pi_1^{(j)}(t_1^{(j)}|t_d^{(j)},t_{d-1}^{(j)}\ldots,t_2^{(j)}).$$

Then, $\mathbf{x}^{(j)} = (x_1^{(j)}, x_2^{(j)},\ldots,x_d^{(j)})$ is uniformly distributed on $\mathcal{X}_1^d$. Define that

$$\begin{aligned} &h(x_1,x_2,\ldots,x_d) \\ &= f(\{\Pi_1^{(j)}\}^{-1}(x_1|x_d,\ldots,x_2),\{\Pi_2^{(j)}\}^{-1}(x_2|x_d,\ldots,x_3),\ldots,\{\Pi_d^{(j)}\}^{-1}(x_d)). \end{aligned}$$

Thus,

$$\frac{\partial h(\mathbf{x})}{\partial x_j} = \sum_{k=1}^{j}\frac{\partial f(\mathbf{t})}{\partial t_k}\cdot\frac{\partial t_k}{\partial x_j} = \sum_{k=1}^{j-1}\frac{\partial f}{\partial t_k}\cdot\frac{\partial t_k}{\partial x_j} + \frac{\partial f}{\partial t_j}\cdot\frac{1}{d\Pi_j^{(j)}(t_j|t_d,\ldots,t_{j+1})}.$$

With the design $\mathbf{x}^{(j)}$ defined, we transform the responses $Y^{(j)}$s to $Z^{(j)}$s by letting $Z^{(0)} = Y^{(0)}$ and for any $j = 1,\ldots,p$,

$$Z^{(j)} = \sum_{k=1}^{j-1} Y^{(k)}\frac{\partial t_k^{(j)}(x_d^{(j)},x_{d-1}^{(j)}\ldots,x_k^{(j)})}{\partial x_j} + \frac{Y^{(j)}}{d\Pi_j^{(j)}(t_j^{(j)}|t_d^{(j)},\ldots,t_{j+1}^{(j)})}.$$

Write

$$\tilde{\sigma}_j^2 = \sum_{k=1}^{j-1}\sigma_k^2\left[\frac{\partial t_k^{(j)}}{\partial x_j}(x_d^{(j)},x_{d-1}^{(j)},\ldots,x_k^{(j)})\right]^2 + \frac{\sigma_j^2}{[d\Pi_j^{(j)}(t_j^{(j)}|t_d^{(j)},\ldots,t_{j+1}^{(j)})]^2}.$$

Then, it is clear that $Z^{(j)} = \partial h/\partial x_j(\mathbf{x}^{(j)}) + \widetilde{\epsilon^{(j)}}$, where the errors $\widetilde{\epsilon^{(j)}}$s are independent centered noises with variance $\tilde{\sigma}_j^2$s.

Second, we consider that not all types of function observations and partial derivatives data share a common design, i.e., $\exists 0 \le j \ne k \le p$ and $1 \le i \le n$ such that $\mathbf{t}_i^{(j)} \ne \mathbf{t}_i^{(k)}$. We require the c(j)ovariates of each $\mathbf{t}^{(j)}$ are independent, that is the density of $\mathbf{t}^{(j)}$ can be

decomposed as:

$$d\Pi^{(j)}(t_1,\ldots,t_d) = d\Pi_1^{(j)}(t_1)d\Pi_2^{(j)}(t_2)\cdots d\Pi_d^{(j)}(t_d)$$

Now let

$$x_1^{(j)} = \Pi_1^{(j)}(t_1^{(j)}), \quad x_2^{(j)} = \Pi_2^{(j)}(t_2^{(j)}), \quad \ldots, \quad x_d^{(j)} = \Pi_d^{(j)}(t_d^{(j)}).$$

Then $\mathbf{x}^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_d^{(j)})$ is uniformly distributed on $\mathcal{X}_1^d$. Define the function

$$h(x_1,\ldots,x_d) = f(\{\Pi_1^{(j)}\}^{-1}(x_1), \{\Pi_2^{(j)}\}^{-1}(x_2), \ldots, \{\Pi_d^{(j)}\}^{-1}(x_d)).$$

Thus, we have

$$\frac{\partial h(\mathbf{x})}{\partial x_j} = \frac{\partial f(\mathbf{t})}{\partial t_j}\cdot\frac{\partial t_j(x_j)}{\partial x_j} = \frac{\partial f(\mathbf{t})}{\partial t_j}\cdot\frac{1}{d\Pi_j^{(j)}(t_j)}.$$

Correspondingly, the responses $Y^{(j)}$ is transformed to $Z^{(j)}$, $0 \le j \le p$, by letting $Z^{(0)} = Y^{(0)}$ and $Z^{(j)} = Y^{(j)}/d\Pi_j^{(j)}(t_j^{(j)})$ for $1 \le j \le d$, and write the transformed variance $\tilde{\sigma}_j^2 = \sigma_j^2/[d\Pi_j^{(j)}(t_j^{(j)})]^2$. □

Lemma A.18. *Suppose that $s \ge 1$, $\beta \ge 0$ and $\beta \ne 1$, and $r \ge 1$. Then*

$$\int_{x_1\cdots x_r\cdot z\le\Xi, x_k\ge 1, z\ge 1} x_1^\beta\cdots x_r^\beta z^\beta(\log z)^s(x_1^2+\cdots+x_r^2)^{-1}dx_1\cdots dx_r dz$$
$$\asymp \Xi^{\beta+1}(\log\Xi)^s, \quad \text{as } \Xi\to\infty.$$

Proof. For any $\tau \ge 1$, we have $\{1 \le z \le \Xi\tau^{-r}, 1 \le x_k \le \tau, k = 1,\ldots,r\} \subset \{x_1\cdots x_r\cdot z \le \Xi, z \ge 1, x_k \ge 1, k = 1,\ldots,r\}$. Thus, if $\Xi\to\infty$,

$$\begin{aligned}
&\int_{x_1\cdots x_r\cdot z\le\Xi, x_k\ge 1, z\ge 1} x_1^\beta\cdots x_r^\beta z^\beta(\log z)^s(x_1^2+\cdots+x_r^2)^{-1}dx_1\cdots dx_r dz\\
&\ge \int_1^{\Xi\tau^{-r}}\int_1^\tau\cdots\int_1^\tau z^\beta(\log z)^s x_1^{\beta-2}\cdots x_r^{\beta-2}dx_1\cdots dx_r dz\\
&\asymp \Xi^{\beta+1}\tau^{-r(\beta+1)}(\log\Xi - r\log\tau)^s\tau^{r(\beta-1)}.
\end{aligned}$$

Let $\tau\to 1$, we have $\int_{x_1\cdots x_r\cdot z\le\Xi, x_k\ge 1, z\ge 1}(\log z)^s(x_1^2+\cdots+x_r^2)^{-1}dx_1\cdots dx_r dz \gtrsim \Xi^{\beta+1}(\log\Xi)^s$.

On the other hand, define $u = x_1 \cdots x_r \cdot z$ and change the variable z to u. We have that as $\Xi \to \infty$,

$$\begin{aligned}
&\int_{x_1 \cdots x_r \cdot z \le \Xi, x_k \ge 1, z \ge 1} x_1^{\beta} \cdots x_r^{\beta} z^{\beta} (\log z)^s (x_1^2 + \cdots + x_r^2)^{-1} dx_1 \cdots dx_r dz \\
&= \int_1^{\Xi} \int_1^{u} \int_1^{u/x_r} \cdots \int_1^{u/(x_r x_{r-1} \cdots x_2)} u^{\beta} (\log u - \log x_r - \cdots - \log x_1)^s \\
&\qquad \cdot \left(x_1^2 + \cdots + x_{r-1}^2 + x_r^2\right)^{-1} x_1^{-1} \cdots x_{r-1}^{-1} x_r^{-1} dx_1 \cdots dx_{r-1} dx_r du \\
&\lesssim \int_1^{\Xi} \int_1^{u} \int_1^{u/x_r} \cdots \int_1^{u/(x_r x_{r-1} \cdots x_2)} u^{\beta} (\log u - \log x_r - \cdots - \log x_1)^s \\
&\qquad \cdot x_1^{-1-2/r} \cdots x_{r-1}^{-1-2/r} x_r^{-1-2/r} dx_1 \cdots dx_{r-1} dx_r du \\
&\lesssim \int_1^{\Xi} u^{\beta} (\log u)^s du \asymp \Xi^{\beta+1} (\log \Xi)^s,
\end{aligned}$$

where the second step is by Lemma A.15. This completes the proof. □

Lemma A.19. *Suppose that $\beta \ge 0$ and $0 < \alpha \le 2$. Then, as $\Xi \to \infty$,*

$$\begin{aligned}
&\int_{x_1 \cdots x_r \le \Xi, x_k \ge 1} \prod_{k=1}^{r} x_k^{\beta} (x_1^{\alpha} + x_2^{\alpha} + \cdots + x_r^{\alpha})^{-1} dx_1 \cdots dx_r \\
&\asymp \begin{cases}
\Xi^{\beta+1-\alpha/r}, \text{ if } r \ge 3; \\
\log(\Xi), \text{ if } r = 2, \beta = \alpha/2 - 1; \quad \Xi^{\beta+1-\alpha/2} \text{ if } r = 2, \beta > \alpha/2 - 1; \\
1, \text{ if } r = 1, \beta < \alpha - 1; \quad \log(\Xi) \text{ if } r = 1, \beta = \alpha - 1; \\
\Xi^{\beta-\alpha+1} \text{ if } r = 1, \beta > \alpha - 1.
\end{cases}
\end{aligned}$$

Proof. By the symmetry of covariates,

$$\begin{aligned}
&\int_{x_1 \cdots x_r \le \Xi, x_k \ge 1} \prod_{k=1}^{r} x_k^{\beta} (x_1^{\alpha} + x_2^{\alpha} + \cdots + x_r^{\alpha})^{-1} dx_1 \cdots dx_r \\
&\asymp \int_{x_1 \cdots x_r \le \Xi, x_1 \ge x_2 \ge \cdots \ge x_r \ge 1} \prod_{k=1}^{r} x_k^{\beta} (x_1^{\alpha} + x_2^{\alpha} + \cdots + x_r^{\alpha})^{-1} dx_r \cdots dx_1 \\
&:= \mathcal{E}.
\end{aligned}$$

First we prove when $r \geq 3$, as $\Xi \to \infty$, we have

$$\mathcal{E} \lesssim \Xi^{\beta+1-\alpha/r}. \tag{A.55}$$

For this, define the set $\mathcal{K} = \left\{0 \leq k \leq r-2 : \left(\frac{\Xi}{x_1 \cdots x_{r-k-1}}\right)^{1/(k+1)} \leq x_{r-k-1}\right\}$. If $\mathcal{K}$ is not empty, we denote the smallest element in $\mathcal{K}$ by k^*. Then $0 \leq k^* \leq r-2$. For any $(x_1, \ldots, x_r) \in \{(x_1, \ldots, x_r) : x_1 \cdots x_r \leq \Xi, x_1 \geq x_2 \geq \cdots \geq x_r \geq 1, x_r \leq x_{r-1} \leq \frac{\Xi}{x_1 \cdots x_{r-1}}\}$, we have

$$\begin{cases} 1 \leq x_{r-k} \leq x_{r-k-1} & \text{for } 0 \leq k \leq k^*-1, \\ 1 \leq x_{r-k^*} \leq \left(\frac{\Xi}{x_1 \cdots x_{r-k^*-1}}\right)^{1/(k^*+1)} & \text{for } k = k^*, \\ x_{r-k} \geq \left(\frac{\Xi}{x_1 \cdots x_{r-k-1}}\right)^{1/(k+1)} & \text{for } k^*+1 \leq k \leq r-2, \\ x_1 \geq \Xi^{1/r} & \text{for } k = r-1. \end{cases} \tag{A.56}$$

Thus, as $\Xi \to \infty$,

$$\begin{aligned} \mathcal{E} &\lesssim \int_{x_1 \cdots x_r \leq \Xi, x_1 \geq x_2 \geq \cdots \geq x_r \geq 1} \\ &\qquad \left\{(x_1)^{\beta-\alpha/(r-1)} \cdots (x_{r-k^*-1})^{\beta-\alpha/(r-1)}\right\} x_{r-k^*}^{\beta} \\ &\qquad\quad \cdot \left\{(x_{r-k^*+1})^{\beta-\alpha/(r-1)} \cdots (x_r)^{\beta-\alpha/(r-1)}\right\} d\mathbf{x} \\ &\asymp \int_{x_1 \cdots x_r \leq \Xi, x_1 \geq x_2 \geq \cdots \geq x_r \geq 1} \\ &\qquad \left\{(x_1)^{\beta-\alpha/(r-1)} \cdots (x_{r-k^*-1})^{\beta-\alpha/(r-1)}\right\} \\ &\qquad\quad \cdot (x_{r-k^*})^{[\beta+1-\alpha/(r-1)]k^*+\beta} dx_{r-k^*} dx_{r-k^*-1} \cdots dx_1 \\ &\asymp \int_{x_1 \cdots x_r \leq \Xi, x_1 \geq x_2 \geq \cdots \geq x_r \geq 1} \\ &\qquad \left\{(x_1)^{-1-\alpha/[(r-1)(k^*+1)]} \cdots (x_{r-k^*-1})^{-1-\alpha/[(r-1)(k^*+1)]}\right\} \\ &\qquad\quad \cdot \Xi^{\beta+1-\alpha k^*/[(r-1)(k^*+1)]} dx_{r-k^*-1} \cdots dx_1 \\ &= \Xi^{\beta+1-\alpha/r}, \end{aligned} \tag{A.57}$$

where the first step uses $x_{r-k^*} \geq 1$ and Lemma A.15, the second step uses $x_{r-k} \leq x_{r-k-1}$ for

all $k \le k^* - 1$ in (A.56), the third step uses the upper bound on x_{r-k^*} in (A.56), the fourth step uses the lowers bounds on x_{r-k} for all $k^* + 1 \le k \le r - 2$ in (A.56). If $\mathcal{K}$ is empty, then for any $(x_1, \dots, x_r) \in \{(x_1, \dots, x_r) : x_1 \cdots x_r \le \Xi, x_1 \ge x_2 \ge \cdots \ge x_r \ge 1, x_r \le x_{r-1} \le \Xi/(x_1 \cdots x_{r-1})\}$, it satisfies

$$1 \le x_k \le x_{k-1} \text{ for any } 2 \le k \le r, \quad \text{and} \quad 1 \le x_1 \le \Xi^{1/r}.$$

Thus, as $\Xi \to \infty$,

$$\begin{aligned} \mathcal{E} &= \int_1^{\Xi^{1/r}} \cdots \int_1^{x_{r-2}} \int_1^{x_{r-1}} \\ &\qquad \prod_{k=1}^{r} x_k^{\beta} (x_1^{\alpha} + x_2^{\alpha} + \cdots + x_{r-1}^{\alpha} + x_r^{\alpha})^{-1} dx_r dx_{r-1} \cdots dx_1 \\ &\lesssim \int_1^{\Xi^{1/r}} \cdots \int_1^{x_{r-2}} \int_1^{x_{r-1}} \\ &\qquad x_1^{\beta - \alpha/r} \cdots x_{r-1}^{\beta - \alpha/r} x_r^{\beta - \alpha/r} dx_r dx_{r-1} \cdots dx_1 \asymp \Xi^{\beta + 1 - \alpha/r}. \end{aligned} \tag{A.58}$$

Combining (A.57) and (A.58) completes the proof for (A.55).

On the other hand, when $r \ge 3$ and as $\Xi \to \infty$,

$$\begin{aligned} \mathcal{E} &\ge \int_1^{\Xi^{1/r}} \cdots \int_1^{x_{r-2}} \int_1^{x_{r-1}} \\ &\qquad \prod_{k=1}^{r} x_k^{\beta} (x_1^{\alpha} + \cdots + x_{r-1}^{\alpha} + x_r^{\alpha})^{-1} dx_r dx_{r-1} \cdots dx_1 \\ &\ge \int_1^{\Xi^{1/r}} \cdots \int_1^{x_{r-2}} \int_1^{x_{r-1}} \\ &\qquad \prod_{k=1}^{r} x_k^{\beta} \cdot r^{-1} x_1^{-\alpha} dx_r dx_{r-1} \cdots dx_1 \asymp \Xi^{\beta + 1 - \alpha/r}. \end{aligned} \tag{A.59}$$

Therefore, combining (A.55) and (A.59) completes the proof of the lemma for $r \ge 3$.

Then we consider for $r = 2$. For $0 < \alpha \leq 2$,

$$\begin{aligned}\mathcal{E} &\leq 2\int_1^{\sqrt{\Xi}}\int_1^{x_1} x_1^{\beta-\alpha}x_2^{\beta}dx_2dx_1 + 2\int_{\sqrt{\Xi}}^{\Xi}\int_1^{\Xi/x_1} x_1^{\beta-\alpha}x_2^{\beta}dx_2dx_1 \\ &\asymp \begin{cases}\log(\Xi) & \text{when } 2\beta+2-\alpha=0 \\ \Xi^{\beta+1-\alpha/2} & \text{when } 2\beta+2-\alpha>0\end{cases} \qquad \text{as } \Xi\to\infty. \end{aligned} \tag{A.60}$$

On the other hand, we have

$$\begin{aligned}\mathcal{E} &\geq \int_1^{\sqrt{\Xi}}\int_1^{x_1} x_1^{\beta}x_2^{\beta}(x_1^{\alpha}+x_2^{\alpha})^{-1}dx_2dx_1 \\ &\geq 2^{-1}\int_1^{\sqrt{\Xi}}\int_1^{x_1} x_1^{\beta-2}x_2^{\beta}dx_2dx_1 \\ &\asymp \begin{cases}\log(\Xi) & \text{when } 2\beta+2-\alpha=0 \\ \Xi^{m} & \text{when } 2\beta+2-\alpha>0\end{cases} \qquad \text{as } \Xi\to\infty. \end{aligned} \tag{A.61}$$

Combining (A.60) and (A.61) completes the proof of the lemma for $r = 2$.

Finally, we consider for $r = 1$. Note that $\int_1^{\Xi} x_1^{\beta}x_1^{-\alpha}dx_1 \asymp 1$ when $0 \leq \beta < \alpha - 1$, and $\int_1^{\Xi} x_1^{\beta}x_1^{-\alpha}dx_1 \asymp \log(\Xi)$ when $\beta = \alpha - 1$, and $\int_1^{\Xi} x_1^{\beta}x_1^{-\alpha}dx_1 \asymp \Xi^{\beta-\alpha+1}$ when $\beta > \alpha - 1$. This complete the proof. □

Lemma A.20. *Suppose that $\beta \leq -1$ and $\alpha > 0$. Then, as $\Xi \to \infty$,*

$$\int_{x_1\cdots x_r\geq\Xi, x_k\geq 1}\prod_{k=1}^{r} x_k^{\beta}(x_1^{\alpha}+x_2^{\alpha}+\cdots+x_r^{\alpha})^{-1}dx_1\cdots dx_r \asymp \Xi^{\beta+1-\alpha/r}.$$

Proof. The proof is similar to the proof for Lemma A.19. We omit the details here. □

Lemma A.21. *Suppose that $m > 1$. Then, as $\Xi \to \infty$,*

$$\int_{x_1^{(m-1)/m}x_2\cdots x_r\leq\Xi, x_k\geq 1}(x_1^2+x_2^2+\cdots+x_r^2)^{-1}x_1^2dx_1\cdots dx_r \asymp \Xi^{m/(m-1)}.$$

Proof. When $r = 1$, the lemma can be verified by direct calculations. In what follows,

assume $r \geq 2$. First, we show that the left-hand side of the formula above is larger than the right-hand side up to some constant. It suffices to consider a subset of $(x_1, x_2, \ldots, x_r)$ which satisfy $x_1^{(m-1)/m} \geq x_2 \geq \cdots \geq x_r \geq 1$. Let $u_1 = x_1^{(m-1)/m}$, and $u_j = u_1 x_2 \cdots x_j$ for $2 \leq j \leq r$. By changing variables $(x_1, x_2, \ldots, x_r)$ to $(u_1, u_2, \ldots, u_r)$, the left-hand side in the lemma satisfies

$$
\begin{aligned}
&\int_{x_1^{(m-1)/m} x_2 \cdots x_r \leq \Xi, x_k \geq 1} (x_1^2 + x_2^2 + \cdots + x_r^2)^{-1} x_1^2 dx_1 \cdots dx_r \\
&\geq \int_{x_1^{(m-1)/m} x_2 \cdots x_r \leq \Xi, x_k \geq 1} (r x_1^2)^{-1} x_1^2 dx_1 \cdots dx_r \\
&= r^{-1} \int_1^{\Xi} \int_{u_r^{(r-1)/r}}^{u_r} \cdots \int_{u_2^{1/2}}^{u_2} u_1^{1/(m-1)} u_1^{-1} \cdots u_{r-1}^{-1} du_1 \cdots du_{r-1} du_r \\
&\asymp \Xi^{m/(m-1)}.
\end{aligned}
$$

Second, we show that right-hand side of the formula above is larger than the left-hand side up to some constant. Note that $(x_1^2 + x_2^2 + \cdots + x_r^2)^{-1} x_1^2 \leq 1$, so the left-hand side satisfies

$$
\begin{aligned}
&\int_{x_1^{(m-1)/m} x_2 \cdots x_r \leq \Xi, x_k \geq 1} (x_1^2 + x_2^2 + \cdots + x_r^2)^{-1} x_1^2 dx_1 \cdots dx_r \\
&\leq \int_{x_1^{(m-1)/m} x_2 \cdots x_r \leq \Xi, x_k \geq 1} 1 dx_1 \cdots dx_r \\
&= r^{-1} \int_1^{\Xi} \int_{u_r^{(r-1)/r}}^{u_r} \cdots \int_{u_2^{1/2}}^{u_2} u_1^{1/(m-1)} u_1^{-1} \cdots u_{r-1}^{-1} du_1 \cdots du_{r-1} du_r \\
&\asymp \Xi^{m/(m-1)}.
\end{aligned}
$$

This completes the proof. □

Appendix B

Appendix For: High-Dimensional Smoothing Splines with Application to Alzheimer's Disease Prediction Using Longitudinal and Heterogeneous Magnetic Resonance Imaging

B.1 ADNI Database Description

The Alzheimer's Disease Neuroimaging Initiative (ADNI) was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organization, as a $60 million, five year public-private partnership.

The Principal Investigator of ADNI is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations. ADNI recruited from over 50 sites across the U.S. and Canada. The initial phase of ADNI recruited 800 adults, aged 55 to 90 and having a study partner able to provide an independent evaluation of functioning, to participate in the research. Among them, there are approximately 200 healthy control older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years. See `www.adni-info.org` for up-to-date information.

The primary goal of ADNI has been to test whether serial Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Criteria for the different diagnostic groups are summarized in Table B.1. Cognitively healthy control (HC) subjects must have no significant cognitive impairment or impaired activities of daily living. Clinical diagnosed Alzheimer's disease patients (AD) must have had mild AD and had to meet the National Institute of Neurological and Communicative Disorders and Stroke–Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) criteria for probable AD in McKhann et al. (1984). The mild cognitive impairment subjects (MCI) should meet defined criteria for MCI but do not meet the criteria in McKhann et al. (1984) and the MCI subjects should have largely intact general cognition as well as functional performance. Study subjects should have given written informed consent at the time of enrollment for imaging and genetic sample collection and completed questionnaires approved by each participating sites Institutional Review Board (IRB).

B.2 Preprocessing of the Brain MRI Used Here

The structural MRI used in this study are cortical gray matter volumes processed using FreeSurfer software version 4.4 longitudinal image processing framework (`https:`

Table B.1: ADNI recruitment criteria of CN, MCI and AD subjects. AD: Alzheimer's disease; CDR: Clinical Dementia Rating; HC: Healthy controls; MCI: Mild cognitive impairment; MMSE: Mini-Mental State Examination; Edu : years of education.

	HC	MCI	AD
Memory complaints	Absent	Present	Present
CDR	0	0.5	0.5-1.0
Delayed recall Logical Memory II subscale of WMSR	16 Edu:≥ 9 8-15 Edu:≥ 5 0-7 Edu:≥ 3	16 Edu:≤ 8 8-15 Edu:≤ 4 0-7 Edu:≤ 2	16 Edu:≤ 8 8-15 Edu:≤ 4 0-7 Edu:≤ 2

`//surfer.nmr.mgh.harvard.edu/`) ("ucsffsl" file). This dataset has been used in, for example, Mah et al. (2015); Toledo et al. (2014); Tosun et al. (2011). Specifically, subjects with a 1.5-T MRI were included in the dataset where the scans were preprocessed by certain correction methods including gradwarp, B1 calibration, N3 correction, and skull-stripping (see, e.g., Jack Jr et al. (2008) for detail), and the FreeSurfer 4.4 implements the symmetric registration Reuter et al. (2010) and unbiased robust template estimation Reuter et al. (2012). Only MRIs which passed the quality control for all the areas were included in our study. There are total 393 ROIs of brain MRI created by FreeSurfer 4.4 and they consist of volumes of brain regions obtained after cortical parcellation and white matter parcellation, surface area of the brain regions and cortical thickness of the brain regions. However, some ROIs are missing more than 90% across all samples due to the preprocessing. In Section 3.3 of the paper, we use 324 ROIs with at most 20% missing values across the preprocessed samples.

B.3 Proof of Theorem 3.1

Denote by $A(b, \beta_1(\cdot) \ldots, \beta_p(\cdot))$ the functional to be minimized in (3.3). It is clear that $A(b, \beta_1(\cdot) \ldots, \beta_p(\cdot))$ is convex and continuous in $\beta_j(\cdot)$s. Denote by $J(\beta_1(\cdot) \ldots, \beta_p(\cdot)) = \lambda \sum_{j=1}^{p} \|\beta_j\|_{\mathcal{H}_K}$, and without loss of generality, we assume $\lambda = 1$. Denote by $c_K = \max_{i,\nu} K^{1/2}(t_{i\nu}, t_{i\nu})$ and $c_x = \max_{j,i,\nu} |x_{ij}(t_{i\nu})|$. By Cauchy-Schwarz inequality, for any $i = 1, \ldots, n$, $\nu =$

$1, \ldots, m_i$,

$$
\begin{aligned}
&|\sum_{j=1}^{p} \beta_j(t_{i\nu}) x_{ij}(t_{i\nu})| = |\langle \sum_{j=1}^{p} \beta_j(\cdot) x_{ij}(t_{i\nu}), K(t_{i\nu}, \cdot)\rangle_{\mathcal{H}_K}| \\
&\leq \| \sum_{j=1}^{p} \beta_j(\cdot) x_{ij}(t_{i\nu})\|_{\mathcal{H}_K} K(t_{i\nu}, t_{i\nu}) \leq c_K \| \sum_{j=1}^{p} \beta_j(\cdot) x_{ij}(t_{i\nu})\|_{\mathcal{H}_K} \leq c_K c_x J(b, \ldots, \beta_p).
\end{aligned} \tag{B.1}
$$

Denote $\rho = \max_{i,\nu}\{y_{i\nu}^2 + |y_{i\nu}| + 1\}$. Consider the set

$$
\Omega = \{\beta_1(\cdot), \ldots, \beta_p(\cdot) \in \mathcal{H}_K, b \in \mathbb{R} : J(\beta_1(\cdot), \ldots, \beta_p(\cdot)) \leq \rho, |b| \leq \rho^{1/2} + (c_K c_x + 1)\rho\}.
$$

Since Ω is closed, convex, and bounded set, there exists a minimizer for (3.3) in Ω. Denote the minimizer by $\tilde{\beta}_0, \tilde{\beta}_1(\cdot), \ldots, \tilde{\beta}_p(\cdot)$. Then, $A(\tilde{\beta}_0, \tilde{\beta}_1(\cdot), \ldots, \tilde{\beta}_p(\cdot)) \leq A(0, 0, \ldots, 0) < \rho$. On the other hand, for any $\beta_1(\cdot), \ldots, \beta_p(\cdot) \in \mathcal{H}_K$ satisfying $J(\beta_1(\cdot), \ldots, \beta_p(\cdot)) > \rho$. It is clear that $A(b, \beta_1(\cdot) \ldots, \beta_p(\cdot)) \geq J(\beta_1(\cdot), \ldots, \beta_p(\cdot)) > \rho$. For any $\beta_1(\cdot), \ldots, \beta_p(\cdot) \in \mathcal{H}_K$ with $J(\beta_1(\cdot), \ldots, \beta_p(\cdot)) \leq \rho$ and $|b| > \rho^{1/2} + (c_K c_x + 1)\rho$, (B.1) implies that for any $i = 1, \ldots, n$, $\nu = 1, \ldots, m_i$,

$$
|b + \sum_{j=1}^{p} \beta_j(t_{i\nu}) x_{ij}(t_{i\nu}) - y_{i\nu}| > \rho^{1/2} + (c_K c_x + 1)\rho - c_K c_x \rho - \rho = \rho^{1/2}.
$$

Hence, $A(b, \beta_1(\cdot), \ldots, \beta_p(\cdot)) > \rho$. Therefore, for any $b, \beta_1(\cdot), \ldots, \beta_p(\cdot) \notin \Omega$, we have that $A(b, \beta_1(\cdot), \ldots, \beta_p(\cdot)) > A(\tilde{\beta}_0, \tilde{\beta}_1(\cdot), \ldots, \tilde{\beta}_p(\cdot))$, where $\tilde{\beta}_0, \tilde{\beta}_1(\cdot), \ldots, \tilde{\beta}_p(\cdot)$ is the minimizer of (3.3). This completes the proof.

B.4 Algorithm

This algorithm is based on Theorem 3.2 whose proof is given later in Appendix B.5. Consider for any fixed $\theta_1, \ldots, \theta_p \geq 0$. If $\theta_j = 0$ for some j, then $\beta_j = 0$ in the optimization (3.4). Without less of generality, let $\theta_1, \ldots, \theta_p > 0$ and (3.4) is equivalent to the smoothing spline

type problem: find $b \in \mathbb{R}, \beta_1(\cdot), \ldots, \beta_p(\cdot) \in \mathcal{H}_K$ to minimize

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{\nu=1}^{m_i}[y_{i\nu} - b - \sum_{j=1}^{p}\beta_j(t_{i\nu})x_{ij}(t_{i\nu})]^2 + \sum_{j=1}^{p}(\tau_0\theta_j^{-1})\|\beta_j\|_{\mathcal{H}_K}^2. \tag{B.2}$$

By the representer lemma wah, $\beta_1(\cdot), \ldots, \beta_p(\cdot)$ have a closed form expression:

$$\beta_j(t) = \sum_{i=1}^{n}\sum_{\nu=1}^{m_i} c_{i\nu}^j K(t_{i\nu}, t), \quad \forall j = 1, \ldots, p.$$

Define a $m_{i_1} \times m_{i_2}$ matrix $\Sigma_j^{(i_1,i_2)}$ by

$$\Sigma_j^{(i_1,i_2)} = \begin{pmatrix} x_{i_1 j}(t_{i_1 1})K(t_{i_2 1}, t_{i_1 1}) & \cdots & x_{i_1 j}(t_{i_1 1})K(t_{i_2 m_{i_2}}, t_{i_1 1}) \\ \vdots & & \vdots \\ x_{i_1 j}(t_{i_1 m_{i_1}})K(t_{i_2 1}, t_{i_1 m_{i_1}}) & \cdots & x_{i_1 j}(t_{i_1 m_{i_1}})K(t_{i_2 m_{i_2}}, t_{i_1 m_{i_1}}) \end{pmatrix}$$

and let Σ_j be a $N \times N$ ($N = \sum_{i=1}^n m_i$) matrix where the (i_1, i_2)th $m_{i_1} \times m_{i_2}$ matrix is $\Sigma_j^{(i_1,i_2)}$. Define kernel matrix Σ by

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_2 & \cdots & \Sigma_p \end{pmatrix} \in \mathbb{R}^{N \times N \cdot p}.$$

Let the unknown coefficient vector c^j be

$$c^j = \begin{pmatrix} c_{11}^j & \cdots & c_{1m_1}^j & \cdots & c_{n1}^j & \cdots & c_{nm_n}^j \end{pmatrix}^\top \in \mathbb{R}^N,$$

and

$$c = \begin{pmatrix} \{c^1\}^\top & \{c^2\}^\top & \cdots & \{c^p\}^\top \end{pmatrix}^\top \in \mathbb{R}^{Np}.$$

Write the response vector y as

$$y = \begin{pmatrix} y_{11} & \cdots & y_{1m_1} & \cdots & y_{n1} & \cdots & y_{nm_n} \end{pmatrix}^\top \in \mathbb{R}^N.$$

Let 1_N be the column vector consisting of N 1's. Then (B.2) becomes

$$\frac{1}{N}\left(y-\Sigma c-b1_N\right)^\top\left(y-\Sigma c-b1_N\right)+\sum_{j=1}^{p}(\tau_0\theta_j^{-1})\{c^j\}^\top\Sigma_j c^j,$$

which has the unique solution given as follows:

$$\begin{aligned}\hat{b}&=[1_N^\top(1_{N\times N}-\Sigma\tilde{\Sigma}^{-1}\Sigma^\top)1_N]^{-1}\cdot 1_N^\top(1_{N\times N}-\Sigma\tilde{\Sigma}^{-1}\Sigma^\top)y,\\ \hat{c}&=\tilde{\Sigma}^{-1}\Sigma^\top(y-1_N\hat{b}),\end{aligned}\tag{B.3}$$

where $\tilde{\Sigma}=\Sigma^\top\Sigma+N\text{diag}\{(\tau_0\theta_1^{-1})\Sigma_1,\ldots,(\tau_0\theta_p^{-1})\Sigma_p\}$.

Note that when $\theta_1,\ldots,\theta_p$ are fixed, (3.4) is equivalent to find $b\in\mathbb{R}, c\in\mathbb{R}^{Np}$ to minimize

$$\frac{1}{N}(y-b1_N-\sum_{j=0}^{p}\theta_j\Sigma_j c^j)^\top\cdot(y-b1_N-\sum_{j=0}^{p}\theta_j\Sigma_j c^j)+\sum_{j=0}^{p}(\tau_0\theta_j)\{c^j\}^\top\Sigma_j c^j.\tag{B.4}$$

The minimizer of (B.4) is

$$b=\hat{b}\quad\text{and}\quad c^j=\theta_j^{-1}\hat{c}^j,\quad j=0,1,\ldots,p,$$

where $\hat{b}$ and $\hat{c}$ are given by (B.3).

On the other hand, consider when c is fixed, then the minimization of (3.4) is equivalent to

$$\min_{\theta,b}\|y-\sum_{j=0}^{p}\theta_j\Sigma_j c^j-b1_N\|^2+N\tau_0\sum_{j=0}^{p}\theta_j\{c^j\}^\top\Sigma_j c^j+N\tau_1\sum_{j=0}^{p}\theta_j,$$
$$\text{s.t. }\theta_j\geq 0, j=0,1,\ldots,p,$$

which can be written as

$$\begin{aligned}\min_{\theta,b}\|y-\sum_{j=0}^{p}\theta_j\Sigma_j c^j-b1_N\|^2+N\tau_0\sum_{j=0}^{p}\theta_j\{c^j\}^\top\Sigma_j c^j,&\\ \text{s.t. }\theta_j\geq 0, j=0,1,\ldots,p;\sum_{j=0}^{p}\theta_j\leq M,&\end{aligned}\tag{B.5}$$

for some $M \geq 0$.

Therefore, we propose the algorithm of iterating (B.4) and (B.5) for giving the minimizer of (3.4). We observe in simulations that the objective function in optimization (3.4) decreases quickly in the first iteration and after the first iteration the objective function is close to the objective function at convergence. This motivates us to consider the following one-step update algorithm:

1. Initialization: fix $\theta_j = 1$ for $j = 0, 1, \ldots, p$.
2. Solve for c and b in (B.4) and tune τ_0 according to the generalized cross-validation (GCV). Fix τ_0 at the chosen value in all later steps.
3. For c and b obtained in step 2, solve for θ in (B.5) with a fixed M.
4. With θ obtained in step 3, solve for c and b in (B.4).

We choose the best M in Step 3 according to the fivefold cross-validation. In the simulations we find that when τ_0 is fixed according to step 2, the optimal M seems to be close to the number of important components. This gives a range to determine the tuning for M.

B.5 Proof of Theorem 3.2

Recall that $A(b, \beta_1(\cdot), \ldots, \beta_p(\cdot))$ denotes the functional in (3.3). Let $B(\theta_1, \ldots, \theta_p; b, \beta_1(\cdot), \ldots, \beta_p(\cdot))$ be the functional in (3.4). Observe that

$$\tau_0 \theta_j^{-1} \|\beta_j\|_{\mathcal{H}_K}^2 + \tau_1 \theta_j \geq 2\tau_0^{1/2} \tau_1^{1/2} \|\beta_j\|_{\mathcal{H}_K} = \lambda^2 \|\beta_j\|_{\mathcal{H}_K}, \ \forall \theta_j \geq 0,$$

and the equality in the above formula holds if and only if $\theta_j = \tau_0^{1/2} \tau_1^{-1/2} \|\beta_j\|_{\mathcal{H}_K}$. Therefore,

$$B(\theta_1, \ldots, \theta_p; b, \beta_1(\cdot), \ldots, \beta_p(\cdot)) \geq A(b, \beta_1(\cdot), \ldots, \beta_p(\cdot)), \ \ \forall \theta_j \geq 0,$$

and the equality holds if and only if $\theta_j = \tau_0^{1/2} \tau_1^{-1/2} \|\beta_j\|_{\mathcal{H}_K}$ for all $j = 1, \ldots, p$. This completes the proof.

Appendix C

Appendix For: Selection and Estimation Optimality in High Dimensions with the TWIN Penalty

This material is organized as follows. Section C.1 contains additional minimax optimality results under orthogonal designs. Section C.2 contains proofs for the main results of the paper. Section C.3 gives key lemmas for the proof of main results. Section C.4 presents two coordinate-wise algorithms for TWIN and Section C.5 presents i) simulation results illustrating the effectiveness of the universal tuning parameter values, ii) additional simulation results for a higher dimension and higher number of active variables, iii) further simulation results investigating the impact of τ on TWIN-a and TWIN-b, and iv) prediction simulation results left out of the main text due to space constraints.

C.1 Additional Theoretical Results

C.1.1 Orthogonal Designs

For orthogonal designs, multiplying both sides of (4.1) by $\boldsymbol{X}'$ results in the Gaussian sequence model

$$\boldsymbol{y} = \boldsymbol{\beta} + \boldsymbol{z}, \quad \boldsymbol{z} \sim N(0, \sigma^2 I_p). \tag{C.1}$$

Note that the above model and (4.1) are statistically equivalent. Sparse mean vector estimation under the above Gaussian sequence model has been widely-studied in the literature; see, for example, Bickel (1981); Donoho and Johnstone (1994); Foster and George (1994). However, to our knowledge, only an implicit lower bound of asymptotic risk under linear sparsity, where $k/p \to \epsilon \in (0, 1)$ as $p \to \infty$, has been established (Johnstone, 2017). The following result gives an explicit lower bound under this linear sparsity, where the proof is given later in Section C.2.2.

Theorem C.1. *Suppose that $k/p \to \epsilon \in (0, 1)$ as $p \to \infty$. Let $\boldsymbol{\beta}$ be from the model (C.1). Then*

$$\infty_{\widetilde{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \ge 2\sigma^2(1-\epsilon)^2 k \log(1/\epsilon),$$

where the infimum is taken over all measurable estimators.

We make the following two remarks for Theorem C.1. First, there is a small difference between the lower bound $2\sigma^2 k(1-\epsilon)^2 \log(1/\epsilon)$ given in Theorem C.1 for orthogonal designs and the lower bound $2\sigma^2 k \log(1/\epsilon)$ given in Theorem 4.11 for random Gaussian designs. This difference vanishes when ϵ is small (close to 0). The difference between two lower bounds shows that it is fundamentally more difficult to estimate unknown coefficients under Gaussian random designs, which is partially due to the sample correlation among the columns of design matrix.

Second, the following implicit lower bound is given in the Theorem 8.20 of Johnstone

(2017):

$$\infty_{\widetilde{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = (1 + o(1))\sigma^2 p \beta_0(\epsilon),$$

where $\beta_0(\cdot)$ is a univariate Bayes minimax risk for all priors satisfying the linear sparsity $k/p \to \epsilon \in (0, 1)$. The Proposition 8.18 of Johnstone (2017) shows that $\beta_0(\epsilon) \ge \epsilon$ for all $0 \le \epsilon \le 1$. Together, they imply

$$\infty_{\widetilde{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \ge (1 + o(1))\sigma^2 k. \tag{C.2}$$

Comparing the lower bound in Theorem C.1 with the lower bound in (C.2), it is clear that $2\sigma^2 k(1-\epsilon)^2 \log(1/\epsilon) > \sigma^2 k$ when $\epsilon < 0.33$. Since $k/p < 0.33$ is a reasonable assumption in many applications, Theorem C.1 provides a sharper lower bound than (C.2).

Now we give the asymptotic risk for TWIN estimators under orthogonal designs and the linear sparsity.

Theorem C.2. *Suppose that $p \to \infty$ with $k/p \to \epsilon$ for some constant $\epsilon > 0$. Let $\widehat{\boldsymbol{\beta}}$ be the TWIN-a or TWIN-b estimator in (4.2) and $\boldsymbol{\beta}$ be from the model (C.1). Let $\min_{t\in\mathbb{R}}\{|t| + P'_{\lambda,\tau}(|t|)\} = \sigma(1-\epsilon)\sqrt{2\log(1/\epsilon)}$. Then,*

$$\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = [1 + c(\epsilon)]\, 2\sigma^2(1-\epsilon)^2 k \log(1/\epsilon),$$

with $c(\epsilon) = \frac{2}{\sqrt{\pi \log(1/\epsilon)}\epsilon^{1-(1-\epsilon)^2}(1-\epsilon)}$.

The proof of Theorem C.2 is given in Section C.2.2. We make the following remarks regarding the above theorems. First, comparing Theorem C.2 with the lower bound result of Theorem C.1, it is clear that TWIN estimators are minimax rate optimal. Second, the constant $c(\epsilon)$ decreases as ϵ decreases, and $c(\epsilon)$ approaches to zero when ϵ is close to zero. For example, $c(\epsilon) = 0.58$ when $\epsilon = 0.01$, and $c(\epsilon) = 1.28$ when $\epsilon = 0.1$. Third, if $\epsilon = 0$ (i.e., $k/p \to 0$ as $p \to \infty$), the lower bound of asymptotic risk is known; see, for example, Donoho

and Johnstone (1994); Johnstone (2017), which is

$$\infty_{\widetilde{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = (1 + o(1))2\sigma^2 k \log(p/k). \tag{C.3}$$

We note that (C.3) is different from the lower bound in Theorem C.1. The reason is that k/p are in different sparsity regimes. The following theorem shows that TWIN estimators achieve the asymptotic minimax risk when $k/p \to 0$ as $p \to \infty$.

Theorem C.3. *Suppose that $k/p \to 0$ as $p \to \infty$. Let $\boldsymbol{\beta}$ be from the model (C.1). Then TWIN with $\min_{t\in\mathbb{R}}\{|t| + P'_{\lambda,\tau}(|t|)\} = \sigma\sqrt{2\log(p/k)}$ achieves the minimax optimal risk*

$$\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = (1 + o(1))\, 2\sigma^2 k \log(p/k).$$

The proof of Theorem C.3 is given in Section C.2.2

C.2 Proofs of Main Results

C.2.1 Proofs for Section 4.3

In this section, we give proofs for Section 4.3 in the following order: Proposition 4.4, Theorem 4.7, Corollary 4.8, Theorem 4.5, Corollary 4.6, Theorem 4.9, and Corollary 4.10.

C.2.1.0.1 Proof of Proposition 4.4 We are testing p hypotheses $H_i : \beta_i = 0, i = 1, \ldots, p$. Suppose that the first $p - k$ hypotheses are null, i.e., $\beta_i = 0$ for $i \le p - k$. We reject H_i if and only if $\widehat{\beta}_i \neq 0$. Let V be the number of false rejections and R be the number of total rejections. Hence,

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \sum_{r=1}^{p} \frac{1}{r} \mathbb{E}\left[\sum_{i=1}^{p-k} 1_{\{H_i \text{ is rejected }\}} 1_{\{R=r\}}\right]. \tag{C.4}$$

Note that when H_i is a null hypothesis,

$$\left\{\boldsymbol{X}'\boldsymbol{z} : H_i \text{ is rejected and } R = r\right\} = \left\{\boldsymbol{X}'\boldsymbol{z} : |y_i| > \min_{t\in\mathbb{R}}\{|t| + P'_{\lambda,\tau}(|t|)\} \text{ and } \widetilde{R} = r-1\right\},$$

where $\widetilde{R}$ is the number of total rejections when applying TWIN with same parameters to the following data

$$\widetilde{\boldsymbol{X}'\boldsymbol{z}} = (\boldsymbol{x}_1'\boldsymbol{z}, \ldots, \boldsymbol{x}_{i-1}'\boldsymbol{z}, \boldsymbol{x}_{i+1}'\boldsymbol{z}, \ldots, \boldsymbol{x}_n'\boldsymbol{z}).$$

Let $\min_{t\in\mathbb{R}}\{|t| + P'_{\lambda,\tau}(|t|)\} = \sigma\Phi^{-1}(1-\alpha/2p)$ for any $\alpha \in [0,1]$, then

$$\begin{aligned}
\mathbb{P}\left(H_i \text{ rejected and } R = r\right) &= \mathbb{P}\left(|\boldsymbol{x}_i'\boldsymbol{z}| > \min_{t\in\mathbb{R}}\{|t| + P'_{\lambda,\tau}(|t|)\} \text{ and } \widetilde{R} = r-1\right) \\
&= \mathbb{P}\left(|\boldsymbol{x}_i'\boldsymbol{z}| > \min_{t\in\mathbb{R}}\{|t| + P'_{\lambda,\tau}(|t|)\}\right)\mathbb{P}\left(\widetilde{R} = r-1\right) \\
&= \frac{\alpha}{p}\mathbb{P}\left(\widetilde{R} = r-1\right).
\end{aligned}$$

where the second equality is from the independence of $\boldsymbol{x}_i'\boldsymbol{z}$ and $\widetilde{\boldsymbol{X}'\boldsymbol{z}}$. Plugging this equality into equation (C.4) gives

$$\text{FDR} = \frac{\alpha(p-k)}{p}\sum_{r=1}^{p}\frac{1}{r}\mathbb{P}(\widetilde{R} = r-1) \le \alpha\left(1 - \frac{k}{p}\right).$$

Given $\min_{t\in\mathbb{R}}\{|t| + P'_{\lambda,\tau}(|t|)\} = \sigma\Phi^{-1}(1-\alpha/2p)$ for any $\alpha \in [0,1]$, it is easy to see that FWER $= \alpha$, which completes the proof.

C.2.1.0.2 Proof of Theorem 4.7 Denote by $\tau_\lambda \equiv \tau/\lambda$. If $\tau_\lambda \ge (1 - \delta^{-1/2} - \vartheta)^{-2}$, then Lemma C.4 implies that

$$\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) - \tau_\lambda^{-1} > 0 \quad \text{with probability at least } 1 - e^{-n\vartheta^2/2}. \tag{C.5}$$

Let λ be fixed and $\lambda_0 = P'_{\lambda,\tau}(0+)$. Define $h(t) \equiv \tau_\lambda^{-1} t^2/2 + P_{\lambda,\tau}(|t|) - \lambda_0|t|$. Note that both TWIN-a and TWIN-b satisfy

$$\infty_{0<t_1<t_2}\{P'_{\lambda,\tau}(t_2) - P'_{\lambda,\tau}(t_1)\}/(t_2 - t_1) = -\tau_\lambda^{-1},$$

then $h(t)$ is a continuously differentiable convex function. Note that the penalized loss in (4.2) can be written as

$$L(\boldsymbol{b}; \lambda, \tau) = \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 - \frac{1}{2\tau_\lambda}\|\boldsymbol{b}\|^2 \right\} + \sum_{j=1}^{p} \{\lambda_0|b_j| + h(|b_j|)\},$$

where on the right-hand side, the first part is strictly convex due to (C.5), and the second part is also strictly convex. Thus, $L(\boldsymbol{b}; \lambda, \tau)$ is strictly convex in $\boldsymbol{b} \in \mathbb{R}^p$. On the other hand, observe that if $\boldsymbol{b} = 0$, $L(\boldsymbol{b}; \lambda, \tau) = \|\boldsymbol{y}\|^2/(2n)$ is bounded, thus $\boldsymbol{y} \to \widehat{\boldsymbol{\beta}}$ maps bounded sets of $\boldsymbol{y} \in \mathbb{R}^n$ to bounded sets of $\widehat{\boldsymbol{\beta}}$ in $\mathbb{R}^p$. Since $L(\boldsymbol{b}; \lambda, \tau)$ is continuous and convex in $\boldsymbol{b}$ and is continuous in $\boldsymbol{y}$, the global minimizer of $L(\boldsymbol{b}; \lambda, \tau)$ is unique and continuous in $\boldsymbol{y}$. Thus, (4.12) is the KKT condition and its solution $\widehat{\boldsymbol{\beta}}$ is unique.

Let $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}$ be the oracle least squares estimator on the true support A^o:

$$\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o} = \underset{\boldsymbol{b}\in\mathbb{R}^k}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}_{A^o}\boldsymbol{b}\|^2. \tag{C.6}$$

We define two sets Ω_1 and Ω_2 as follows, where $\lambda_1 \le \lambda_2$ are two positive parameters:

$$\Omega_1(\lambda_1) \equiv \left\{ \max_{j\notin A^o} |\boldsymbol{x}'_j(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o})| < \lambda_1 \right\},$$
$$\Omega_2(\lambda_2) \equiv \left\{ \min_{j\in A^o} \mathrm{sgn}(\beta_j)\widehat{\beta}^{\mathrm{LS}}_j > \gamma\lambda_2 \right\}.$$

For any $j \in A^o$, it is clear that $\boldsymbol{x}'_j(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}) = 0$. If $|\widehat{\beta}^{\mathrm{LS}}_j| \ge \gamma\lambda_2$, then by definition of TWIN-b, $P'_{\lambda,\tau}(|\widehat{\beta}^{\mathrm{LS}}_j|) = 0$. Thus, the vector which is equal to $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}$ for the components corresponding to A^o and 0 otherwise is the unique solution of the KKT condition (4.12) and $\mathrm{sgn}(\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}) = \mathrm{sgn}(\boldsymbol{\beta}_{A^o})$ for all $\lambda_1 \le \lambda \le \lambda_2$ in the intersection of Ω_1 and Ω_2.

Observe that $\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}$ is equal to the projection of $\boldsymbol{z}$ onto the orthogonal complement of $\boldsymbol{X}_{A^o}$. Conditional on $\boldsymbol{z}$ and $\boldsymbol{X}_{A^o}$, $\boldsymbol{X}'_{\bar{A}^o}(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o})$ is distributed as i.i.d. centered Gaussian random variables with variance $\|\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}\|^2/n$. Hence, we can write

$$\boldsymbol{X}'_{\bar{A}^o}(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}) \stackrel{d}{=} \frac{\|\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}\|}{\sqrt{n}}(\zeta_1, \ldots, \zeta_{p-k})',$$

where the ζ_j terms are i.i.d. $N(0,1)$ independent of $\|\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}\|$. Observe that $\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}$ is equal to the projection of $\boldsymbol{z}$ onto the orthogonal complement of $\boldsymbol{X}_{A^o}$, and the orthogonal complement of $\boldsymbol{X}_{A^o}$ of dimension $n-k$ has uniform orientation. Thus by Lemma C.5, we know that for an arbitrary small constant $\vartheta > 0$,

$$\mathbb{P}\left\{\|\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}\| > (1+\vartheta)\sigma\sqrt{n-k}\right\} \le e^{-\vartheta^2(n-k)\sigma^2/2}. \tag{C.7}$$

Observe that

$$\begin{aligned}
&\mathbb{P}\left\{\max_{1\le j\le p-k} \zeta_j^2 > (2+4\vartheta)\log(p-k)\right\} \\
&\le \frac{\mathbb{E}e^{\max_{1\le j\le p-k}\zeta_j^2/(2+\vartheta)}}{e^{(1+\vartheta)\log(p-k)}} \le \frac{\sum_{1\le j\le p-k}\mathbb{E}e^{\zeta_j^2/(2+\vartheta)}}{e^{(1+\vartheta)\log(p-k)}} \le \frac{\sqrt{\pi\vartheta}(p-k)}{e^{(1+\vartheta)\log(p-k)}} = \frac{\sqrt{\pi\vartheta}}{(p-k)^{\vartheta}},
\end{aligned} \tag{C.8}$$

where the first step is by the Markov's inequality. By (C.7) and (C.8), we have that

$$\|\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}\| \le (1+\vartheta)\sigma\sqrt{n-k}, \quad \max_{1\le j\le p-k}|\zeta|_j^2 \le (2+4\vartheta)\log(p-k)$$

hold simultaneously with probability at least $1 - e^{-\vartheta^2(n-k)\sigma^2/2} - \sqrt{\pi\vartheta}/(p-k)^{\vartheta}$. Thus, if

$$\begin{aligned}
\lambda_1 &\ge (1+3\vartheta)\sigma\sqrt{2\log p}\sqrt{1-\frac{\epsilon}{\delta}} \\
&= (1+3\vartheta)\sigma\sqrt{2\log(p-k)}\sqrt{1-\frac{\epsilon}{\delta}} + o(1), \quad \text{with } \epsilon < \delta,
\end{aligned} \tag{C.9}$$

we have

$$1 - \mathbb{P}\left\{\Omega_1(\lambda_1)\right\} \le e^{-\vartheta^2(n-k)\sigma^2/2} + \frac{\sqrt{\pi\vartheta}}{(p-k)^{\vartheta}}. \tag{C.10}$$

By Lemma C.4, we know the minimum singular value of $\boldsymbol{X}_{A^o}$ satisfies

$$\sigma_{\min}(\boldsymbol{X}_{A^o}) \geq 1 - \sqrt{k/n} - \vartheta \quad \text{with probability at least } 1 - e^{-n\vartheta^2/2}. \tag{C.11}$$

Denote by $\|A\|_{\max}$ the maximum absolute value of entries in any matrix A. From (C.11), with probability at least $1 - e^{-n\vartheta^2/2}$, we have

$$\begin{aligned} \|(\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}\|_{\max} &\leq \|(\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}\|_2 \\ &\leq \sigma_{\min}^{-2}(\boldsymbol{X}_{A^o}) \leq (1 - \sqrt{\epsilon/\delta} - \vartheta)^{-2} < \infty, \end{aligned} \tag{C.12}$$

Let $\mathcal{M}(\epsilon, \delta) := (1 - \sqrt{\epsilon/\delta} - \vartheta)^{-2}$. Conditioning on $\boldsymbol{X}_{A^o}$,

$$(\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}\boldsymbol{X}'_{A^o}\boldsymbol{z} \equiv \sigma \cdot (\xi'_1, \ldots, \xi'_k) \sim \sigma \cdot N(0, (\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}),$$

where the terms ξ'_j are independent of $\boldsymbol{z}$. Since $\forall 1 \leq j \leq k$, $\mathbb{E}e^{|\xi'_j|^2/(2+\vartheta)\mathcal{M}(\epsilon,\delta)} \leq \sqrt{\pi\vartheta}$, we have by Markov's inequality,

$$\begin{aligned} &\mathbb{P}\left\{\max_{1\leq j\leq k}(\xi'_j)^2 > (2+4\vartheta)\mathcal{M}(\epsilon,\delta)\log k\right\} \\ &\leq \frac{\mathbb{E}e^{\max_{1\leq j\leq k}(\xi'_j)^2/(2+\vartheta)\mathcal{M}(\epsilon,\delta)}}{e^{(1+\vartheta)\log k}} = \frac{\sum_{1\leq j\leq k}\mathbb{E}e^{|\xi'_j|^2/(2+\vartheta)\mathcal{M}(\epsilon,\delta)}}{k^{1+\vartheta}} \leq \frac{\sqrt{\pi\vartheta}}{k^{\vartheta}} \to 0, \end{aligned}$$

which together with (C.12) implies that

$$\|(\boldsymbol{X}_{A^o}\boldsymbol{X}_{A^o})^{-1}\boldsymbol{X}'_{A^o}\boldsymbol{z}\|_\infty = \sigma \max_{1\leq j\leq k}|\xi'_j| \leq \sigma\sqrt{(2+4\vartheta)\log k}(1 - \sqrt{\epsilon/\delta} - \vartheta)^{-1} \tag{C.13}$$

with probability at least $1 - e^{-n\vartheta^2/2} - \sqrt{\pi\vartheta}/k^{\vartheta}$. Recall that $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o} = \boldsymbol{\beta}_{A^o} + (\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}\boldsymbol{X}'_{A^o}\boldsymbol{z}$, and if

$$|\beta_j| > \gamma\lambda_2 + \sigma\sqrt{(2+4\vartheta)\log k}(1 - \sqrt{\epsilon/\delta} - \vartheta)^{-1}, \quad \forall j \in A^o,$$

we have

$$\min_{j\in A^o}\operatorname{sgn}(\beta_j)\widehat{\beta}^{\mathrm{LS}}_j \geq \min_{j\in A^o}|\beta_j| - \|(\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}\boldsymbol{X}'_{A^o}\boldsymbol{z}\|_\infty > \gamma\lambda_2 \tag{C.14}$$

with probability at least $1 - e^{-n\vartheta^2/2} - \sqrt{\pi\vartheta}/k^{\vartheta}$. The result follows by combining (C.5), (C.10) and (C.14).

C.2.1.0.3 Proof of Corollary 4.8 This corollary can be directly justified from the Theorem 4.7.

C.2.1.0.4 Proof of Theorem 4.5 Denote by $\widehat{\boldsymbol{\beta}}_{A^o}$ the solution to the reduced penalization problem with the oracle support:

$$\widehat{\boldsymbol{\beta}}_{A^o} = \underset{\boldsymbol{b}\in\mathbb{R}^k}{\arg\min} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_{A^o}\boldsymbol{b}\|^2 + \sum_{j=1}^{k} P_{\lambda,\tau}(|b_j|) \right\}. \tag{C.15}$$

As discussed in the proof of Theorem 4.7, if $\tau_\lambda \equiv \tau/\lambda \geq (1 - \delta^{-1/2} - \vartheta)^{-2}$, then (4.12) is the KKT condition and the TWIN estimator is unique with probability at least $1 - e^{-n\vartheta^2/2}$.

By Borell's inequality, we know that with probability at least $1 - e^{-n\vartheta^2/2}$,

$$\|\boldsymbol{X}'_{\bar{A}^o}\boldsymbol{z}\|_\infty \leq \max_{1\leq i\leq n} |x_{1i}| \sqrt{\sum_{i=1}^{n} z_i^2} \leq \sigma(1+\vartheta)\sqrt{2\log p}. \tag{C.16}$$

Observe that

$$\boldsymbol{X}'_{\bar{A}^o}\boldsymbol{X}_{A^o}(\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}) = \boldsymbol{X}'_{\bar{A}^o}\boldsymbol{X}_{A^o}(\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}(\boldsymbol{X}'_{A^o}(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}_{A^o}) - \boldsymbol{X}'_{A^o}\boldsymbol{z}).$$

Conditional on $\boldsymbol{z}$, the terms in $\boldsymbol{X}'_{A^o}\boldsymbol{z}$ are distributed as i.i.d. centered Gaussian random variables with variance $\|\boldsymbol{z}\|^2/n$. Write

$$\boldsymbol{X}'_{A^o}\boldsymbol{z} \overset{d}{=} \frac{\|\boldsymbol{z}\|}{\sqrt{n}}(\zeta_1, \ldots, \zeta_k)',$$

where ζ_j are i.i.d. $N(0,1)$ independent of $\|\boldsymbol{z}\|$. By Lemma C.5 we have,

$$\mathbb{P}\left\{\|\boldsymbol{z}\| \geq \sigma\sqrt{n}(1+\vartheta)\right\} \leq e^{-\vartheta^2 n\sigma^2/2} \text{ and } \mathbb{P}\left\{\sqrt{\zeta_1^2 + \cdots + \zeta_k^2} \geq \sqrt{k}(1+\vartheta)\right\} \leq e^{-k\vartheta^2/2}.$$

Thus with probability at least $1 - e^{-\vartheta^2 n\sigma^2/2} - e^{-k\vartheta^2/2}$,

$$\|\boldsymbol{X}'_{A^o}\boldsymbol{z}\| \leq \sigma\sqrt{k}(1+\vartheta)^2. \tag{C.17}$$

From the proof of Theorem 4.7, we know that if

$$|\beta_j| > \gamma\lambda + \sigma\sqrt{(2+4\vartheta)\log k}(1-\sqrt{\epsilon/\delta}-\vartheta)^{-1}, \quad \forall j \in A^o,$$

and

$$|P'_{\lambda,\tau}(|\widehat{\beta}_j|)| = o(\lambda), \quad \forall j \in A^o, \tag{C.18}$$

then with probability at least $1 - e^{-n\vartheta^2/2} - \sqrt{\pi\vartheta}/k^{\vartheta}$,

$$\min_{j\in A^o} \operatorname{sgn}(\beta_j)\widehat{\beta}_j \geq \min_{j\in A^o}|\beta_j| - \|(\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}\boldsymbol{X}'_{A^o}\boldsymbol{z}\|_\infty - |P'_{\lambda,\tau}(|\widehat{\beta}_j|)| > \gamma\lambda. \tag{C.19}$$

On the other hand, if (C.19) is true, then by the definition of TWIN-a and the uniqueness of the estimator, we conclude that (C.18) holds. Combining (C.17) and (C.18), then

$$\|\boldsymbol{X}'_{A^o}(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}_{A^o}) - \boldsymbol{X}'_{A^o}\boldsymbol{z}\| \leq \|P'_{\lambda,\tau}(|\widehat{\boldsymbol{\beta}}_{A^o}|)\| + \|\boldsymbol{X}'_{A^o}\boldsymbol{z}\| \leq \sqrt{k}\cdot o(\lambda) + (1+\vartheta)^2\sigma\sqrt{k}$$

with probability at least $1 - e^{-\vartheta^2 n\sigma^2/2} - e^{-k\vartheta^2/2} - e^{-n\vartheta^2/2} - \sqrt{\pi\vartheta}/k^{\vartheta}$. By Lemma C.4,

$$\|\boldsymbol{X}_{A^o}(\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}\| \leq \frac{1}{1-\sqrt{\epsilon/\delta}-\vartheta}$$

holds with probability at least $1 - e^{-n\vartheta^2/2}$. Therefore, we have

$$\begin{aligned}
&\|\boldsymbol{X}'_{\bar{A}^o}\boldsymbol{X}_{A^o}(\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o})\|_\infty \\
&= \|\boldsymbol{X}'_{\bar{A}^o}\boldsymbol{X}_{A^o}(\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}(\boldsymbol{X}'_{A^o}(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}_{A^o}) - \boldsymbol{X}'_{A^o}\boldsymbol{z})\|_\infty \\
&\leq \sqrt{\frac{2\log(p-k)}{n}}\|\boldsymbol{X}_{A^o}(\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}(\boldsymbol{X}'_{A^o}(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}_{A^o}) - \boldsymbol{X}'_{A^o}\boldsymbol{z})\| \\
&\leq \left[(1-\vartheta)\sqrt{\delta/\epsilon} - 1\right]^{-1}[\sigma(1+\vartheta)^2 + o(\lambda)]\sqrt{2\log p}.
\end{aligned} \tag{C.20}$$

with probability at least $1 - e^{-n\sigma^2\vartheta^2/2} - e^{-k\vartheta^2/2} - 2e^{-n\vartheta^2/2} - \sqrt{\pi\vartheta}/k^{\vartheta}$. Combining (C.16) and (C.20), with probability at least $1 - e^{-n\sigma^2\vartheta^2/2} - e^{-k\vartheta^2/2} - 3e^{-n\vartheta^2/2} - \sqrt{\pi\vartheta}/k^{\vartheta}$,

$$\begin{aligned} &\|\boldsymbol{X}'_{\bar{A}^o}(\boldsymbol{y} - \boldsymbol{X}_{A^o}\widehat{\boldsymbol{\beta}}_{A^o})\|_\infty \\ &\leq \|\boldsymbol{X}'_{\bar{A}^o}\boldsymbol{X}_{A^o}(\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o})\|_\infty + \|\boldsymbol{X}'_{\bar{A}^o}\boldsymbol{z}\|_\infty \\ &\leq \left\{[(1-\vartheta)\sqrt{\delta/\epsilon} - 1]^{-1}[\sigma(1+\vartheta)^2 + o(\lambda)] + \sigma(1+\vartheta)\right\}\sqrt{2\log p}. \end{aligned} \tag{C.21}$$

Hence, by letting $\lambda \geq \{[(1-\vartheta)\sqrt{\delta/\epsilon} - 1]^{-1}(1+\vartheta) + 1\}(1+\vartheta)\sigma\sqrt{2\log p}$, we complete the proof.

C.2.1.0.5 Proof of Corollary 4.6 Note that $\sqrt{\delta/\epsilon} - 1 > \sqrt{\delta}$ for $\epsilon \leq 1/4$. The rest of the proof follows directly from Theorem 4.5.

C.2.1.0.6 Proof of Theorem 4.9 For $m \geq 1$, we define a semi-norm for any $\boldsymbol{v} \in \mathbb{R}^n$ as

$$\zeta(\boldsymbol{v}; m, A^o) \equiv \max\left\{\frac{\|(\boldsymbol{P}_A - \boldsymbol{P}_{A^o})\boldsymbol{v}\|}{\sqrt{m}} : A^o \subseteq A \subseteq \{1,\ldots,p\}, |A| = m + k\right\},$$

where $\boldsymbol{P}_A$ is the orthogonal projection from $\mathbb{R}^n$ to the span of $\{\boldsymbol{x}_j : j \in A\}$ and recall that $\{\boldsymbol{x}_j\}$ are columns of the design matrix $\boldsymbol{X}$. Let $\{\lambda^{(x)} : x \in [0,\infty)\}$ be a continuous path with $\lambda^{(0)} = +\infty$ and $\lim_{x\to\infty}\lambda^{(x)} = 0$ and $\boldsymbol{\beta}^{(x)}$ be the TWIN estimator corresponding to $\lambda = \lambda^{(x)}$. Let $x_1 = \infty_{x\geq 0}\{\lambda^{(x)} < \max\{(1+3\vartheta)\sqrt{1-\epsilon\delta^{-1}}\sigma\sqrt{2\log p}, 2[1+\vartheta+\sqrt{(\epsilon/\delta+1)/2}]\zeta(\boldsymbol{y}; (k-n)/2, A^o)\}\}$ and $\vartheta > 0$. Note that $P_{\lambda,\tau}$ satisfies $\lambda(1 - t/(\tau_\lambda\lambda))_+ \leq |P'_{\lambda,\tau}(t)| \leq \lambda$, where recall that $\tau_\lambda \equiv \tau/\lambda$.

By Lemma 1 and Remark 5 in Zhang (2010), together with Lemma C.4, we have that with probability at least $1 - e^{-n\vartheta^2/2}$,

$$\#\{j \notin A^o : \widehat{\beta}_j^{(x)} \neq 0\} < K_* k \quad \text{for } 0 \leq x \leq x_1,$$

where K_* is a constant satisfying

$$K_* \leq [1+\vartheta+\sqrt{(1+K_*)\epsilon\delta^{-1}}]^2[1-\vartheta-\sqrt{(1+K_*)\epsilon\delta^{-1}}]^{-2}-1/2. \tag{C.22}$$

We further require that K^* satisfies

$$(K_*+1)k \leq (k+n)/2. \tag{C.23}$$

A sufficient condition for the existence of K_* satisfying (C.22) and (C.23) is by letting

$$[1+\vartheta+\sqrt{(1+K_*)\epsilon\delta^{-1}}]^2[1-\vartheta-\sqrt{(1+K_*)\epsilon\delta^{-1}}]^{-2}-1/2 \leq \{(k+n)/2\}/k-1.$$

In order that the above inequality holds, it suffices to require

$$\epsilon/\delta \leq 0.12.$$

We let

$$\tau_\lambda \geq \left(1-\vartheta-\sqrt{(\epsilon/\delta+1)/2}\right)^{-2},$$

then from the proof of Theorem 4.7, we know (C.23) ensures the uniqueness of the TWIN estimator. On the other hand, $\zeta(\boldsymbol{y};(n-k)/2,A^o)=\zeta(\boldsymbol{z};(n-k)/2,A^o)$ as shown in the proof of Theorem 6 in Zhang (2010). By Lemma 2 in Zhang (2010), we have that

$$\mathbb{P}\left\{\zeta(\boldsymbol{z};(n-k)/2,A^o) \leq \sigma\sqrt{2\log\tilde{p}_\theta}\right\} \geq 1-\frac{\theta/2}{\sqrt{\log\tilde{p}_\theta}} \quad \forall\theta\in(0,1],$$

where $\tilde{p}_\theta$ is defined as the solution of

$$2\log\tilde{p}_\theta - 1 - \log(2\log\tilde{p}_\theta) = \frac{4}{n-k}\left\{\log\binom{p-k}{(n-k)/2}+\log\left(\frac{1}{\theta}\right)\right\}. \tag{C.24}$$

Therefore, we can let

$$\begin{cases} \lambda \geq \max\left\{(1+3\vartheta)\sqrt{1-\epsilon\delta^{-1}}\sigma\sqrt{2\log p}, 2[1+\vartheta+\sqrt{(\epsilon/\delta+1)/2}]\sigma\sqrt{2\log\tilde{p}_\theta}\right\}, \\ |\beta_j| > \gamma\lambda + \sigma\sqrt{(2+4\vartheta)\log k}(1-\epsilon^{1/2}\delta^{-1/2}-\vartheta)^{-1}. \end{cases}$$

From the proof of Theorem 4.7, it is known that the unique TWIN estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{A^o}$, the oracle LSE defined in (C.6). When n is large, we have

$$\begin{aligned} \mathbb{P}\left\{\widehat{A} \neq A^o\right\} &\leq \mathbb{P}\left\{\widehat{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}^o \text{ or } \mathrm{sgn}(\widehat{\boldsymbol{\beta}}) \neq \mathrm{sgn}(\boldsymbol{\beta})\right\} \\ &\leq e^{-\vartheta^2(n-k)\sigma^2/2} + 2e^{-n\vartheta^2/2} + \sqrt{\pi\vartheta}(p-k)^{-\vartheta} + \sqrt{\pi\vartheta}k^{-\vartheta} \\ &\quad + \frac{\theta}{(2\log\tilde{p}_\theta - 1 + 2/(n-k))\sqrt{(n-k)\pi/2}}. \end{aligned}$$

Now let $\theta = 1$; it is known from (C.24) that when n is large,

$$\begin{aligned} \log\tilde{p}_1 &= \frac{1}{\delta}\left[(1-\epsilon)\log(1-\epsilon) - (\delta-\epsilon)\log(\delta-\epsilon) - (1-\delta)\log(1-\delta)\right] + \frac{1}{2} + o(1) \\ &= \tilde{c} + \frac{1}{2} + o(1), \end{aligned}$$

where $\tilde{c} \equiv [(1-\epsilon)\log(1-\epsilon) - (\delta-\epsilon)\log(\delta-\epsilon) - (1-\delta)\log(1-\delta)]/\delta$.

C.2.1.0.7 Proof of Corollary 4.10 The corollary follows by directly verifying the conditions of Theorem 4.9. In particular, note that $\tilde{c} \equiv [(1-\epsilon)\log(1-\epsilon) - (\delta-\epsilon)\log(\delta-\epsilon) - (1-\delta)\log(1-\delta)]/\delta$ increases as ϵ decreases. Hence $\tilde{c} \leq \log\delta^{-1} + (\delta^{-1}+1)\log(1-\delta)^{-1} \equiv \tilde{c}'$, which is a constant. Then as $p \to \infty$, $\sigma\sqrt{2\log p} > 4\sigma\sqrt{2\tilde{c}'+1} \geq 2[1+\vartheta+\sqrt{(\epsilon/\delta+1)/2}]\sigma\sqrt{2\tilde{c}+1}$.

C.2.2 Proofs for Section C.1

Now we give proofs for estimation properties of TWIN under the orthogonal designs in Section C.1. The proofs for results are organized in the following order: Theorem C.1, Theorem C.2, and Theorem C.3.

C.2.2.0.1 Proof of Theorem C.1 From the scale invariance, we only need to prove for $\sigma = 1$. By the max-min inequality (see, e.g., Johnstone (2017)), we have for any $\vartheta > 0$,

$$\infty_{\widetilde{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}_\Pi \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \ge \sup_{\|\Pi\|_0 \le k} \infty_{\widetilde{\boldsymbol{\beta}}} \mathbb{E}_\Pi \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2, \tag{C.25}$$

Here, Π denotes any distribution on $\mathbb{R}^p$ such that any realization $\boldsymbol{\beta}$ obeys $\|\boldsymbol{\beta}\|_0 \le k$. Without loss of generality, assume that $p/k = 1/\epsilon$ is an integer since otherwise we can replace p with $p' = k\lfloor p/k \rfloor$, where $\lfloor p/k \rfloor$ denotes the integer part of p/k, and let Π be supported on $\{1, \ldots, p'\}$. We decompose $\{1, \ldots, p\}$ into k blocks as follows

$$\{1, \ldots, 1/\epsilon\}, \ \{1/\epsilon + 1, \ldots, 2/\epsilon\}, \ \ldots, \ \{(k-1)/\epsilon + 1, \ldots, k/\epsilon\},$$

and choose a particular prior Π as to uniformly random select a coordinate in each block and set its amplitude to $\sqrt{2\log(1/\epsilon)}$. Then the total loss can be decomposed as $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = L_1 + \cdots + L_k$, where L_j is the loss from coordinates on the jth block $\{(j-1)/\epsilon + 1, \ldots, j/\epsilon\}$.

We first prove for any $j = 1, \ldots, k$,

$$\infty_{\widetilde{\boldsymbol{\beta}}} \mathbb{E}[L_j] \ge 2(1-\epsilon)^2 \log(1/\epsilon). \tag{C.26}$$

Without loss of generality, let $j = 1$ and I be the index in $\{1, \ldots, 1/\epsilon\}$ whose amplitude is set to $\sqrt{2\log(1/\epsilon)}$ by Π. Observe that

$$\begin{aligned}
\mathbb{E}[L_1] = \sum_{i=1}^{1/\epsilon} \mathbb{E}[\widetilde{\beta}_i - \beta_i]^2 &= \sum_{i=1}^{1/\epsilon} \left[(1-\epsilon)\mathbb{E}_{\beta_i = 0}[\widetilde{\beta}_i^2] + \epsilon \mathbb{E}_{\beta_i = \sqrt{2\log(1/\epsilon)}}[\widetilde{\beta}_i - \beta_i]^2 \right] \\
&\ge \mathbb{E}_{\beta_1 = \sqrt{2\log(1/\epsilon)}}[\widetilde{\beta}_1 - \beta_1]^2 = t_\epsilon^2 \mathbb{E}_{\beta_1 = \sqrt{2\log(1/\epsilon)}}[\mathbb{P}(I = 1|\boldsymbol{y}) - 1]^2,
\end{aligned}$$

where the last step uses the fact $\mathbb{E}[\beta_i|\boldsymbol{y}] = \sqrt{2\log(1/\epsilon)}\mathbb{P}(I = i|\boldsymbol{y})$. Now we only need to study $\mathbb{P}(I = 1|\boldsymbol{y})$. Recall that if $I = 1$, then $y_1 = \sqrt{2\log(1/\epsilon)} + z_1$, $y_i = z_i$ for $2 \le i \le 1/\epsilon$,

and

$$\mathbb{P}_\Pi\{I=1|\boldsymbol{y}\} = \frac{e^{\sqrt{2\log(1/\epsilon)}y_1}}{\sum_{i=1}^{1/\epsilon} e^{\sqrt{2\log(1/\epsilon)}y_i}} = \frac{e^{\sqrt{2\log(1/\epsilon)}z_1+2\log(1/\epsilon)}}{\sum_{i=1}^{1/\epsilon} e^{\sqrt{2\log(1/\epsilon)}z_i}}$$
$$= \left[1+\{(1/\epsilon-1)^{-1}e^{-\log(1/\epsilon)}\sum_{i=2}^{1/\epsilon} e^{\sqrt{2\log(1/\epsilon)}z_i}\}\{(1/\epsilon-1)e^{-\log(1/\epsilon)-\sqrt{2\log(1/\epsilon)}z_1}\}\right]^{-1},$$

By Jensen's inequality and the independence among the z_i terms, we have

$$\mathbb{E}[\mathbb{P}_\Pi\{I=1|\boldsymbol{y}\}]$$
$$\geq [1+\mathbb{E}\{(1/\epsilon-1)^{-1}e^{-\log(1/\epsilon)}\sum_{i=2}^{1/\epsilon} e^{\sqrt{2\log(1/\epsilon)}z_i}\}\{(1/\epsilon-1)e^{-\log(1/\epsilon)-\sqrt{2\log(1/\epsilon)}z_1}\}]^{-1} = \epsilon,$$

and similarly, $\mathbb{E}[\mathbb{P}_\Pi\{I=1|\boldsymbol{y}\}]^2 \geq \epsilon^2$. Then, (C.26) follows and by the independence of $L_1,\ldots,L_k$,

$$\sup_{\|\Pi\|_0\leq k} \infty_{\widetilde{\boldsymbol{\beta}}}\mathbb{E}_\Pi\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2 \geq 2(1-\epsilon)^2 k\log(1/\epsilon).$$

We complete the proof by (C.29).

C.2.2.0.2 Proof of Theorem C.2 In this proof, we set $\sigma=1$ and assume without loss of generality that $\beta_j \geq 0$. Let the risk function be $r(\lambda,\tau;\beta_j) = \mathbb{E}[\widehat{\beta}_j-\beta_j]^2$ and $r(\lambda,\tau;\boldsymbol{\beta}) = \sum_{j=1}^p r(\lambda,\tau;\beta_j) = \mathbb{E}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2$. Let $\kappa := \kappa(\lambda,\tau) = \min_{t\in\mathbb{R}}\{|t|+P'_{\lambda,\tau}(|t|)\}$. For a given $x\in\mathbb{R}$, the TWIN estimator $\widehat{\beta}(x)$ is the solution of

$$\widehat{\beta}(x) = \text{sgn}(x)\left(|x|-P'_{\lambda,\tau}(|\widehat{\beta}(x)|)\right)_+.$$

Denote $\phi(\cdot)$ as the standard normal density function. By breaking the range of integration into regions $(-\infty,-\kappa),[-\kappa,\kappa],(\kappa,\infty)$, we have

$$\begin{aligned} r(\lambda,\tau;\beta_j) = &\int_{-\infty}^{-\kappa}\left[x+P'_{\lambda,\tau}(|\widehat{\beta}_j(x)|)-\beta_j\right]^2\phi(x-\beta_j)dx \\ &+\beta_j^2\int_{-\kappa}^{\kappa}\phi(x-\beta_j)dx + \int_{\kappa}^{\infty}\left[x-P'_{\lambda,\tau}(|\widehat{\beta}_j(x)|)-\beta_j\right]^2\phi(x-\beta_j)dx. \end{aligned} \tag{C.27}$$

First, we consider $\beta_j = 0$. Note that $\lambda = O(\kappa)$ and we can rewrite (C.27) as

$$\begin{aligned} r(\lambda, \tau; 0) &= 2 \int_{\kappa}^{\infty} \left[x - P'_{\lambda,\tau}(|\widehat{\beta}_j(x)|) \right]^2 \phi(x) dx \\ &\leq 4 \int_{\kappa}^{\infty} x^2 \phi(x) dx + 4 \int_{\kappa}^{\infty} \{P'_{\lambda,\tau}(|\widehat{\beta}_j(x)|)\}^2 \phi(x) dx \\ &\leq 4 \int_{\kappa}^{\infty} x^2 \phi(x) dx + 4\lambda^2 \int_{\kappa}^{\infty} \phi(x) dx \\ &= 4\kappa\phi(\kappa) \quad \text{as } \kappa \to \infty. \end{aligned}$$

where the last step is due to $\int_{\kappa}^{\infty} \phi(x) dx \leq \kappa^{-1}\phi(\kappa)$.

Then, we consider $0 < \beta_j \leq \kappa$ where (C.27) can be bounded by

$$\begin{aligned} r(\lambda, \tau; \beta_j) &\leq \beta_j^2 \int_{\beta_j - \kappa}^{\beta_j + \kappa} \phi(x) dx + 1 - \int_{-\kappa - \beta_j}^{\kappa - \beta_j} \left[x + P'_{\lambda,\tau}(\widehat{\beta}_j(x + \beta_j)) \right]^2 \phi(x) dx \\ &\leq 1 + \beta_j^2 \int_{\beta_j - \kappa}^{\infty} \phi(x) dx \leq 1 + \kappa^2. \end{aligned}$$

For the last case, we consider $\beta_j \geq \kappa$. Set $\alpha = \beta_j - \lambda \geq 0$ and define $g(\alpha) = (\kappa + \alpha)^2 \int_{\alpha}^{\infty} \phi(x) dx$. We have

$$g'(\alpha) = (\lambda + \alpha)\phi(\alpha)h(\alpha), \text{ where } h(\alpha) = 2 \left(\int_{\alpha}^{\infty} \phi(x) dx / \phi(\alpha) \right) - \kappa - \alpha,$$

and $h(0) = \sqrt{2\pi} - \lambda$. When $\lambda \geq \sqrt{2\pi}$, then $h(0) \leq 0$. By direct calculation and the fact that $\int_{\alpha}^{\infty} \phi(x) dx \leq \phi(\alpha)/\alpha$, we know that h is decreasing and hence $h(\alpha) \leq 0$ on $[0, \infty)$. Thus, $g(\alpha) \leq g(0) = \kappa^2/2$. Now, (C.27) can be bounded by

$$\begin{aligned} r(\lambda, \tau; \beta_j) &\leq \beta_j^2 \int_{\beta_j - \kappa}^{\beta_j + \kappa} \phi(x) dx + 1 - \int_{-\kappa - \beta_j}^{0} \left[x + P'_{\lambda,\tau}(\widehat{\beta}_j(x + \beta_j)) \right]^2 \phi(x) dx \\ &\leq 1 + \beta_j^2 \int_{\beta_j - \kappa}^{\infty} \phi(x) dx \leq 1 + \kappa^2/2. \end{aligned}$$

Therefore, $r(\lambda, \tau; \beta_j) \leq 1 + \kappa^2$ for any $\beta_j \neq 0$ and we can bound the risk as follows

$$\begin{aligned} r(\lambda, \tau; \boldsymbol{\beta}) &\leq (p-k) \cdot r(\lambda, \tau; 0) + k \cdot \sup_{\beta_j} r(\lambda, \tau; \beta_j) \\ &\leq 4p\kappa\phi(\kappa) + k(\kappa^2 + 1). \end{aligned} \tag{C.28}$$

Let $\kappa = (1-\epsilon)\sqrt{2\log(1/\epsilon)}$. Then $\phi(\kappa) = \phi(0)e^{-\kappa^2/2} = \phi(0)\epsilon^{(1-\epsilon)^2}$. Consequently,

$$\begin{aligned} \sup_{\|\boldsymbol{\beta}\|_0 \leq k} r(\lambda, \tau; \boldsymbol{\beta}) &\leq 4p(1-\epsilon)\sqrt{2\log(1/\epsilon)} \cdot \phi(0)\epsilon^{(1-\epsilon)^2} + k[2(1-\epsilon)^2\log(1/\epsilon) + 1] \\ &= (1-\epsilon)^2 2k\log(1/\epsilon)\left(1 + \frac{2}{\sqrt{\pi\log(1/\epsilon)}\epsilon^{1-(1-\epsilon)^2}(1-\epsilon)}\right). \end{aligned}$$

This completes the proof.

C.2.2.0.3 Proof of Theorem C.3 The proof here is the same as the proof for Theorem C.2 except that we should let $\min_{t\in\mathbb{R}}\{|t| + P'_{\lambda,\tau}(|t|)\} = \sigma\sqrt{2\log(p/k)}$ in the risk upper bound (C.28).

C.2.3 Proofs for Section 4.4

In this section, we present proofs for the estimation properties of TWIN under the random Gaussian designs discussed in Section 4.4. The proofs for results are organized in the following order: Theorem 4.11, Theorem 4.12, and Theorem 4.14.

C.2.3.0.1 Proof of Theorem 4.11 From the scale invariance, we only need to prove for $\sigma = 1$. By the max-min inequality , we have for any $\vartheta > 0$,

$$\infty_{\widetilde{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \leq k} \mathbb{P}\left\{\frac{\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k\log(1/\epsilon)} > 1 - \vartheta\right\} \geq \sup_{\|\Pi\|_0 \leq k} \infty_{\widetilde{\boldsymbol{\beta}}} \mathbb{P}_\Pi\left\{\frac{\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k\log(1/\epsilon)} > 1 - \vartheta\right\}. \tag{C.29}$$

Here, Π denotes the prior on $\mathbb{R}^p$ such that any realization $\boldsymbol{\beta}$ obeys $\|\boldsymbol{\beta}\|_0 \leq k$, and $\mathbb{P}_\Pi\{\cdot\}$ denotes that $\boldsymbol{\beta}$ follows the prior Π. As in the proof of Theorem C.1, we assume that p/k is

an integer and then decompose $\{1, \ldots, p\}$ into k blocks:

$$\{1, \ldots, 1/\epsilon\}, \ \{1/\epsilon + 1, \ldots, 2/\epsilon\}, \ \ldots, \ \{(k-1)/\epsilon + 1, \ldots, k/\epsilon\}.$$

Let L_j be the loss from coordinates on the jth block $\{(j-1)/\epsilon + 1, \ldots, j/\epsilon\}$ for $j = 1, \ldots, k$. Define the prior Π such that we select a coordinate in each block at uniformly random and set the selected coordinate's amplitude to

$$t_\epsilon = \sqrt{2\log(1/\epsilon)} - \log\sqrt{2\log(1/\epsilon)}. \tag{C.30}$$

For any $\vartheta > 0$, there exists ϵ_0 such that for any $0 < \epsilon < \epsilon_0$,

$$2(1-\vartheta)\log(1/\epsilon) \leq (1-\vartheta/2)t_\epsilon^2.$$

From (C.29), it suffices to derive the following result in order to complete the proof for Theorem 4.11:

$$\sup_{\|\Pi\|_0 \leq k} \infty_{\widetilde{\boldsymbol{\beta}}} \mathbb{P}_\Pi \left\{ \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/(kt_\epsilon^2) > 1 - \vartheta/2 \right\} = 1. \tag{C.31}$$

We first prove that for any $\vartheta > 0$,

$$\sup_{0 \leq \epsilon \leq 1} \infty_{\widetilde{\boldsymbol{\beta}}} \mathbb{P}_\Pi \left\{ L_1/t_\epsilon^2 > 1 - \vartheta/2 \right\} = 1. \tag{C.32}$$

Let I be the index in $\{1, \ldots, 1/\epsilon\}$ whose amplitude is set to t_ϵ by Π. Note that $L_1 = \sum_{i=1}^{1/\epsilon} \widetilde{\beta}_i^2 + t_\epsilon^2 - 2t_\epsilon\widetilde{\beta}_I$, then

$$L_1 \leq (1-\vartheta/2)t_\epsilon^2 \iff \widetilde{\beta}_I \geq \frac{2\sum_{i=1}^{1/\epsilon}\widetilde{\beta}_i^2 + \vartheta t_\epsilon^2}{4t_\epsilon}. \tag{C.33}$$

Denote $\mathcal{D}$ as the set of indices $i \in \{1, \ldots, 1/\epsilon\}$ such that $\widetilde{\beta}_i \geq (2\sum_{i=1}^{1/\epsilon}\widetilde{\beta}_i^2 + \vartheta t_\epsilon^2)/(4t_\epsilon)$. Let $\widetilde{\beta}_{\min}$ be the minimum value of these $\widetilde{\beta}_i$ and then

$$\widetilde{\beta}_{\min} \geq \frac{2\sum_{i=1}^{1/\epsilon}\widetilde{\beta}_i^2 + \vartheta t_\epsilon^2}{4t_\epsilon} \geq \frac{2|\mathcal{D}|\widetilde{\beta}_{\min}^2 + \vartheta t_\epsilon^2}{4t_\epsilon} \geq \frac{2\sqrt{2|\mathcal{D}|\widetilde{\beta}_{\min}^2 \cdot \vartheta t_\epsilon^2}}{4t_\epsilon},$$

which gives $|\mathcal{D}| \leq 2\vartheta^{-1}$. Observe that we can rewrite the linear equation as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z} = \boldsymbol{X}^{(1)}\boldsymbol{\beta}^{(1)} + \boldsymbol{X}^{-(1)}\boldsymbol{\beta}^{-(1)} + \boldsymbol{z},$$

where $\boldsymbol{X}^{(1)}$ and $\boldsymbol{\beta}^{(1)}$ are the first $1/\epsilon$ columns of $\boldsymbol{X}$ and $\boldsymbol{\beta}$, respectively. Then

$$\boldsymbol{X}^{-(1)}\boldsymbol{\beta}^{-(1)} + \boldsymbol{z} \sim N(0, (t_\epsilon^2(k-1)/n + 1)I_n),$$

and $\boldsymbol{X}^{-(1)}\boldsymbol{\beta}^{-(1)} + \boldsymbol{z}$ is independent of $\boldsymbol{X}^{(1)}$ and $\boldsymbol{\beta}^{(1)}$. Since $t_\epsilon^2(k-1)/n = t_\epsilon^2\epsilon/\delta + o_{\mathbb{P}}(t_\epsilon^2)$, we can write

$$\boldsymbol{y} = \boldsymbol{X}^{(1)}\boldsymbol{\beta}^{(1)} + (t_\epsilon\sqrt{\epsilon/\delta} + 1) \cdot \widetilde{\boldsymbol{z}}, \text{ where } \widetilde{\boldsymbol{z}} \sim N(0,1).$$

By conditioning on $\boldsymbol{X}$ and $\boldsymbol{y}$, then

$$\mathbb{P}_\Pi\left\{L_1/t_\epsilon^2 \leq 1 - \vartheta/2\right\} = \mathbb{P}_\Pi\left\{I \in \mathcal{D}\right\} = \frac{\sum_{i\in\mathcal{D}} \exp\{(t_\epsilon \boldsymbol{x}_i'\boldsymbol{y} - t_\epsilon^2\|\boldsymbol{x}_i\|^2/2)/(t_\epsilon\sqrt{\epsilon/\delta}+1)^2\}}{\sum_{i=1}^{1/\epsilon} \exp\{(t_\epsilon \boldsymbol{x}_i'\boldsymbol{y} - t_\epsilon^2\|\boldsymbol{x}_i\|^2/2)/(t_\epsilon\sqrt{\epsilon/\delta}+1)^2\}},$$

which implies that $\mathbb{P}_\Pi\left(L_1/t_\epsilon^2 \leq (1-\vartheta/2)\right)$ is maximized if $\mathcal{D}$ is the set of indices i corresponding to the largest values of $(t_\epsilon \boldsymbol{x}_i'\boldsymbol{y} - t_\epsilon^2\|\boldsymbol{x}_i\|^2/2)/(t_\epsilon\sqrt{\epsilon/\delta}+1)^2$. Hence,

$$\mathbb{P}_\Pi\left\{L_1/t_\epsilon^2 \leq 1 - \vartheta/2\right\} \leq \mathbb{P}_\Pi\left\{\boldsymbol{x}_I'\boldsymbol{y} - t_\epsilon\|\boldsymbol{x}_I\|^2/2 \text{ is at least the } [2\vartheta^{-1}]\text{th largest}\right\}. \tag{C.34}$$

We now study that right-hand side of (C.34). Without loss of generality, let $I = 1$. Observe that for $i \neq 1$,

$$\begin{aligned}
&(\boldsymbol{x}_i'\boldsymbol{y} - t_\epsilon\|\boldsymbol{x}_i\|^2/2) - (\boldsymbol{x}_1'\boldsymbol{y} - t_\epsilon\|\boldsymbol{x}_1\|^2/2) \\
&= [\boldsymbol{x}_i'((t_\epsilon\sqrt{\epsilon/\delta}+1)\cdot\widetilde{\boldsymbol{z}} + t_\epsilon\boldsymbol{x}_1) - t_\epsilon\|\boldsymbol{x}_i\|^2/2] - [(t_\epsilon\sqrt{\epsilon/\delta}+1)\boldsymbol{x}_1'\widetilde{\boldsymbol{z}} + t_\epsilon\|\boldsymbol{x}_1\|^2/2].
\end{aligned}$$

Denote $C_1 \equiv= \boldsymbol{x}_1'\tilde{\boldsymbol{z}}$ and $\|\boldsymbol{x}_1\|^2 \equiv C_1/\sqrt{n}$, then

$$(t_\epsilon\sqrt{\epsilon/\delta}+1)\boldsymbol{x}_1'\tilde{\boldsymbol{z}} + t_\epsilon\|\boldsymbol{x}_1\|^2/2 = (t_\epsilon\sqrt{\epsilon/\delta}+1)\cdot C_1 + (1 + C_2/\sqrt{n})t_\epsilon/2. \tag{C.35}$$

Note that

$$\begin{aligned}&\boldsymbol{x}_i'((t_\epsilon\sqrt{\epsilon/\delta}+1)\cdot\widetilde{\boldsymbol{z}}+t_\epsilon\boldsymbol{x}_1)-t_\epsilon\|\boldsymbol{x}_1\|^2/2\\&=_d\|(t_\epsilon\sqrt{\epsilon/\delta}+1)\cdot\widetilde{\boldsymbol{z}}+t_\epsilon\boldsymbol{x}_1\|x_{i,1}-t_\epsilon x_{i,1}^2/2-t_\epsilon\|\boldsymbol{x}_{i,-1}\|^2/2,\end{aligned}\tag{C.36}$$

where $\boldsymbol{x}_{i,-1}$ is $\boldsymbol{x}_i$ without the entry $x_{i,1}$. Now we study the terms on the right-hand side of (C.36) separately. First, since $\|\boldsymbol{x}_{i,-1}\|^2=x_{i,2}^2+\cdots+x_{i,n}^2\to_p 1$, we know that with probability approaching one,

$$\left|\frac{\#\{2\le i\le 1/\epsilon:\|\boldsymbol{x}_{i,-1}\|\le 1\}}{1/\epsilon-1}-\frac{1}{2}\right|\le\frac{c_1}{2}\sqrt{\frac{\epsilon}{1-\epsilon}},\tag{C.37}$$

where c_1 is some positive constant. Using the normal approximation, we know that, for example, if $c_1=3$, then (C.37) holds with probability equals to 99.7%

Second, observe that

$$\begin{aligned}&\mathbb{P}\left\{\max_{1\le i\le 1/\epsilon}x_{i,1}^2\ge\frac{2\log(1/\epsilon)}{n}\right\}\le\frac{1}{2}-\frac{1}{2}\left[1-2\left\{1-\Phi\left(\sqrt{2\log(1/\epsilon)}\right)\right\}\right]^{1/\epsilon}\\&\le\frac{1}{\epsilon}\left[1-\Phi\left(\sqrt{2\log(1/\epsilon)}\right)\right]\le\frac{1}{\epsilon}\frac{\phi(\sqrt{2\log(1/\epsilon)})}{\sqrt{2\log(1/\epsilon)}}=\frac{1}{\sqrt{2\pi}\sqrt{2\log(1/\epsilon)}},\end{aligned}$$

which implies

$$\max_{1\le i\le 1/\epsilon}x_{i,1}^2\le\frac{2\log(1/\epsilon)}{n}\text{ holds with probability }\ge 1-[2\sqrt{\pi\log(1/\epsilon)}]^{-1}.\tag{C.38}$$

Third, it is clear that

$$\begin{aligned}&\|(t_\epsilon\sqrt{\epsilon/\delta}+1)\cdot\widetilde{\boldsymbol{z}}+t_\epsilon\boldsymbol{x}_1\|\ge\|(t_\epsilon\sqrt{\epsilon/\delta}+1)\cdot\widetilde{\boldsymbol{z}}\|-t_\epsilon\|\boldsymbol{x}_1\|\\&\quad\ge(t_\epsilon\sqrt{\epsilon/\delta}+1)\cdot\sqrt{n}-t_\epsilon\qquad\text{holds with probability one.}\end{aligned}\tag{C.39}$$

Combining (C.37), (C.38), and (C.39), with probability at least $1-[2\sqrt{\pi\log(1/\epsilon)}]^{-1}$, in order to prove the supreme of the right-hand side of (C.34) over $0\le\epsilon\le 1$ is zero, it suffices to

show that $\sup_{0\le\epsilon\le 1}\mathbb{P}\left\{Q\le 2\vartheta^{-1}\right\}=0$, where

$$Q\equiv\#\left\{2\le i\le\frac{1/\epsilon-1}{2}+\frac{c_1\sqrt{1/\epsilon-1}}{2}:\right.$$
$$\left.\sqrt{n}x_{i,1}>\left[t_\epsilon+\frac{t_\epsilon\log(1/\epsilon)}{n}+(t_\epsilon\sqrt{\epsilon/\delta}+1)C_1+\frac{C_2t_\epsilon}{\sqrt{n}}\right]\Big/\left[(t_\epsilon\sqrt{\epsilon/\delta}+1)-t_\epsilon/\sqrt{n}\right]\right\}.$$

As $n\to\infty$,

$$\left[t_\epsilon+\frac{t_\epsilon\log(1/\epsilon)}{n}+(t_\epsilon\sqrt{\epsilon/\delta}+1)C_1+\frac{C_2t_\epsilon}{\sqrt{n}}\right]\Big/\left[(t_\epsilon\sqrt{\epsilon/\delta}+1)-t_\epsilon/\sqrt{n}\right]$$
$$\to(\sqrt{\epsilon/\delta}+t_\epsilon^{-1})^{-1}+C_1,$$

where $C_1\equiv\boldsymbol{x}_1'\tilde{\boldsymbol{z}}$ is defined in (C.35). Let $\xi_i\equiv\sqrt{n}x_{i,1}\sim N(0,1)$. Then for any $i\ge 1$,

$$\mathbb{P}\left\{\xi_i>(\sqrt{\epsilon/\delta}+t_\epsilon^{-1})^{-1}+C_1\right\}=1-\Phi\left(\frac{t_\epsilon/\sqrt{2}}{t_\epsilon\sqrt{\epsilon/\delta}+1}\right)$$
$$\ge c\frac{t_\epsilon\sqrt{\epsilon/\delta}+1}{t_\epsilon/\sqrt{2}}\cdot\exp\left(-\frac{1}{2}\left(\frac{t_\epsilon/\sqrt{2}}{t_\epsilon\sqrt{\epsilon/\delta}+1}\right)^2\right)$$

for some constant c, and we have

$$\begin{aligned}&\sup_{0\le\epsilon\le 1}\left\{\mathbb{P}\left\{\xi_i>(\sqrt{\epsilon/\delta}+t_\epsilon^{-1})^{-1}+C_1\right\}\cdot\epsilon^{-1}\right\}\\&\quad=\sup_{0\le\epsilon\le 1}\left\{\left[1-\Phi\left(\sqrt{\delta/(2\epsilon)}\right)\right]\cdot\epsilon^{-1}\right\}=+\infty.\end{aligned}\tag{C.40}$$

Observe that

$$\begin{aligned}
&\sup_{0\le\epsilon\le 1}\mathbb{P}\left\{\#\left\{2\le i\le\frac{1/\epsilon-1}{2}+\frac{c_1\sqrt{1/\epsilon-1}}{2}:\xi_i>\sqrt{\delta/\epsilon}+C_1\right\}>\frac{2}{\vartheta}\right\}\\
&\ge 1-\infty_{0\le\epsilon\le 1}\Phi\left(\frac{\frac{2}{\vartheta}-\mathbb{P}(\xi_2>\sqrt{\delta/\epsilon}+C_1)(\frac{1/\epsilon-3}{2}+\frac{c_1\sqrt{1/\epsilon-1}}{2})}{\sqrt{\mathbb{P}(\xi_2>\sqrt{\delta/\epsilon}+C_1)[1-\mathbb{P}(\xi_2>\sqrt{\delta/\epsilon}+C_1)]\cdot(\frac{1/\epsilon-3}{2}+\frac{c_1\sqrt{1/\epsilon-1}}{2})}}\right)\\
&\ge \sup_{0\le\epsilon\le 1}\Phi\left(\sqrt{\mathbb{P}(\xi_2>\sqrt{\delta/\epsilon}+C_1)\cdot(\frac{1/\epsilon-3}{2}+\frac{c_1\sqrt{1/\epsilon-1}}{2})}\right.\\
&\qquad\qquad\left.-\frac{2/\vartheta}{\sqrt{\mathbb{P}(\xi_2>\sqrt{\delta/\epsilon}+C_1)\cdot(\frac{1/\epsilon-3}{2}+\frac{3\sqrt{1/\epsilon-1}}{2})}}\right)\\
&=1
\end{aligned}$$

where the last step is by (C.40). Hence, $\sup_{0\le\epsilon\le 1}\mathbb{P}\{Q\le 2\vartheta^{-1}\}=0$ and by (C.34), we complete the proof for (C.32).

Now we complete the proof for (C.31) by observing that

$$\begin{aligned}
\mathbb{P}_\Pi\left\{\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2/(kt_\epsilon^2)>1-\vartheta/2\right\}&\ge 1-\left[\mathbb{P}_\Pi\left\{L_1/t_\epsilon^2\le 1-\vartheta/2\right\}\right]^k\\
&\ge 1-k\left[1-\mathbb{P}_\Pi\left\{L_1/t_\epsilon^2>1-\vartheta/2\right\}\right]\\
&=k\mathbb{P}_\Pi\left\{L_1/t_\epsilon^2>1-\vartheta/2\right\}-(k-1),
\end{aligned}\tag{C.41}$$

and $\|\Pi\|_0\le k$ is equivalent to $0\le\epsilon\le 1$. Then, we obtain (C.31) by applying (C.32) to (C.41). This completes the proof.

C.2.3.0.2 Proof of Theorem 4.12 Define

$$\widetilde{\boldsymbol{\beta}}_{A^o}=\underset{\boldsymbol{b}\in\mathbb{R}^k}{\operatorname{argmin}}\left\{\frac{1}{2}||\boldsymbol{\beta}_{A^o}+\boldsymbol{X}'_{A^o}\boldsymbol{z}-\boldsymbol{b}||^2+\sum_{j=1}^{k}P_{\lambda,\tau}(|b_j|)\right\}.$$

For any $j\in A^o$, $\widetilde{\beta}_{A^o,j}=\beta_j+\boldsymbol{X}'_j\boldsymbol{z}-P'_{\lambda,\tau}(|\widetilde{\beta}_{A^o,j}|)\text{sgn}(\widetilde{\beta}_{A^o,j})$ and

$$|\widetilde{\beta}_{A^o,j}-\beta_j|=|\boldsymbol{X}'_j\boldsymbol{z}-P'_{\lambda,\tau}(|\widetilde{\beta}_{A^o,j}|)\text{sgn}(\widetilde{\beta}_{A^o,j})|\le|\boldsymbol{X}'_j\boldsymbol{z}|+|P'_{\lambda,\tau}(|\widetilde{\beta}_{A^o,j}|)|.$$

Let $\tau_\lambda \equiv \tau/\lambda \geq (1-\delta^{-1/2}-\vartheta)^{-2}$. As shown in the proof of Theorem 4.5, the TWIN-a estimator is unique with probability at least $1-e^{-n\vartheta^2/2}$. By (C.21), since $\lambda = \{[(1-\vartheta)\sqrt{\delta/\epsilon}-1]^{-1}(1+\vartheta)+1\}(1+\vartheta)\sigma\sqrt{2\log p}$, we have

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = \|\widehat{\boldsymbol{\beta}}_{A^o} - \boldsymbol{\beta}_{A^o}\|,$$

and

$$\mathbb{P}\{|P'_{\lambda,\tau}(|\widetilde{\beta}_{A^o,j}|)| \leq [(\sqrt{\delta/\epsilon}-1)^{-1}+1]\sigma\sqrt{2\log p}\} \to 1.$$

From the proof of Theorem 4.5, we know $|\boldsymbol{X}_j'\boldsymbol{z}| = o_{\mathbb{P}}\{[(\sqrt{\delta/\epsilon}-1)^{-1}+1]\sigma\sqrt{2\log p}\}$ and then

$$\mathbb{P}\left\{\frac{\|\widetilde{\boldsymbol{\beta}}_{A^o} - \boldsymbol{\beta}_{A^o}\|^2}{[(\sqrt{\delta/\epsilon}-1)^{-1}+1]^2 2\sigma^2 k\log p} \leq 1\right\} \to 1. \tag{C.42}$$

Now we show that

$$\mathbb{P}\left\{\frac{\|\widetilde{\boldsymbol{\beta}}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2}{\frac{3[(\sqrt{\delta/\epsilon}-1)^{-1}+1]^2}{(1-\tau_\lambda^{-1})\sqrt{\delta/\epsilon}-2} 2\sigma^2 k\log p} \leq 1\right\} \to 1. \tag{C.43}$$

By definition of $\widehat{\boldsymbol{\beta}}_{A^o}$ and $\widetilde{\boldsymbol{\beta}}_{A^o}$, they minimize the following $L_1(\boldsymbol{b})$ and $L_2(\boldsymbol{b})$ with $\boldsymbol{b}\in\mathbb{R}^k$, respectively,

$$\begin{aligned} L_1(\boldsymbol{b}) &:= \frac{1}{2}\|\boldsymbol{X}_{A^o}(\boldsymbol{\beta}_{A^o}-\boldsymbol{b})\|^2 + \boldsymbol{z}'\boldsymbol{X}_{A^o}(\boldsymbol{\beta}_{A^o}-\boldsymbol{b}) + \sum_{j=1}^k P_{\lambda,\tau}(|b_j|), \\ L_2(\boldsymbol{b}) &:= \frac{1}{2}\|\boldsymbol{\beta}_{A^o}-\boldsymbol{b}\|^2 + \boldsymbol{z}'\boldsymbol{X}_{A^o}(\boldsymbol{\beta}_{A^o}-\boldsymbol{b}) + \sum_{j=1}^k P_{\lambda,\tau}(|b_j|). \end{aligned} \tag{C.44}$$

By Lemma C.4, we know that all the eigenvalues of $\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o}$ lie in $(1-\sqrt{\epsilon/\delta}, 1+\sqrt{\epsilon/\delta})$

with overwhelming probability. Thus,

$$L_2(\widetilde{\boldsymbol{\beta}}_{A^o}) - \frac{\sqrt{\epsilon/\delta}}{2}\|\boldsymbol{\beta}_{A^o} - \widetilde{\boldsymbol{\beta}}_{A^o}\|^2 \le L_1(\widetilde{\boldsymbol{\beta}}_{A^o}) \le L_2(\widetilde{\boldsymbol{\beta}}_{A^o}) + \frac{\sqrt{\epsilon/\delta}}{2}\|\boldsymbol{\beta}_{A^o} - \widetilde{\boldsymbol{\beta}}_{A^o}\|^2,$$
$$L_2(\widehat{\boldsymbol{\beta}}_{A^o}) - \frac{\sqrt{\epsilon/\delta}}{2}\|\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2 \le L_1(\widehat{\boldsymbol{\beta}}_{A^o}) \le L_2(\widehat{\boldsymbol{\beta}}_{A^o}) + \frac{\sqrt{\epsilon/\delta}}{2}\|\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2.$$

Thus,

$$\begin{aligned} L_2(\widetilde{\boldsymbol{\beta}}_{A^o}) + \frac{\sqrt{\epsilon/\delta}\|\boldsymbol{\beta}_{A^o} - \widetilde{\boldsymbol{\beta}}_{A^o}\|^2}{2} &\ge L_1(\widetilde{\boldsymbol{\beta}}_{A^o}) \ge L_1(\widehat{\boldsymbol{\beta}}_{A^o}) \\ &\ge L_2(\widehat{\boldsymbol{\beta}}_{A^o}) - \frac{\sqrt{\epsilon/\delta}\|\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2}{2}. \end{aligned} \tag{C.45}$$

Note that $\tau_\lambda \ge (1 - \delta^{-1/2} - \vartheta)^{-2} > 1$, then L_2 is strongly convex and

$$L_2(\widehat{\boldsymbol{\beta}}_{A^o}) \ge L_2(\widetilde{\boldsymbol{\beta}}_{A^o}) + \frac{(1 - \tau_\lambda^{-1})\|\widetilde{\boldsymbol{\beta}}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2}{2}.$$

Plugging the above formula into the right-hand side of (C.45) gives

$$\frac{(1 - \tau_\lambda^{-1})\|\widetilde{\boldsymbol{\beta}}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2}{2} - \frac{\sqrt{\epsilon/\delta}\|\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2}{2} \le \frac{\sqrt{\epsilon/\delta}\|\boldsymbol{\beta}_{A^o} - \widetilde{\boldsymbol{\beta}}_{A^o}\|^2}{2}.$$

Since $\sqrt{\epsilon/\delta}\|\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2/2 \le \sqrt{\epsilon/\delta}\|\widetilde{\boldsymbol{\beta}}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2 + \sqrt{\epsilon/\delta}\|\boldsymbol{\beta}_{A^o} - \widetilde{\boldsymbol{\beta}}_{A^o}\|^2$, we have

$$\|\widetilde{\boldsymbol{\beta}}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2 \le \frac{3\sqrt{\epsilon/\delta}}{1 - \tau_\lambda^{-1} - 2\sqrt{\epsilon/\delta}}\|\boldsymbol{\beta}_{A^o} - \widetilde{\boldsymbol{\beta}}_{A^o}\|^2.$$

Together with (C.42), then (C.43) follows. Now combining (C.42) with (C.43), we complete the proof for this theorem.

C.2.3.0.3 Proof of Corollary 4.13 Since $\sqrt{\delta/\epsilon} - 1 > \sqrt{\delta}$ for $\epsilon \le 1/4$ and now $\lambda = \lambda_{a,\text{univ}} = (1 + \delta^{-1/2})\sigma\sqrt{2\log p}$, we have $\mathbb{P}\{|P'_{\lambda,\tau}(|\widetilde{\beta}_{A^o,j}|)| \le (1 + \delta^{-1/2})\sigma\sqrt{2\log p}\} \to 1$. Following the proof of Theorem 4.12, we obtain that with probability approaching one,

$$\|\widetilde{\boldsymbol{\beta}}_{A^o} - \boldsymbol{\beta}_{A^o}\|^2 \le (1 + \delta^{-1/2})^2 2\sigma^2 k\log p,$$
$$\|\widetilde{\boldsymbol{\beta}}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2 \le \frac{3(1 + \delta^{-1/2})^2}{(1 - \lambda_{a,\text{univ}}\tau_{a,\text{univ}}^{-1})\sqrt{\delta/\epsilon} - 2} 2\sigma^2 k\log p.$$

Then,

$$\begin{aligned}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2 &= \|\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2 \\ &\leq \left\{ \frac{\sqrt{3}}{[(1-\lambda_{a,\text{univ}}\tau_{a,\text{univ}}^{-1})\delta^{1/2}\epsilon^{-1/2}-2]^{1/2}} + 1 \right\}^2 (1+\delta^{-1/2})^2 2\sigma^2 k \log p \\ &\leq \left\{ \frac{\sqrt{3}}{[(1.98-\delta^{-1/2})\epsilon^{-1/2}-2]^{1/2}} + 1 \right\}^2 (1+\delta^{-1/2})^2 2\sigma^2 k \log p,\end{aligned}$$

which completes the proof.

C.2.3.0.4 Proof of Theorem 4.14 We follow the proof of Theorem 4.12. By (C.9), $\mathbb{P}\{|P'_{\lambda,\tau}(|\widetilde{\beta}_{A^o,j}|)| \leq \sigma\sqrt{1-\epsilon/\delta}\sqrt{2\log p}\} \to 1$. Recall that $L_1(\boldsymbol{b})$ and $L_2(\boldsymbol{b})$ are defined in (C.44). Note that $L_2(\cdot)$ is strongly convex on a open connected neighborhood containing $\widehat{\boldsymbol{\beta}}_{A^o}$ and $\widetilde{\boldsymbol{\beta}}_{A^o}$ where $P_{\lambda,\tau}(\cdot) \equiv 0$. This implies

$$L_2(\widehat{\boldsymbol{\beta}}_{A^o}) \geq L_2(\widetilde{\boldsymbol{\beta}}_{A^o}) + \frac{\|\widetilde{\boldsymbol{\beta}}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2}{2}.$$

Plugging the above formula into the right-hand side of (C.45) gives

$$\frac{\|\widetilde{\boldsymbol{\beta}}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2}{2} - \frac{\sqrt{\epsilon/\delta}\|\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2}{2} \leq \frac{\sqrt{\epsilon/\delta}\|\boldsymbol{\beta}_{A^o} - \widetilde{\boldsymbol{\beta}}_{A^o}\|^2}{2}.$$

Thus,

$$\|\widetilde{\boldsymbol{\beta}}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2 \leq \frac{3\sqrt{\epsilon/\delta}}{1-2\sqrt{\epsilon/\delta}}\|\boldsymbol{\beta}_{A^o} - \widetilde{\boldsymbol{\beta}}_{A^o}\|^2.$$

The rest of proof follows the idea in the proof of Theorem 4.12.

C.2.3.0.5 Proof of Corollary 4.15 Since $\lambda = \lambda_{b,\text{univ}} = \sigma\sqrt{2\log p}$, we have $\mathbb{P}\{|P'_{\lambda,\tau}(|\widetilde{\beta}_{A^o,j}|)| \leq \sigma\sqrt{2\log p}\} \to 1$. Following the proof of Theorem 4.12, we obtain that with probability approaching to one,

$$\|\widetilde{\boldsymbol{\beta}}_{A^o} - \boldsymbol{\beta}_{A^o}\|^2 \leq 2\sigma^2 k\log p, \;\; \|\widetilde{\boldsymbol{\beta}}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2 \leq \frac{3}{\sqrt{\delta/\epsilon}-2} 2\sigma^2 k \log p.$$

Then,

$$\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\beta}_{A^o} - \widehat{\boldsymbol{\beta}}_{A^o}\|^2 \leq \left[\frac{\sqrt{3}}{(\delta^{-1/2}\epsilon^{-1/2} - 2)^{1/2}} + 1\right]^2 2\sigma^2 k \log p,$$

which completes the proof.

C.3 Key Lemmas

Lemma C.4. *For any $A \subset \{1, \ldots, p\}$ with $|A| < n$, the largest and smallest eigenvalues of $\boldsymbol{X}_A'\boldsymbol{X}_A$ satisfy for any $\vartheta > 0$,*

$$\lambda_{\max}(\boldsymbol{X}_A'\boldsymbol{X}_A) < (1 + \sqrt{|A|/n} + \vartheta)^2, \;\; \lambda_{\min}(\boldsymbol{X}_A'\boldsymbol{X}_A) > (1 - \sqrt{|A|/n} - \vartheta)^2$$

with probability at least $1 - e^{-n\vartheta^2/2}$.

Proof. Classical theory on Wishart matrices gives that the smallest singular value of $\boldsymbol{X}_A$ is larger than $1 - \sqrt{|A|/n} - \vartheta$ with probability at least $1 - e^{-n\vartheta^2/2}$ and the largest singular value of $\boldsymbol{X}_A$ is smaller than $1 + \sqrt{|A|/n} + \vartheta$ with probability at least $1 - e^{-n\vartheta^2/2}$; see, for example, Vershynin (2012). We complete the proof by noticing that $n/p \to \delta$. □

For ease of presentation of our proofs, we re-state in Lemmas C.5 and C.6 two lemmas presented in the Supplement to Su et al. (2017). We use these lemmas throughout our proofs.

Lemma C.5. *For any positive integer d and $\vartheta \geq 0$, we have*

$$\mathbb{P}(\sqrt{\chi_d^2} \geq \sqrt{d} + \vartheta) \leq e^{-\vartheta^2/2}.$$

Lemma C.6. *For any positive integer d and any $\vartheta \geq 0$, we have*

$$\mathbb{P}\left(\chi_d^2 \leq \vartheta d\right) \leq (e\vartheta)^{d/2}.$$

C.4 Coordinate Descent Algorithms for TWIN

In order to develop algorithms for computation of penalized regression problems with TWIN penalties, we focus on minimization of the following univariate penalized regression problem:

$$J(b) = \frac{1}{2}(z-b)^2 + P_{\lambda,\tau}(|b|), \tag{C.46}$$

where $z = \boldsymbol{x}'\boldsymbol{y}$. In the typical coordinate descent fashion, we propose to loop through each of the variables and minimize with respect to its corresponding coefficient and hence the emphasis on (C.46). A coordinate descent algorithm for the TWIN class of penalties is described in Algorithm 2. Algorithm 2 follows Algorithm 1 of Mazumder et al. (2011) for nonconvex penalties with a few modifications. It is known that a cyclic coordinate descent algorithm for penalties with discontinuous thresholding operations may not be convergent (Mazumder et al., 2011; Patrascu and Necoara, 2015). However, in practice we find with a few modifications, coordinate descent can be quite effective. In particular, similar to ideas in Patrascu and Necoara (2015) we randomize the coordinate updates instead of cycling through in a deterministic ordering of variables. Secondly, we do not take full steps in the direction of the univariate minimizers of (C.46). Instead, we only take partial steps, as guided by the parameter α in Algorithm 2. We find that choosing $\alpha = 1/2$ works well in practice. This encourages less "greedy" updates and thus the iterates are less likely to get stuck in poor local minima. Studying the theoretical properties of this approach is an interesting direction of future work. We also note that since the TWIN penalty is non-convex, two different random seeds could potentially yield different local solutions. As such, one potential optimization strategy is to run Algorithm 2 several times and choose the solution with the best loss.

Algorithm 2 works well when the sample size is not too small, however, we have found that with very small sample sizes TWIN, and in particular TWIN-b, can be unstable when τ and λ are such that TWIN is discontinuous. To mitigate instability in these scenarios, we developed a coordinate descent-based hybrid local linear approximation (LLA) algorithm

based on ideas from the local linear approximation algorithm of Zou and Li (2008). The basic idea is to construct a local linear approximation to the penalty function for small to medium sized coefficients:

$$P_{\lambda,\tau}(|b_j|) \approx \begin{cases} P_{\lambda,\tau}(|b_j|) \text{ if } b_j > \tau \text{ and } P''_{\lambda,\tau}(|b_j|) \geq 0 \\ P_{\lambda,\tau}(|\beta_j|) + P'_{\lambda,\tau}(|\beta_j|)(|b_j - |\beta_j||) \text{ otherwise, for } b_j \approx \beta_j. \end{cases}$$

At iteration k we then use this approximation to replace minimization of C.46 with minimization of

$$\widetilde{J}(b_j) = \begin{cases} \frac{1}{2}(z - b_j)^2 + P_{\lambda,\tau}(|b_j|) \text{ if } b_j > \tau \text{ and } P''_{\lambda,\tau}(|b_j|) \geq 0 \\ \frac{1}{2}(z - b_j)^2 + P'_{\lambda,\tau}(|\widetilde{\beta}_j^{(k-1)}|)|b_j| \text{ otherwise.} \end{cases} \tag{C.47}$$

The resulting algorithm, which we call the MCLLA algorithm for mixed coordinate local linear approximation, is described in Algorithm 3.

C.5 Additional Simulation Results

C.5.1 Simulation Illustrating Universal Tuning Parameters

In this section we conduct a simulation to evaluate the finite sample validity of the universal tuning parameters $\lambda_{a,\text{univ}}$, $\lambda_{b,\text{univ}}$, and τ_{univ} from Corollaries 4.6 and 4.8 from the main text. We simulate data under model (4.1) and generate $\boldsymbol{X}$ from a multivariate Gaussian distribution with identity covariance matrix. The nonzero terms in $\boldsymbol{\beta}$ are generated from independent uniform random variables on $[-2, -1] \cup [1, 2]$. The tuning parameters for both TWIN-a and TWIN-b are chosen as the universal values from Corollaries 4.6 and 4.8 from the main text. We vary the sample size n, the number of variables p, the number of active variables k, and the standard deviation σ of the error term. For implementation purposes, we take δ to be n/p instead of the limit of n/p. In this simulation we use the true data-generating noise level σ, however it can be straightforwardly estimated using the approach presented in

Algorithm 2 Coordinate descent for (4.2) with TWIN penalties

1. Input a grid of decreasing λ values $\{\lambda_0, \ldots, \lambda_{L-1}\}$ and a grid of decreasing τ values $\{\tau_0, \tau_0/2, \tau_0/2^2, \ldots, \tau_0/2^{T-1}\}$.
2. For each combination $(\tau_0/2^t, \lambda_\ell) \in \{\tau_0, \tau_0/2, \tau_0/2^2, \ldots, \tau_0/2^{T-1}\} \times \{\lambda_0, \ldots, \lambda_{L-1}\}$, repeat the following procedure:
   (i) At iteration k loop through the following univariate updates for each $j \in \mathcal{P}_k(1, \ldots, p)$

$$\widetilde{\beta}_j^{(k),\dagger} \leftarrow S_{\tau_t}\left(\sum_{i=1}^{n}(y_i - \widetilde{y}_i^j)x_{ij}, \lambda_\ell\right)$$
$$\widetilde{\beta}_j^{(k)} \leftarrow \alpha\widetilde{\beta}_j^{(k),\dagger} + (1-\alpha)\left[(1 - I(\widetilde{\beta}_j^{(k-1),\dagger} = 0, \widetilde{\beta}_j^{(k),\dagger} = 0))\widetilde{\beta}_j^{(k-1)}\right] \quad \text{(C.48)}$$

   where $\widetilde{y}_i^j = \sum_{m \neq j} x_{ik}\widetilde{\beta}_m^{(k-1)}$, $S_\tau(\widetilde{\beta}, \lambda) = \operatorname{argmin}_\beta J(\beta)$, where $J(\beta)$ is defined in (C.46), $\alpha \in (0, 1]$, and $\mathcal{P}_k(1, \ldots, p)$ are permutations of the variable indexes, until the update vectors $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)$ converge to $\boldsymbol{\beta}^*$. The term $1 - I(\widetilde{\beta}_j^{(k-1),\dagger} = 0, \widetilde{\beta}_j^{(k),\dagger} = 0)$ is to allow estimates to be exactly 0 if two successive thresholding iterates are 0. The permutations $\mathcal{P}_k(1, \ldots, p)$ may be identity mappings, i.e. $\mathcal{P}_k(1, \ldots, p) = (1, \ldots, p)$, and thus result in a repeated ordered cycling through the variables, but we find that uniformly at random permutations are more effective.
   (ii) Set $\widehat{\boldsymbol{\beta}}_{\tau_t, \lambda_\ell} \leftarrow \boldsymbol{\beta}^*$

Section 5 of Zhang (2010). This simulation is low-dimensional, however in high-dimensional scenarios, the degrees of freedom must be estimated. The approach of Theorems 7 and 8 of Zhang (2010) can be extended to TWIN for such a purpose.

The simulation is replicated 500 times. For each simulation we record the resulting FDR and TDR values. The average FDR and TDR values for simulations with $n = 1000$ are presented in Table C.1 and for $n = 2000$ in Table C.2. In almost all settings the FDR is nearly zero and the TDR is quite high, even under the more difficult scenarios with large k and large σ. The results improve across all settings as the sample size increases. From the results, it appears that the universal tuning parameter values are conservative in terms of the FDR, rarely yielding any false discoveries.

Algorithm 3 Mixed coordinate local linear approximation descent for (4.2) with TWIN penalties

1. Input a grid of decreasing λ values $\{\lambda_0, \ldots, \lambda_{L-1}\}$ and a grid of decreasing τ values $\{\tau_0, \tau_0/2, \tau_0/2^2, \ldots, \tau_0/2^{T-1}\}$.
2. For each combination $(\tau_0/2^t, \lambda_\ell) \in \{\tau_0, \tau_0/2, \tau_0/2^2, \ldots, \tau_0/2^{T-1}\} \times \{\lambda_0, \ldots, \lambda_{L-1}\}$, repeat the following procedure:
   (i) At iteration k loop through the following univariate updates for each $j \in \mathcal{P}_k(1, \ldots, p)$

$$\widetilde{\beta}_j^{(k),\dagger} \leftarrow S_{\tau_t,\ell_1}\left(\sum_{i=1}^{n}(y_i - \widetilde{y}_i^j)x_{ij}, \lambda_\ell\right)$$
$$\widetilde{\beta}_j^{(k)} \leftarrow \alpha\widetilde{\beta}_j^{(k),\dagger} + (1-\alpha)\left[(1 - I(\widetilde{\beta}_j^{(k-1),\dagger} = 0, \widetilde{\beta}_j^{(k),\dagger} = 0))\widetilde{\beta}_j^{(k-1)}\right]$$

   where $\widetilde{y}_i^j = \sum_{m \neq j} x_{ik}\widetilde{\beta}_m^{(k-1)}$, $S_{\tau,\ell_1}(\widetilde{\beta}, \lambda) = \operatorname{argmin}_\beta \widetilde{J}(\beta)$, where $\widetilde{J}(\beta)$ is defined in (C.47), $\alpha \in (0, 1]$, and $\mathcal{P}_k(1, \ldots, p)$ are permutations of the variable indexes, until the update vectors $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)$ converge to $\boldsymbol{\beta}^*$.
   (ii) Set $\widehat{\boldsymbol{\beta}}_{\tau_t,\lambda_\ell} \leftarrow \boldsymbol{\beta}^*$

C.5.2 Additional Simulation Settings

In this section we provide additional simulation results extending the simulations from Section 5.3 of the main text. We keep the simulation settings the same as in the main text with a few changes. We increase the dimension to $p = 2000$ and use sample sizes of 250 and 500. We generate data under Models 1 and 2 and under two similar models with an increased number of active variables ($k = 100$). In this simulation we generate the covariates with a block diagonal covariance matrix where each block is size 1000 and is constructed in the same way as the full covariance matrix $\boldsymbol{\Sigma}$, i.e. with element in the ith row and jth column equal to $\rho^{|i-j|}$. In addition to Models 1 and 2 from Section 5 of the main text, we also simulate data under the following two models, Models 5 and 6, which both have 100 active variables:

Model 5 A randomly chosen 100 elements of $\boldsymbol{\beta}$ are generated as independent uniform random variables on $[-2, 0.5] \cup [0.5, 2]$ and the rest are 0.

				TWIN-a				TWIN-b			
				FDR		TDR		FDR		TDR	
n	p	k	σ	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1000	100	10	1	0.00000	0.00000	0.95700	0.06972	0.00000	0.00000	0.95740	0.06939
			3	0.00000	0.00000	0.93380	0.07855	0.00000	0.00000	0.93380	0.07855
			5	0.00000	0.00000	0.91700	0.08358	0.00000	0.00000	0.91720	0.08365
		50	1	0.00000	0.00000	0.88224	0.04798	0.00000	0.00000	0.88260	0.04782
			3	0.00000	0.00000	0.87516	0.05047	0.00000	0.00000	0.87552	0.05026
			5	0.00023	0.00225	0.85768	0.05181	0.00023	0.00225	0.85812	0.05144
		100	1	0.00000	0.00000	0.81698	0.04142	0.00000	0.00000	0.81750	0.04130
			3	0.00000	0.00000	0.80868	0.04139	0.00000	0.00000	0.80910	0.04149
			5	0.00000	0.00000	0.80046	0.04237	0.00000	0.00000	0.80080	0.04217
	500	10	1	0.00000	0.00000	0.95340	0.07056	0.00000	0.00000	0.95360	0.07055
			3	0.00000	0.00000	0.94180	0.07045	0.00000	0.00000	0.94180	0.07045
			5	0.00000	0.00000	0.92480	0.08318	0.00000	0.00000	0.92480	0.08318
		50	1	0.00018	0.00205	0.88124	0.05069	0.00018	0.00205	0.88128	0.05069
			3	0.00044	0.00310	0.87604	0.05143	0.00044	0.00310	0.87604	0.05143
			5	0.00244	0.00723	0.86428	0.05627	0.00244	0.00723	0.86432	0.05633
		100	1	0.00720	0.00929	0.81198	0.04427	0.00720	0.00929	0.81200	0.04423
			3	0.00998	0.01103	0.81218	0.04226	0.01000	0.01105	0.81222	0.04230
			5	0.01958	0.01593	0.79904	0.04292	0.01958	0.01593	0.79908	0.04296

Table C.1: FDR and TDR averaged over 500 simulation replications with sample sizes $n = 1000$ and tuning parameters set as their universal values. The values in the "SD" columns are standard deviations, not standard errors.

Model 6 A randomly chosen 100 elements of $\boldsymbol{\beta}$ are $(-0.975)^{j-1}$ for $j = 1, \ldots, 100$ and the rest are 0.

The results from these simulations are consistent with the simulation results from the main text and thus we do not discuss them in-depth.

C.5.3 Extensive Evaluation of the TWIN's τ Tuning Parameter

In this simulation we present expanded results comparing TWIN with different values for the τ tuning parameter. In particular, we present simulation studies under all of the simulation settings described in Section 5 of the main text. The FDR-TPR results for TWIN-a are presented in Figures C.9, C.10, C.11, and C.12 and the RMSE-model size results for TWIN-a are presented in Figures C.13, C.14, C.15, and C.16. Results for TWIN-b mirror results for TWIN-a and are thus not included, but can be made available by contacting the authors. The FDR-TPR results for TWIN-b are presented in Figures C.17, C.18, C.19, and

				TWIN-a				TWIN-b			
				FDR		TDR		FDR		TDR	
n	p	k	σ	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2000	100	10	1	0.00000	0.00000	0.96900	0.05608	0.00000	0.00000	0.97040	0.05376
			3	0.00000	0.00000	0.95620	0.06626	0.00000	0.00000	0.95700	0.06557
			5	0.00000	0.00000	0.94360	0.06978	0.00000	0.00000	0.94480	0.06930
		50	1	0.00000	0.00000	0.92192	0.03920	0.00000	0.00000	0.92252	0.03904
			3	0.00000	0.00000	0.91928	0.04208	0.00000	0.00000	0.91976	0.04215
			5	0.00000	0.00000	0.90976	0.04200	0.00000	0.00000	0.91024	0.04187
		100	1	0.00000	0.00000	0.88312	0.03750	0.00000	0.00000	0.88374	0.03740
			3	0.00000	0.00000	0.87766	0.03623	0.00000	0.00000	0.87814	0.03618
			5	0.00000	0.00000	0.87062	0.03471	0.00000	0.00000	0.87120	0.03464
	500	10	1	0.00000	0.00000	0.97300	0.05382	0.00000	0.00000	0.97300	0.05382
			3	0.00000	0.00000	0.96040	0.05830	0.00000	0.00000	0.96040	0.05830
			5	0.00000	0.00000	0.94940	0.06594	0.00000	0.00000	0.94940	0.06594
		50	1	0.00000	0.00000	0.92344	0.04067	0.00000	0.00000	0.92348	0.04067
			3	0.00000	0.00000	0.91752	0.03801	0.00000	0.00000	0.91752	0.03801
			5	0.00000	0.00000	0.90772	0.04513	0.00000	0.00000	0.90788	0.04499
		100	1	0.00006	0.00083	0.87956	0.03699	0.00006	0.00083	0.87964	0.03701
			3	0.00007	0.00088	0.88002	0.03524	0.00007	0.00088	0.88016	0.03521
			5	0.00032	0.00224	0.87226	0.03451	0.00032	0.00224	0.87242	0.03445

Table C.2: FDR and TDR averaged over 500 simulation replications with sample sizes $n = 2000$ and tuning parameters set as their universal values. The values in the "SD" columns are standard deviations, not standard errors.

C.20. For the sake of space, results for prediction performance for TWIN-b are left out, but can be provided by contacting the authors.

C.5.4 Remaining Prediction Results from Main Text

In this section we include in Figure C.21 the prediction results under Models 3 and 4 that were excluded from the main text.

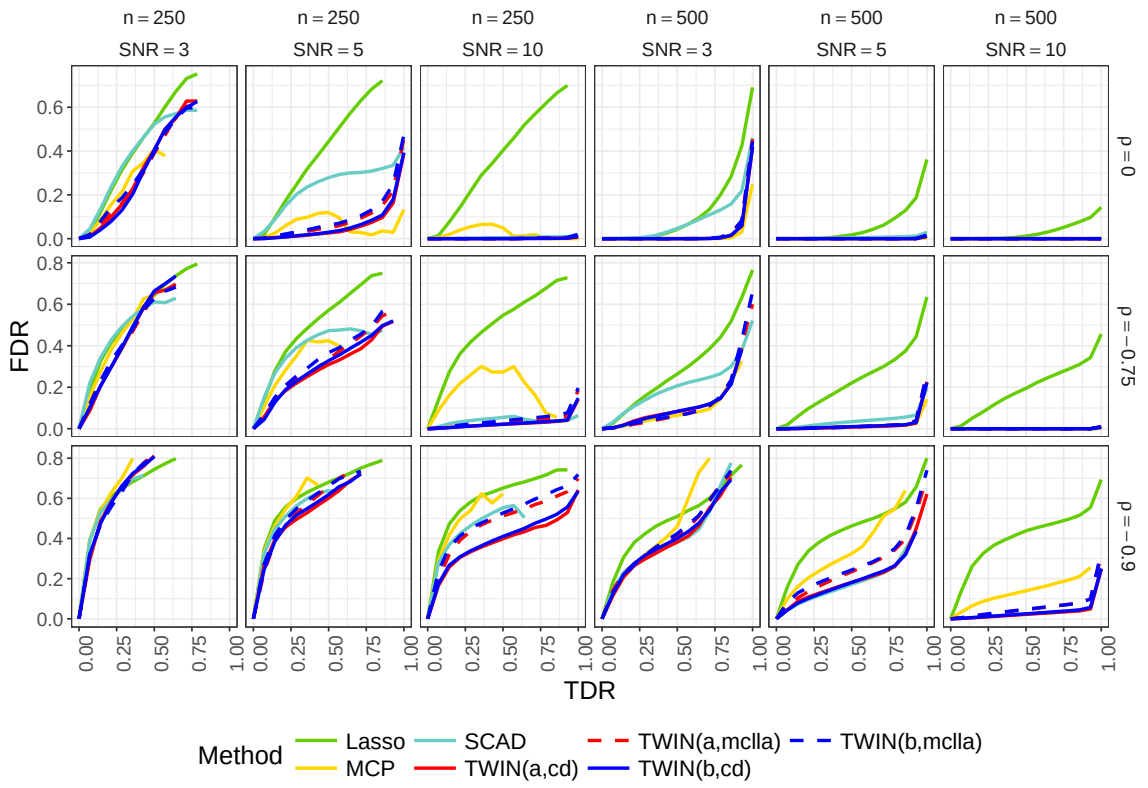

Figure C.1: The results above are for a simulation with data generated under Model 1 with $p = 2000$.

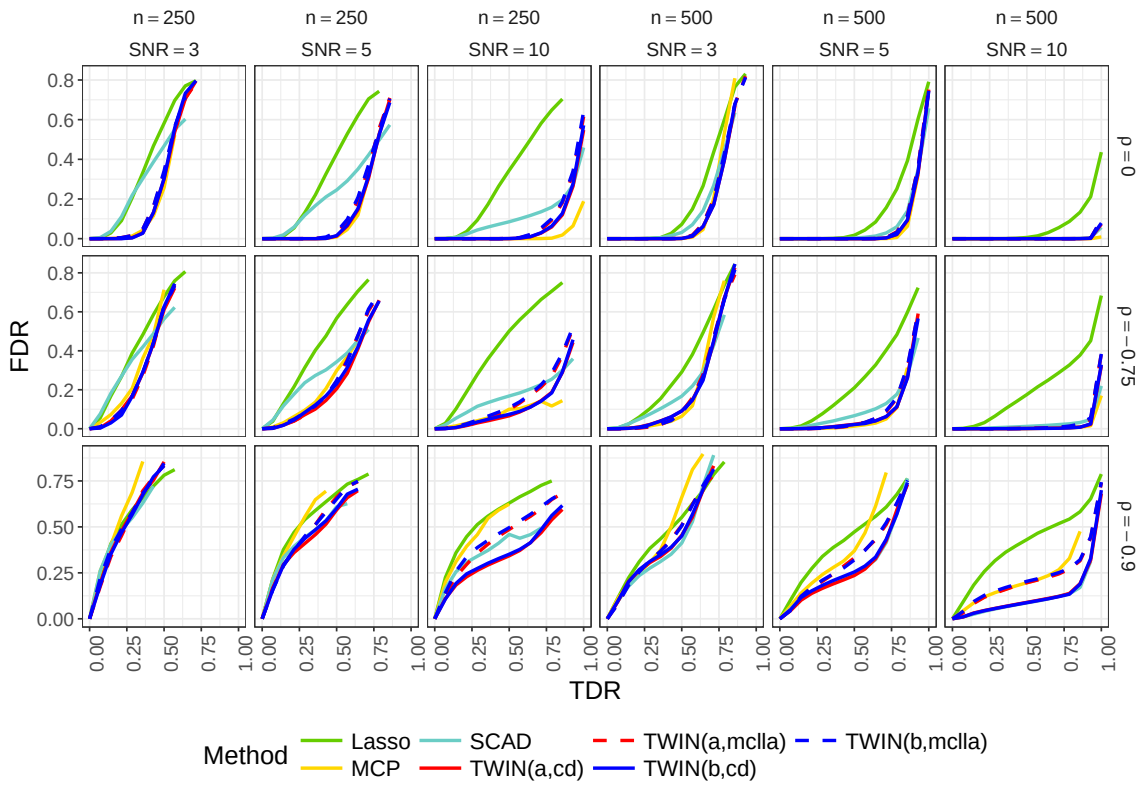

Figure C.2: The results above are for a simulation with data generated under Model 2 with $p = 2000$.

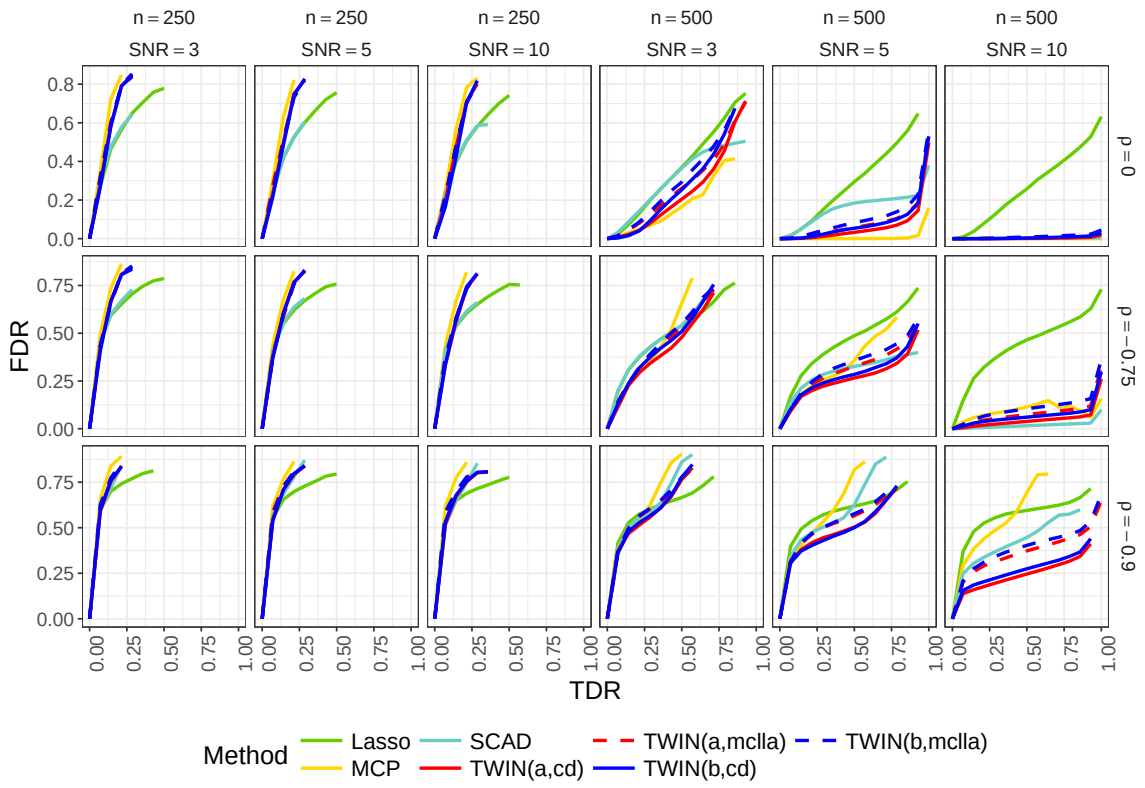

Figure C.3: The results above are for a simulation with data generated under Model 5 with $p = 2000$.

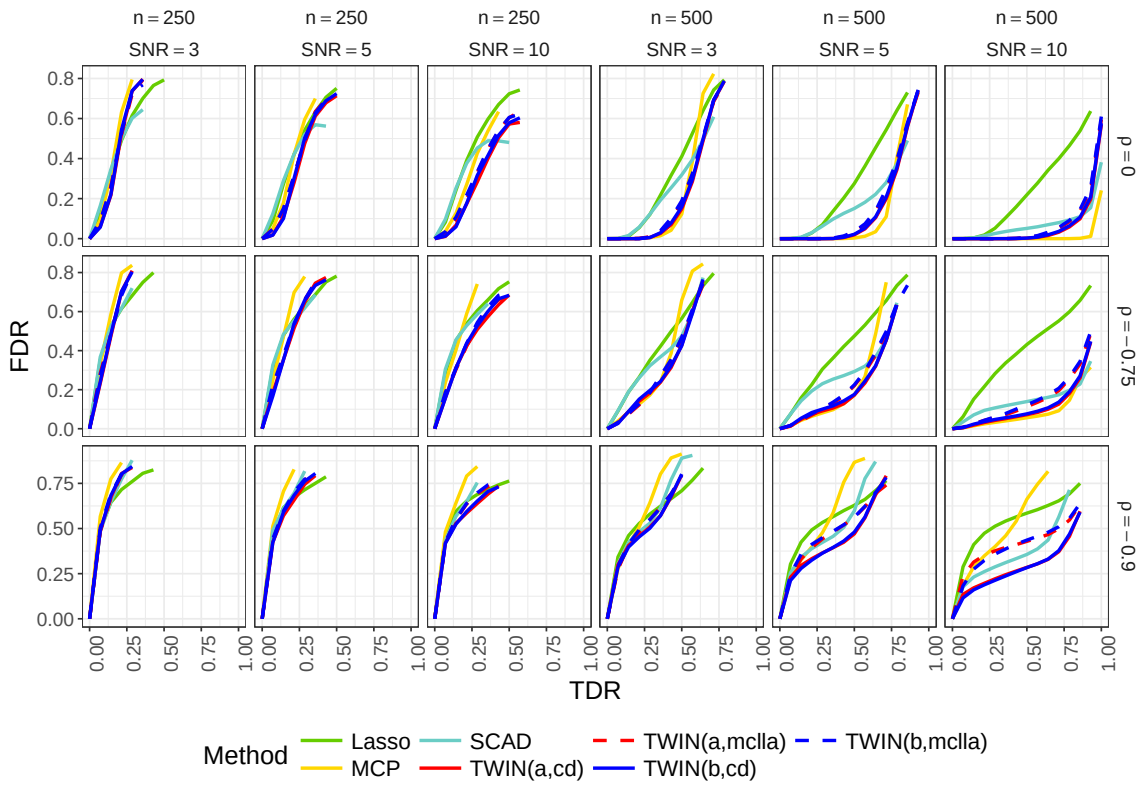

Figure C.4: The results above are for a simulation with data generated under Model 6 with $p = 2000$.

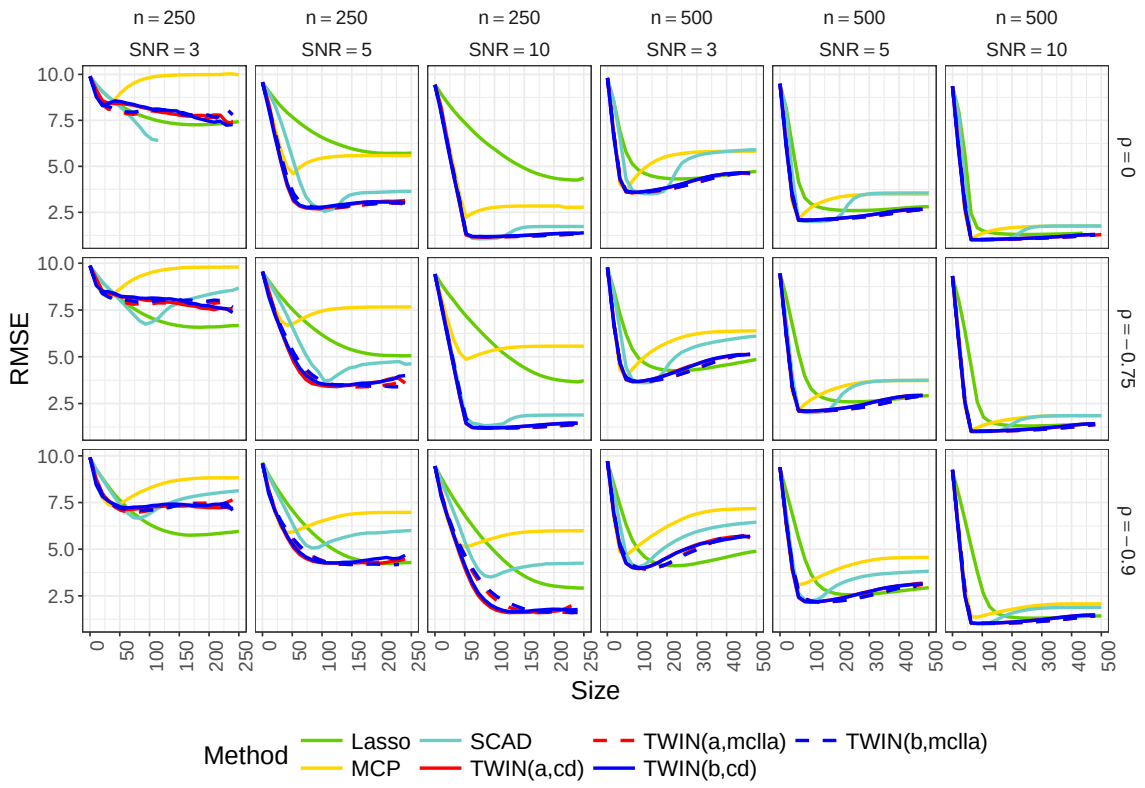

Figure C.5: The results above are for a simulation with data generated under Model 1 with $p = 2000$.

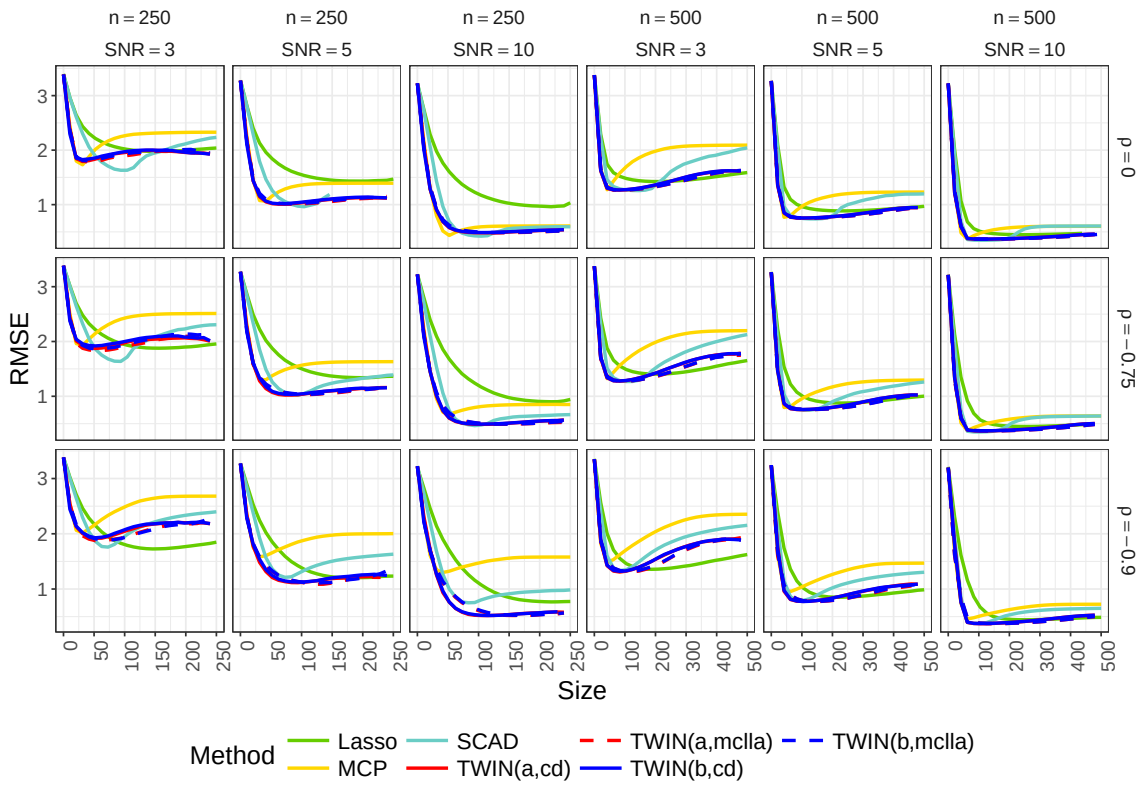

Figure C.6: The results above are for a simulation with data generated under Model 2 with $p = 2000$.

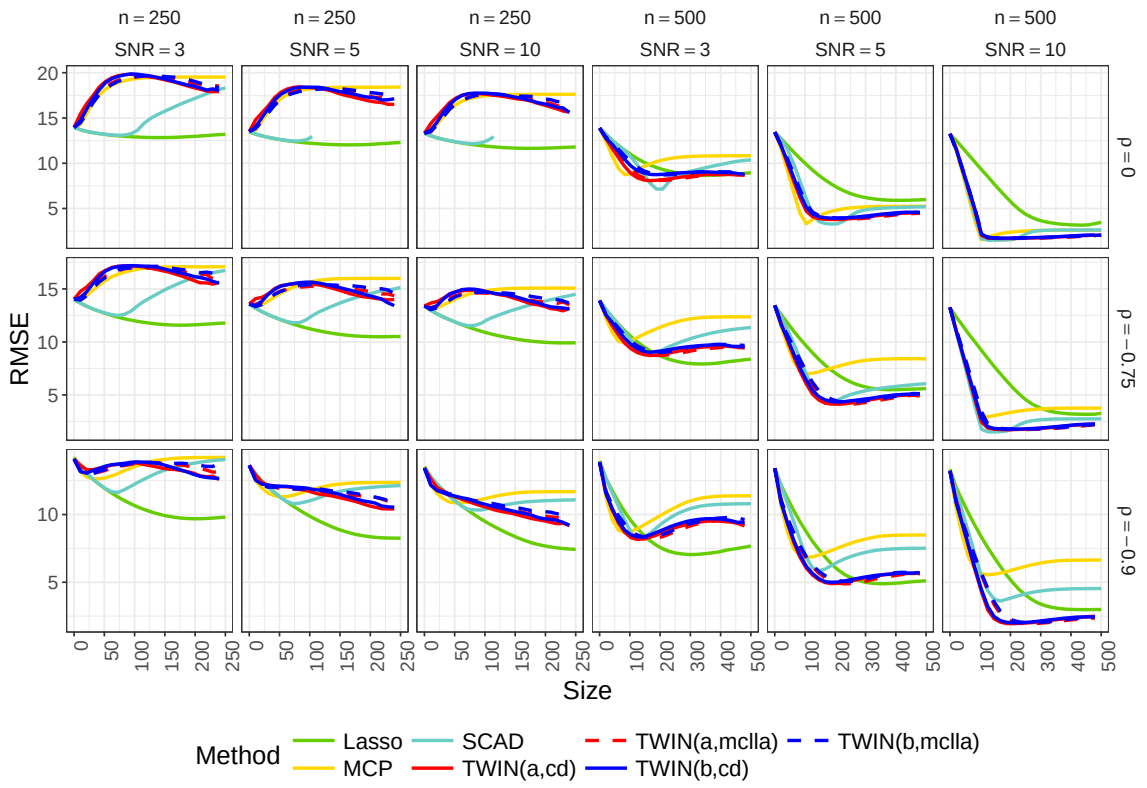

Figure C.7: The results above are for a simulation with data generated under Model 5 with $p = 2000$.

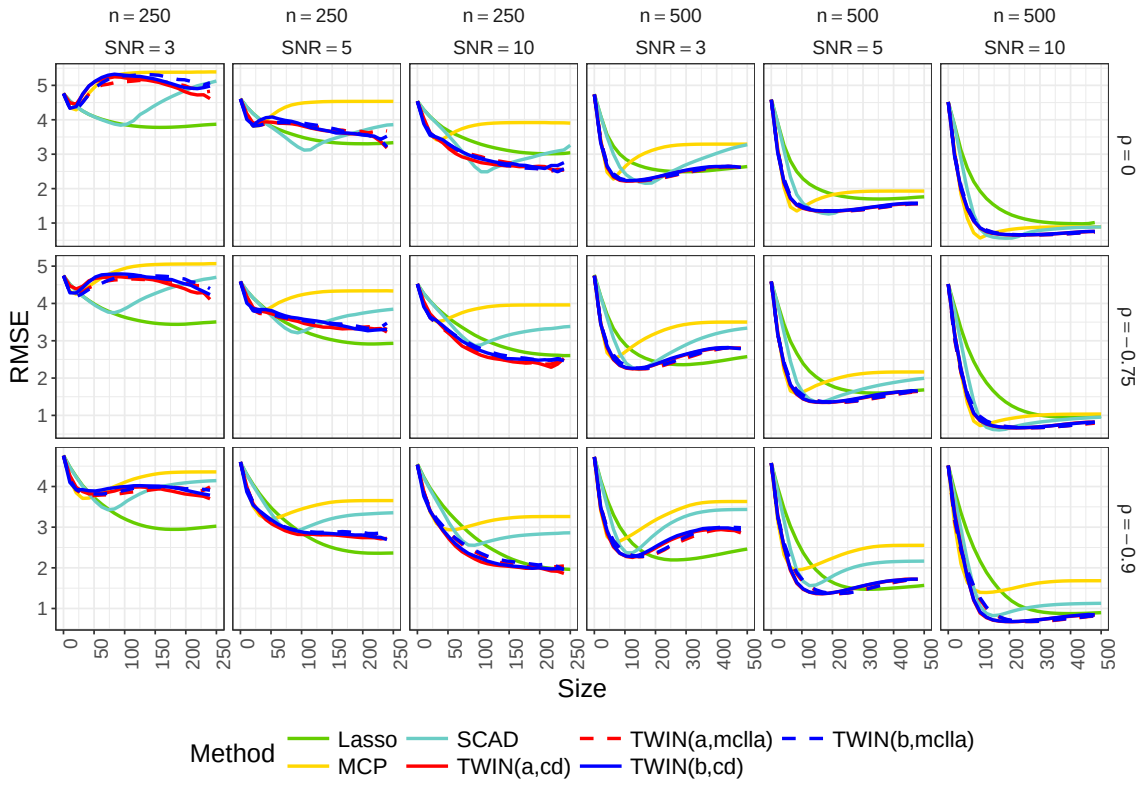

Figure C.8: The results above are for a simulation with data generated under Model 6 with $p = 2000$.

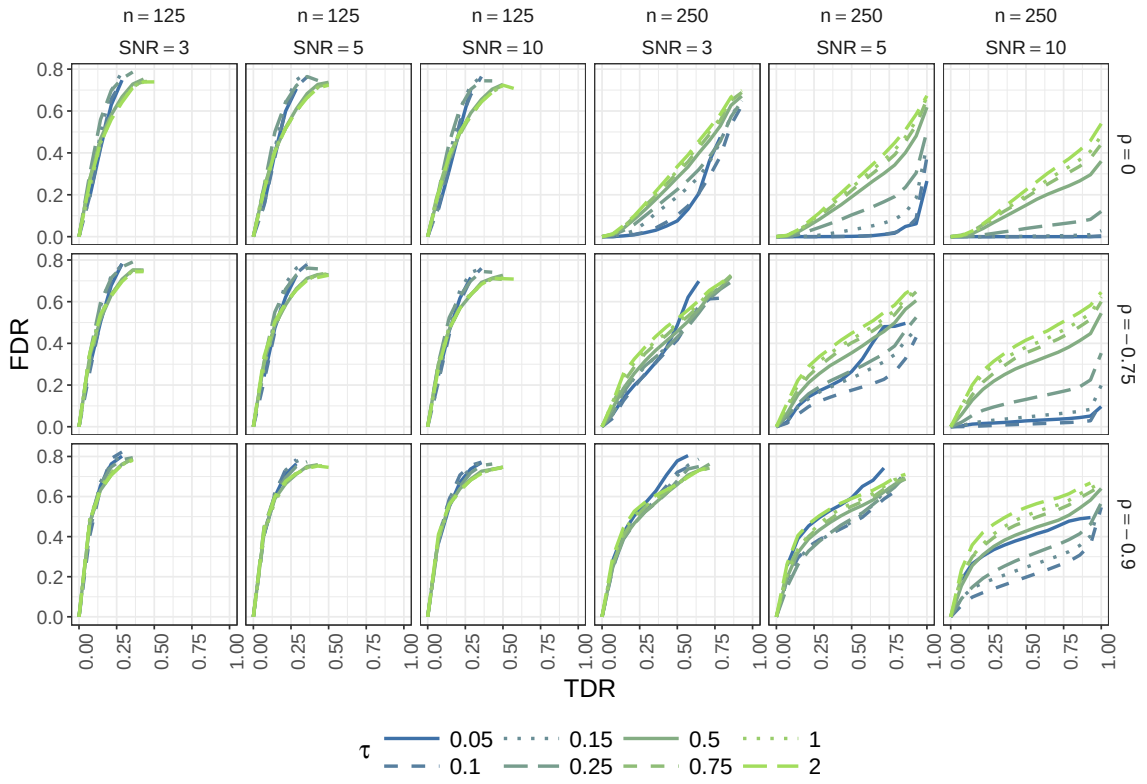

Figure C.9: The results above are for a simulation for TWIN-a with data generated under Model 1.

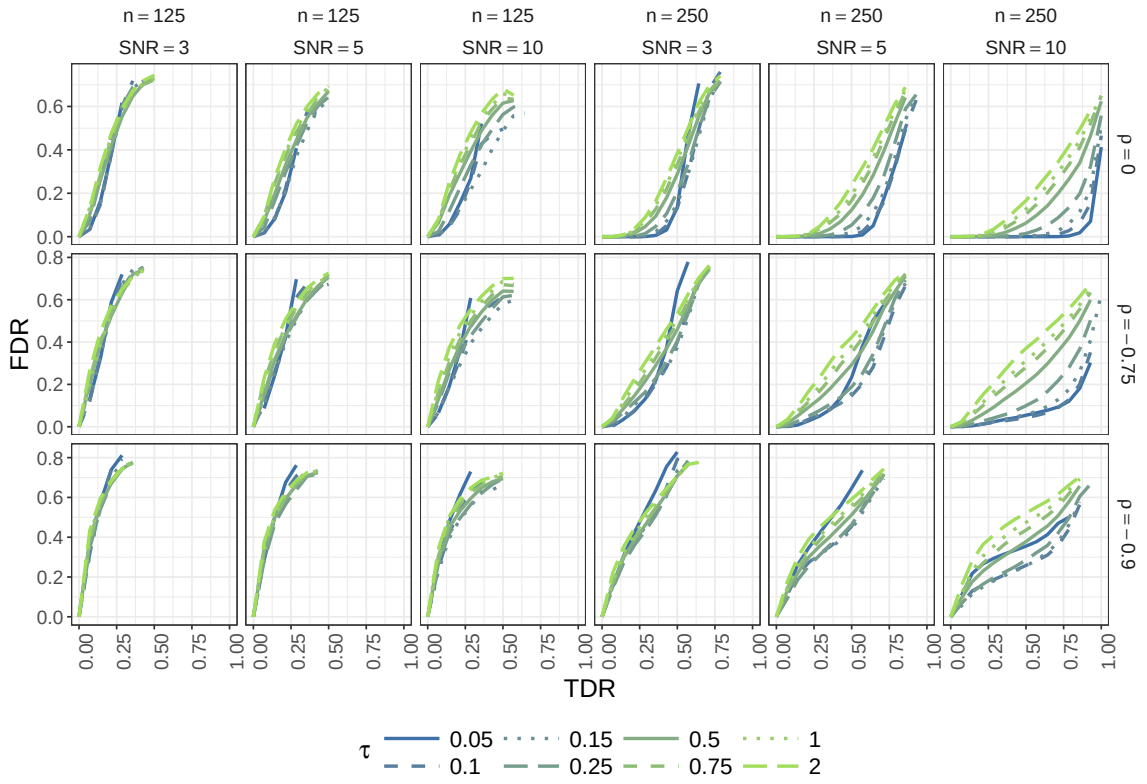

Figure C.10: The results above are for a simulation for TWIN-a with data generated under Model 2 with $p = 1000$.

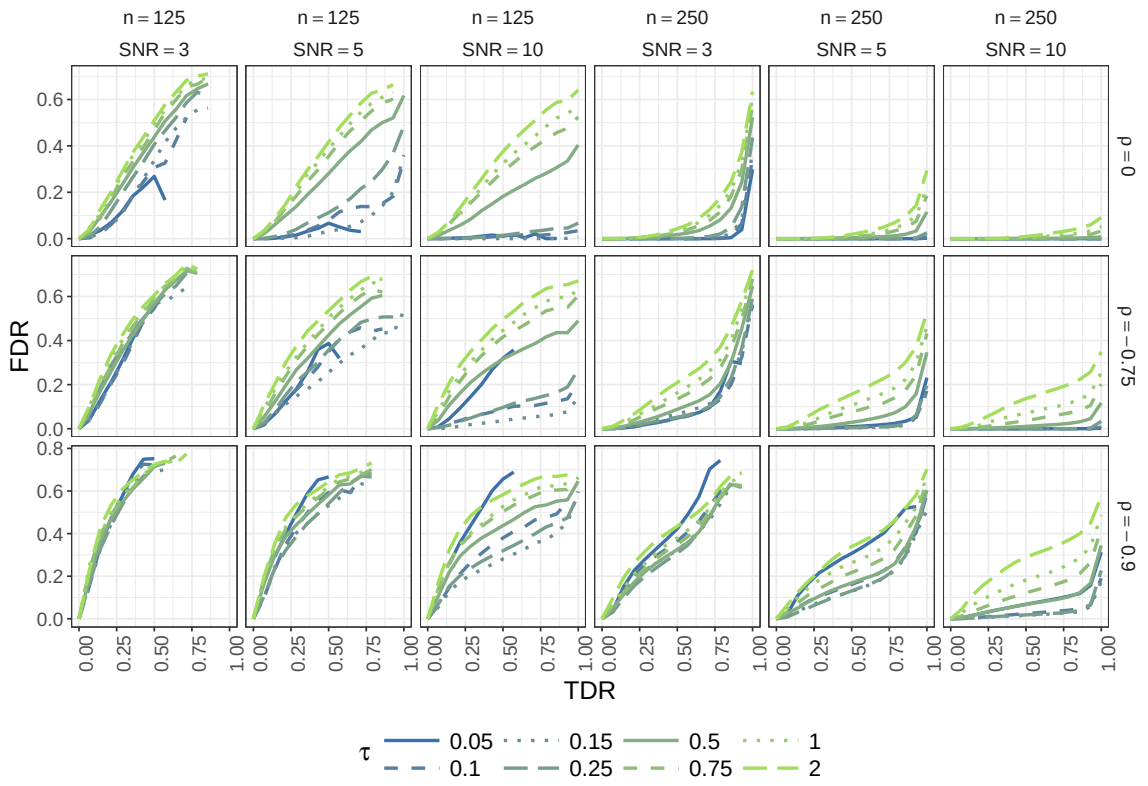

Figure C.11: The results above are for a simulation for TWIN-a with data generated under Model 3 with $p = 1000$.

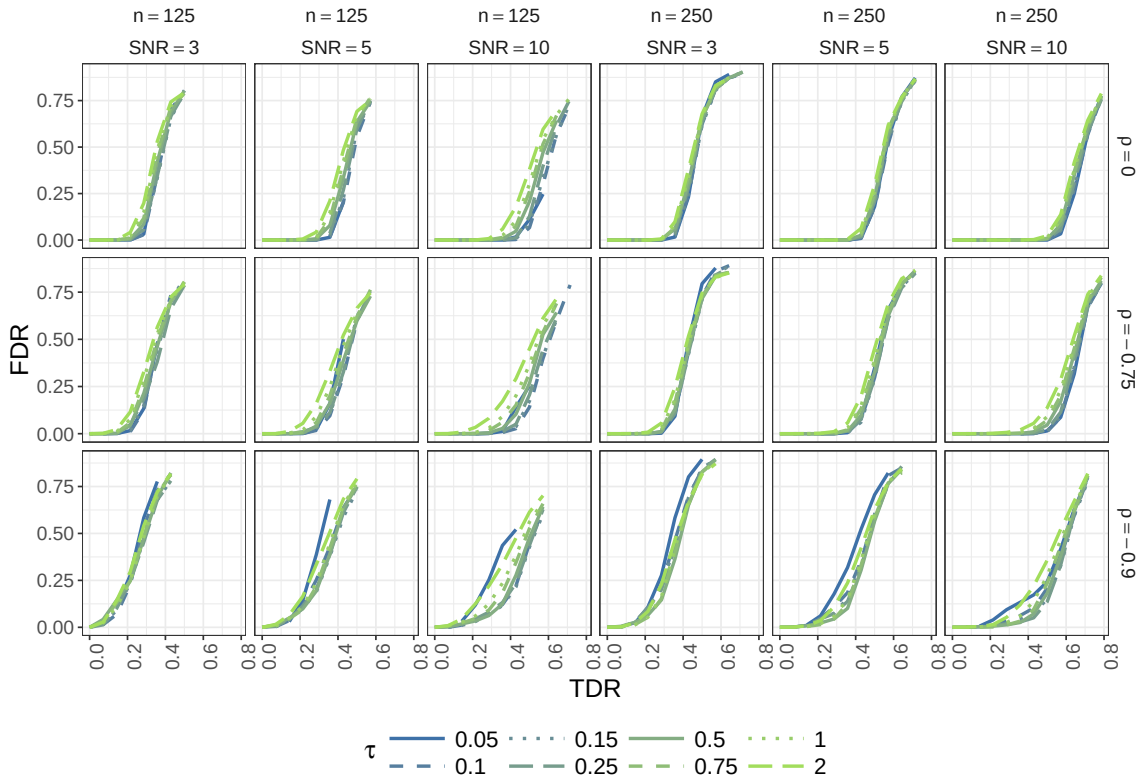

Figure C.12: The results above are for a simulation for TWIN-a with data generated under Model 4 with $p = 1000$.

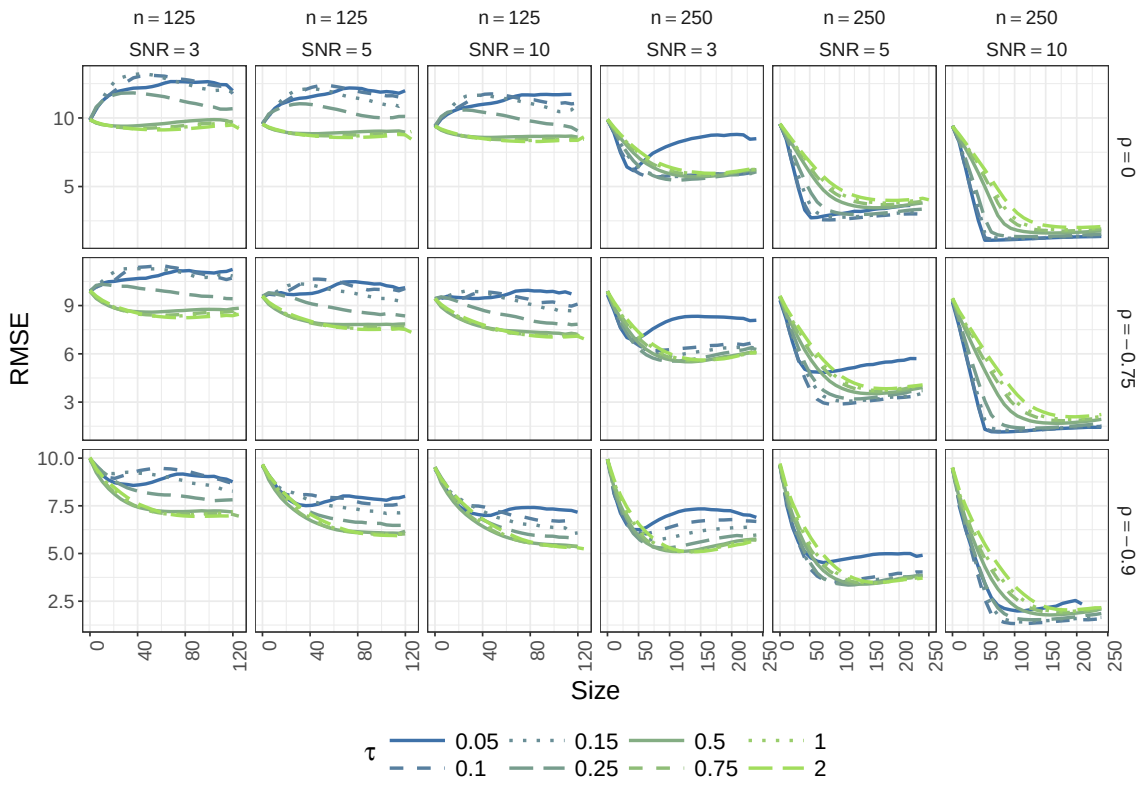

Figure C.13: The results above are for a simulation for TWIN-a with data generated under Model 1 with $p = 1000$.

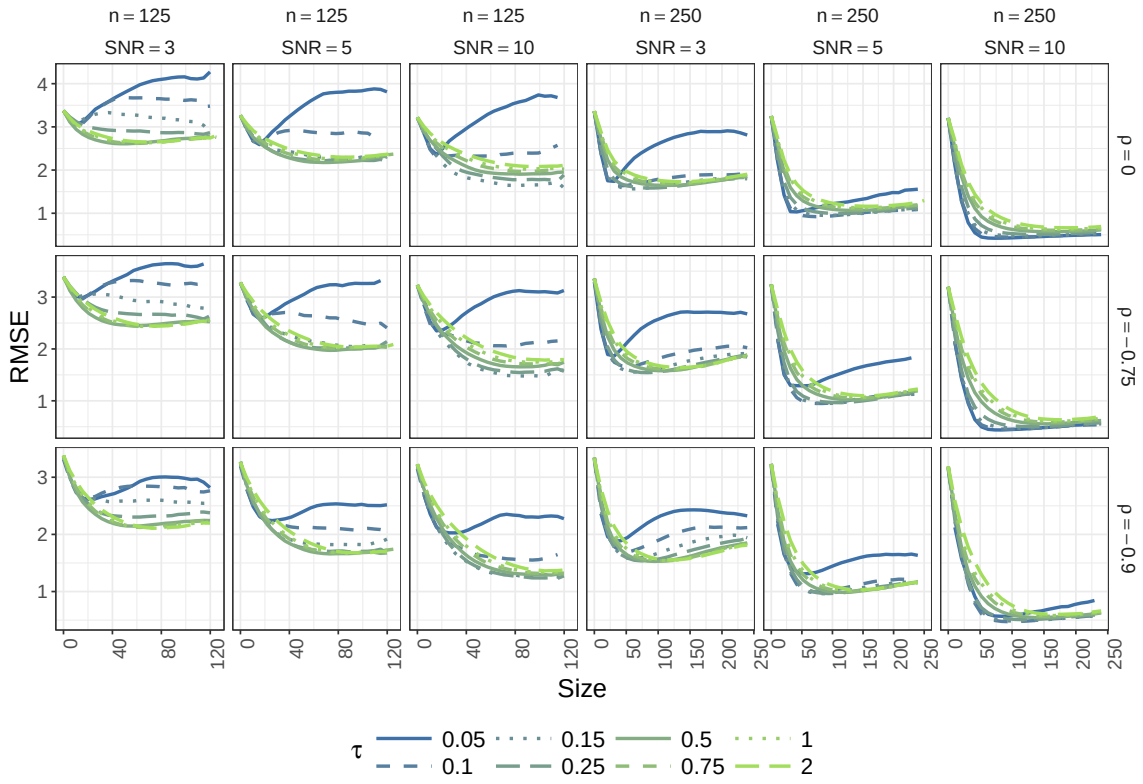

Figure C.14: The results above are for a simulation for TWIN-a with data generated under Model 2 with $p = 1000$.

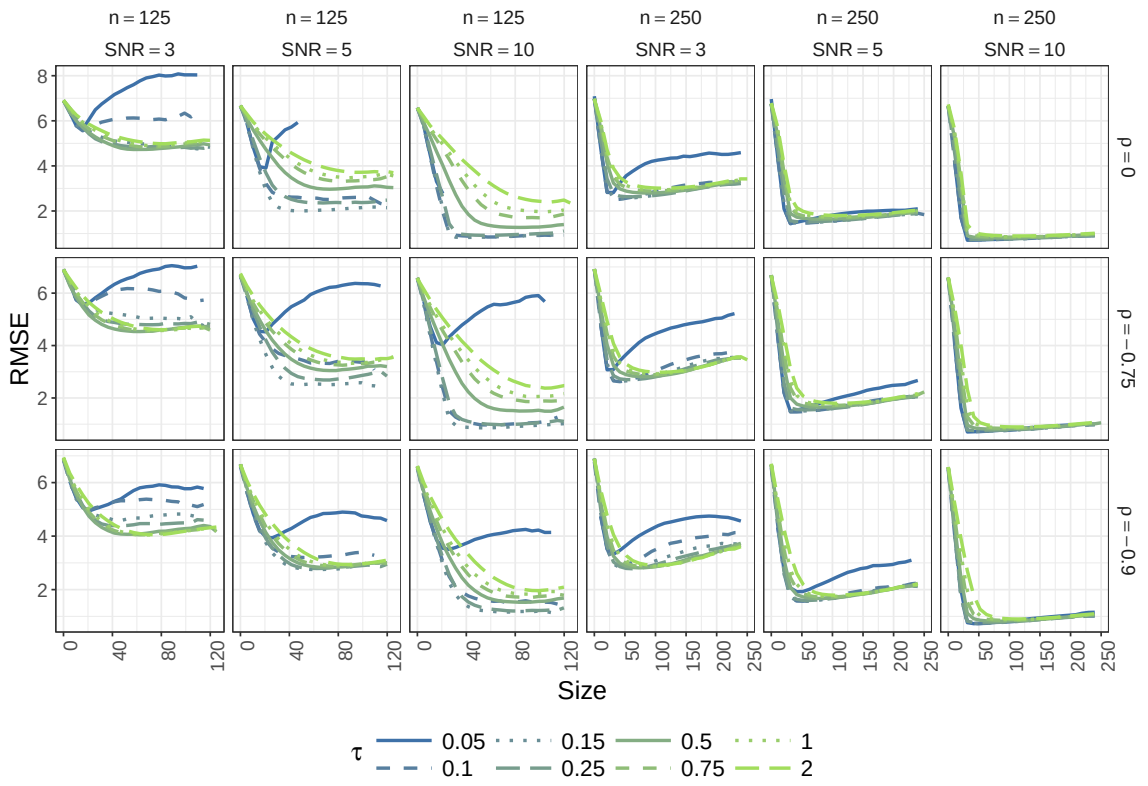

Figure C.15: The results above are for a simulation for TWIN-a with data generated under Model 3 with $p = 1000$.

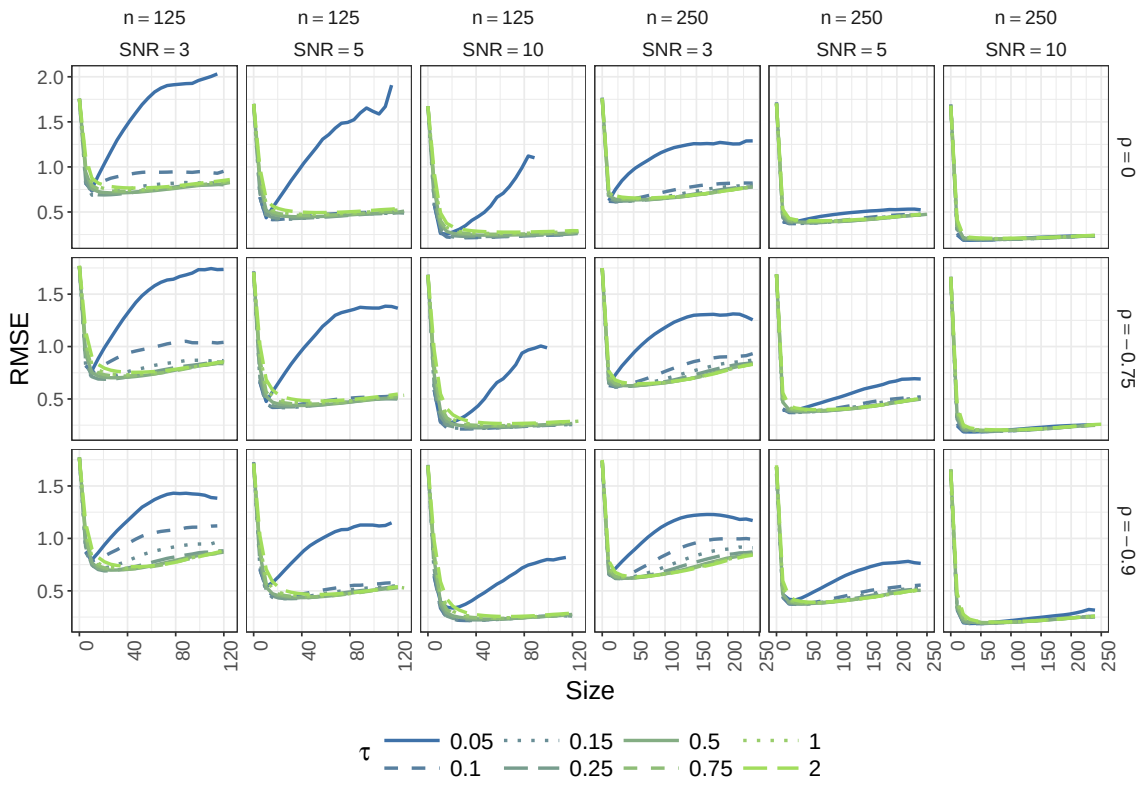

Figure C.16: The results above are for a simulation for TWIN-a with data generated under Model 4 with $p = 1000$.

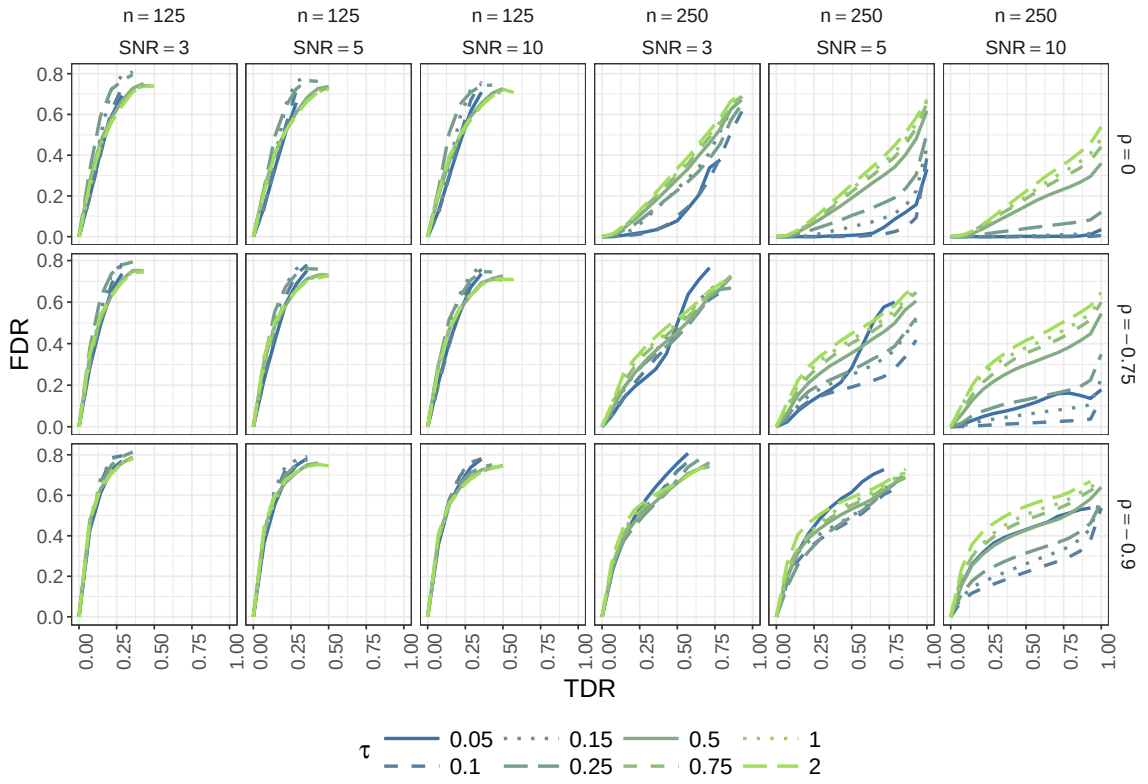

Figure C.17: The results above are for a simulation for TWIN-b with data generated under Model 1.

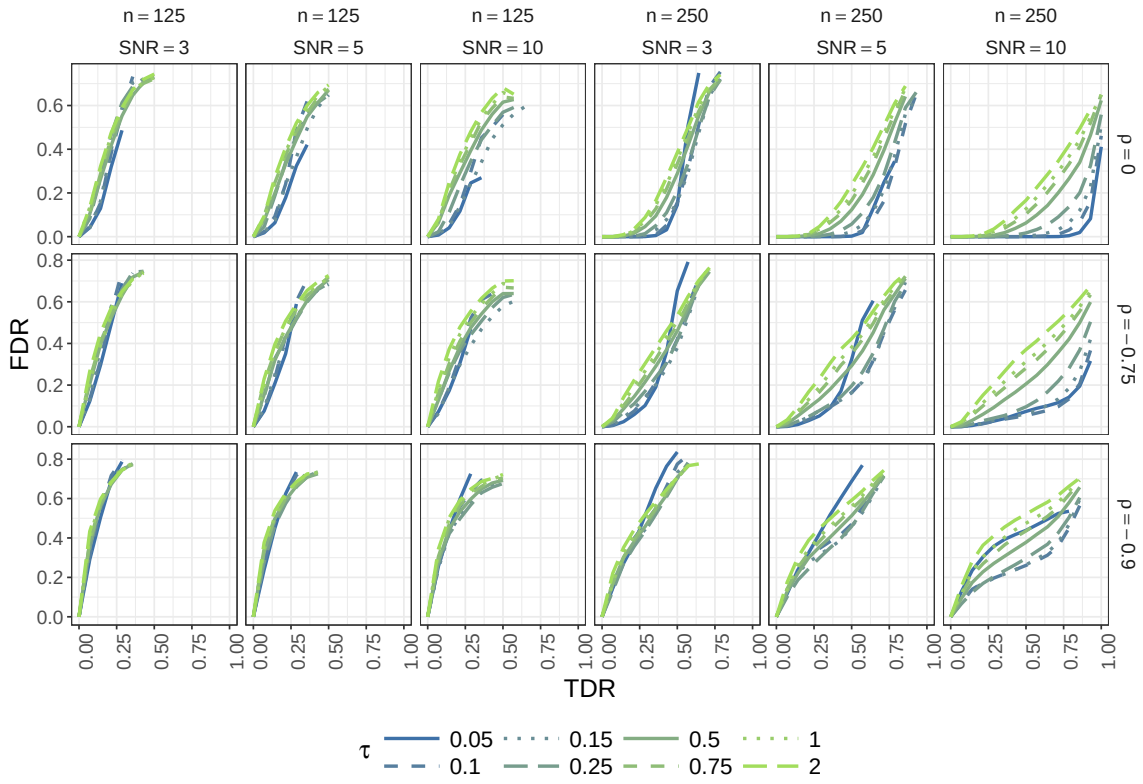

Figure C.18: The results above are for a simulation for TWIN-b with data generated under Model 2 with $p = 1000$.

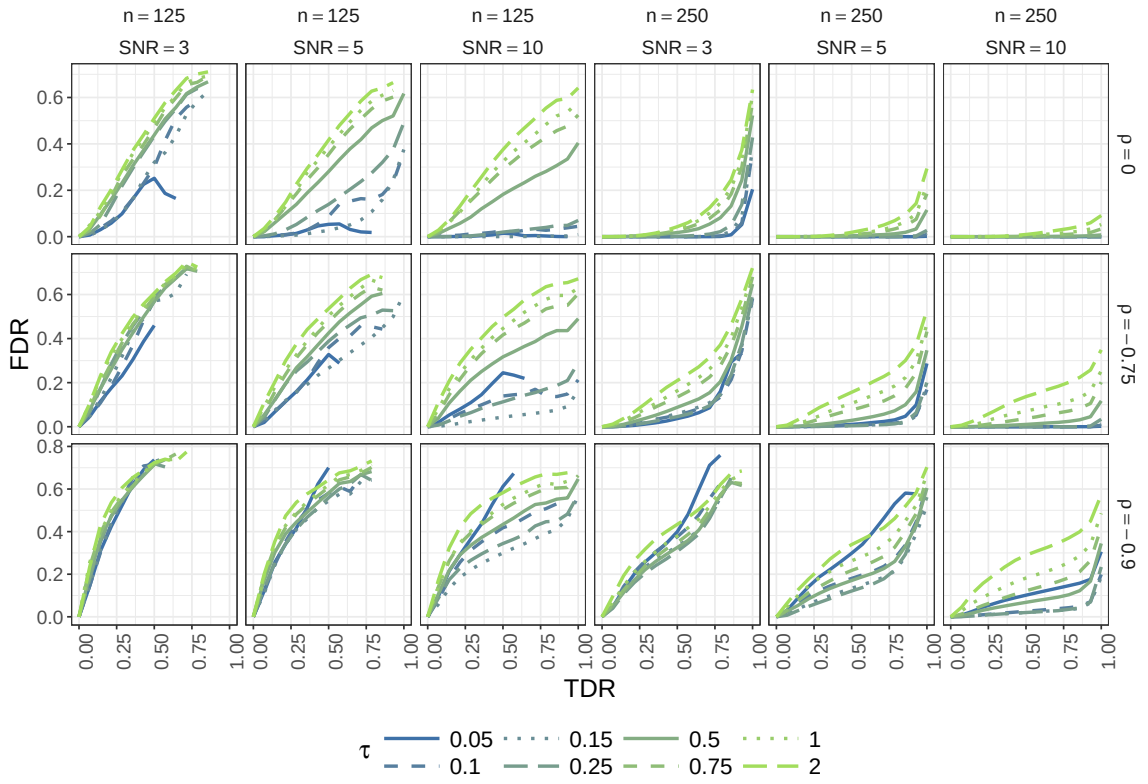

Figure C.19: The results above are for a simulation for TWIN-b with data generated under Model 3 with $p = 1000$.

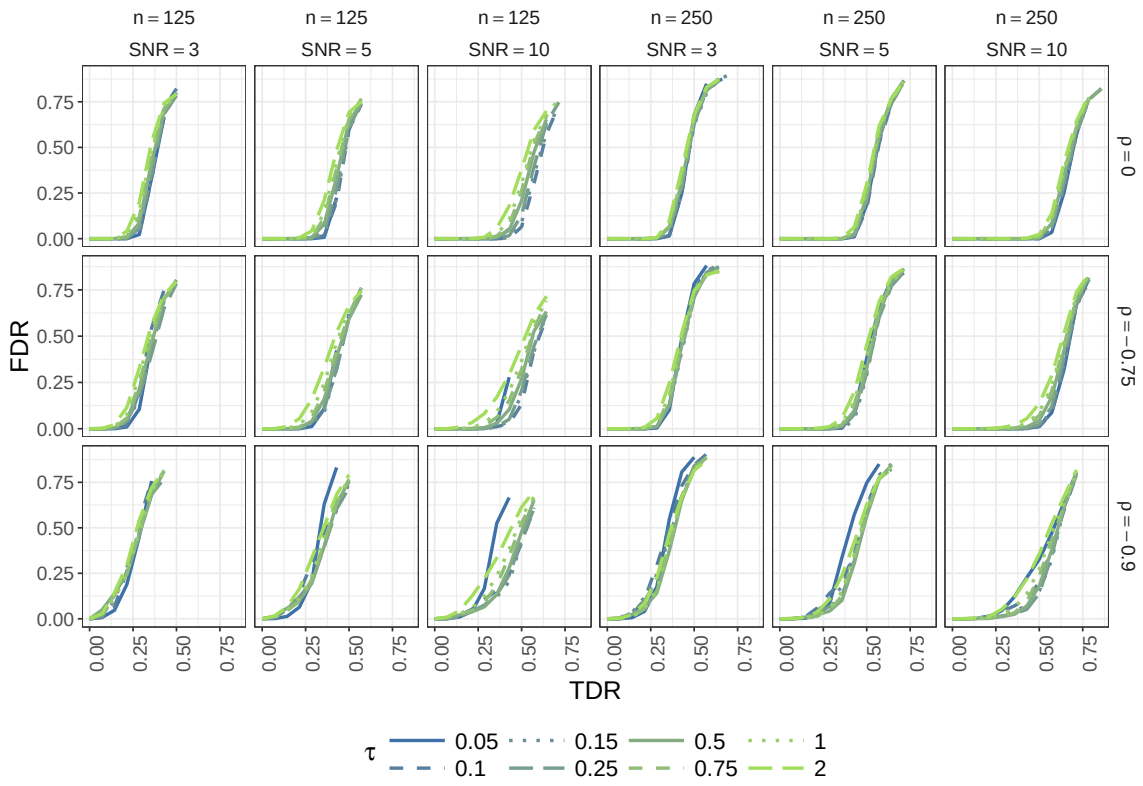

Figure C.20: The results above are for a simulation for TWIN-b with data generated under Model 4 with $p = 1000$.

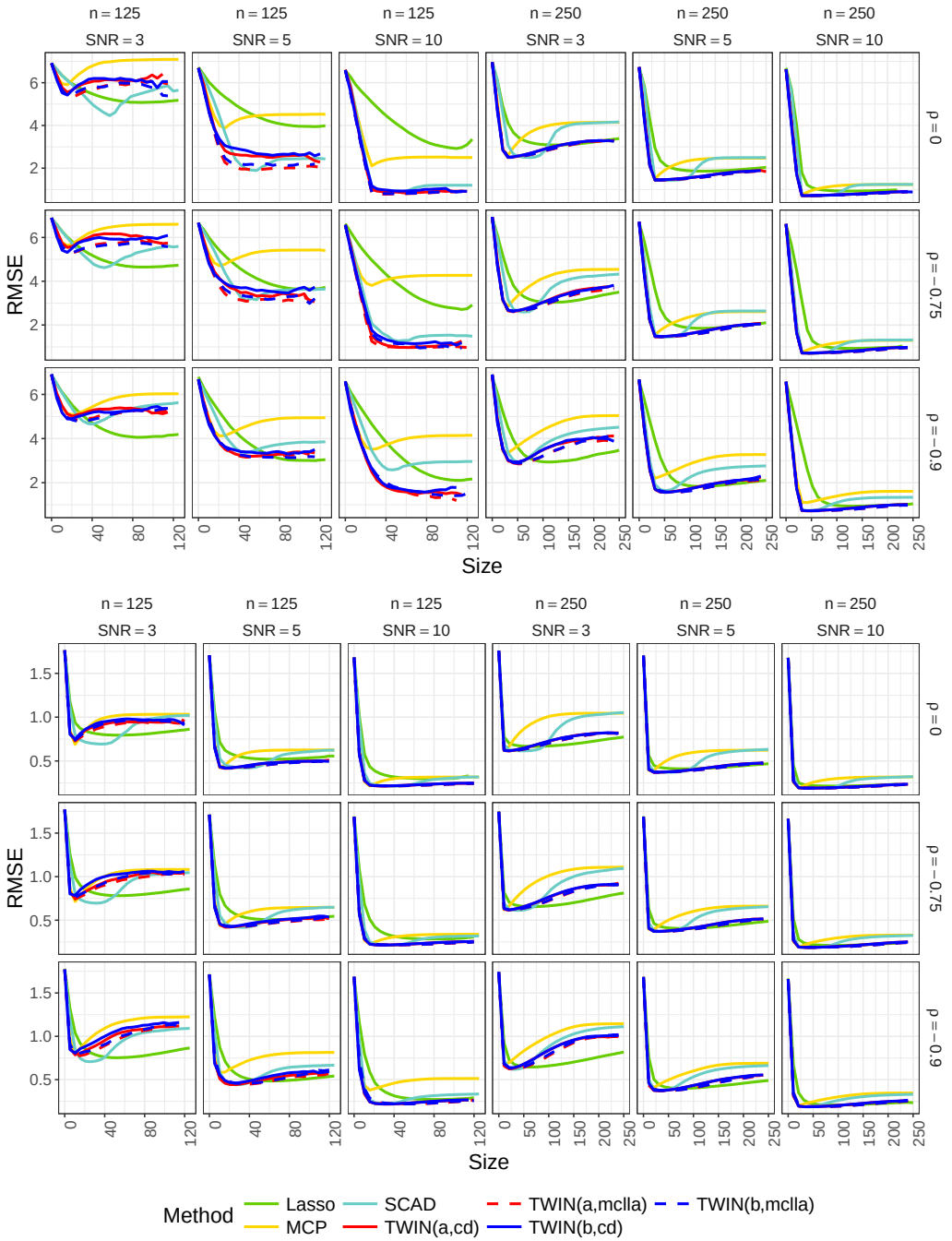

Figure C.21: The results above are for a simulation with data generated under Model 3 (top panel) and Model 4 (bottom panel) with $p = 1000$.

Appendix D

Appendix For: Towards Theoretical Understanding of Large Batch Training in Stochastic Gradient Descent

D.1 Proof of (5.2)

For the first part, without less of generality, we consider $\mathbf{w}$ is in any bounded domain of $\mathbb{R}$. Then

$$\begin{aligned}\nabla L(\mathbf{w}) = \frac{d}{d\mathbf{w}}\mathbb{E}[L_n(\mathbf{w})] &= \lim_{h\to 0}\frac{1}{h}\left\{\mathbb{E}[L_n(\mathbf{w}+h)] - \mathbb{E}[L_n(\mathbf{w})]\right\} \\ &= \lim_{h\to 0}\mathbb{E}\left\{\frac{L_n(\mathbf{w}+h)-L_n(\mathbf{w})}{h}\right\} = \lim_{h\to 0}\mathbb{E}\left\{\nabla L_n(\mathbf{w}+\tau(h))\right\}.\end{aligned}$$

where the last step is by the mean value theorem with some $0 < \tau(h) < h$. Due to continuity of ∇L_n, we can use the dominated convergence theorem and have

$$\lim_{h\to 0}\mathbb{E}\left\{\nabla L_n(\mathbf{w}+\tau(h))\right\} = \mathbb{E}\left\{\lim_{h\to 0}\nabla L_n(\mathbf{w}+\tau(h))\right\} = \mathbb{E}\left\{\nabla L_n(\mathbf{w})\right\}.$$

This completes the proof of the first part. By assuming the iid of the data, we have $\text{Var}[\widehat{\mathbf{g}}^{(B)}] = M^{-1}\boldsymbol{\sigma}^2(\mathbf{w})$. This completes the proof of the second part.

D.2 Proof of Lemma 5.1

We start to consider when $\beta(\mathbf{w}) \equiv \beta$ is a constant and follow the strategy in Kolpas et al. (2007) to derive the Fokker-Planck equation. First, consider $\mathbf{W}(t) = W(t) \in \mathbb{R}$. Note that for SGD the corresponding $W(t)$ is a Markov process, then the Chapman-Kolmogorov equation gives

$$p\left(W(t_3)|W(t_1)\right) = \int_{-\infty}^{+\infty} p\left(W(t_3)|W(t_2) = w\right) p\left(W(t_2) = w|W(t_1)\right) dw.$$

Consider the integral

$$I = \int_{-\infty}^{+\infty} h(w)\partial_t p(w,t|W)dw,$$

where $h(w)$ is a smooth function with compact support. Observe that

$$\int_{-\infty}^{+\infty} h(w)\partial_t p(w,t|W)dw = \lim_{\Delta t\to 0}\int_{-\infty}^{+\infty} h(w)\left(\frac{p(w,t+\Delta t|W) - p(w,t|W)}{\Delta t}\right)dw.$$

Letting Z be an intermediate point. Applying the Chapman-Kolmogorov identity on the right hand side yields

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left(\int_{-\infty}^{+\infty} h(w)\int_{-\infty}^{+\infty} p(w,\Delta t|Z)p(Z,t|W)dZdw - \int_{-\infty}^{+\infty} h(w)p(w,t|W)dw\right).$$

By changing the limits of integration in the first term and letting w approach Z in the second term, we obtain

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left(\int_{-\infty}^{+\infty} p(Z,t|W)\int_{-\infty}^{+\infty} p(w,\Delta t|Z)(h(w) - h(Z))dwdZ\right).$$

Expand $h(w)$ as a Taylor series about Z, we can write the above integral as

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left(\int_{-\infty}^{+\infty}p(Z,t|W)\int_{-\infty}^{+\infty}p(w,\Delta t|Z)\sum_{n=1}^{\infty}h^{(n)}(Z)\frac{(w-Z)^n}{n!}\right)dwdZ.$$

Now we define the function

$$D^{(n)}(Z)=\frac{1}{n!}\frac{1}{\Delta t}\int_{-\infty}^{+\infty}p(w,\Delta t|Z)(w-Z)^n dw.$$

We can write the integral I as

$$\int_{-\infty}^{+\infty}h(w)\partial_t p(w,t|W)dw=\int_{-\infty}^{+\infty}p(Z,t|W)\sum_{n=1}^{\infty}D^{(n)}(Z)h^{(n)}(Z)dZ.$$

Integrating by parts n times gives

$$\partial_t p(w,t)=\sum_{n=1}^{\infty}-\frac{\partial^n}{\partial Z^n}\left[D^{(n)}(Z)p(Z,t|W)\right].$$

Let $D^{(1)}(w)=-L(w)$, $D^{(2)}(w)=-\gamma(t)\beta/[2M(t)]$ and $D^{(n)}(w)=0$ for all $n\geq 3$, the above equation yields

$$\partial_t p(w,t)=\frac{\partial}{\partial w}\left[\nabla L(w)p(w,t)\right]+\frac{\partial}{\partial w^2}\left[\frac{\gamma(t)\beta}{2M(t)}p(w,t)\right],$$

which is the Fokker-Planck equation in one variable. For the multidimensional case $\mathbf{W}=(W_1,W_2,\ldots,W_p)\in\mathbb{R}^p$, the above procedure can be easily generalized to get

$$\begin{aligned}\partial_t p(\mathbf{w},t)&=\sum_{i=1}^{p}\frac{\partial}{\partial w_i}\left[\nabla L(\mathbf{w})p(\mathbf{w},t)\right]+\sum_{i=1}^{p}\frac{\partial^2}{\partial w_i^2}\left[\frac{\gamma(t)\beta}{2M(t)}p(\mathbf{w},t)\right]\\&=\nabla\cdot\left(\nabla L(\mathbf{w})p+\frac{\gamma(t)\beta}{2M(t)}\nabla p\right).\end{aligned}\tag{D.1}$$

Since $\mathbf{W}(0) = \mathbf{w}_0$, $p(\mathbf{w}, 0) = \delta(\mathbf{w}_0)$. This completes the derivation of the Fokker-Planck equation for constant $\beta(\mathbf{w}) = \beta$. For deriving (5.5), we can apply (D.1) and notice that

$$\nabla \left[\frac{\gamma(t)\beta(\mathbf{w})}{2M(t)} p \right] = \nabla \left[\frac{\gamma(t)\beta(\mathbf{w})}{2M(t)} \right] p + \frac{\gamma(t)\beta(\mathbf{w})}{2M(t)} \nabla p.$$

This completes the proof.

D.3 Discussion on the Main Assumptions (A.1) – (A.3).

We verify (A.1) and (A.2) for the L_2 loss and the mean cross entropy loss. Denote by $\{(\mathbf{x}_n, y_n), 1 \leq n \leq N\}$ the set of training data. Without loss of generality, consider $\text{Var}[y_n|\mathbf{x}_n] = 1$.

First, we consider the L_2 loss: $L(\mathbf{w}) = (\mathbf{w} - \mathbf{w}^0)^\top \mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top](\mathbf{w} - \mathbf{w}^0) + 1$. By assumption that $\boldsymbol{\sigma}^2(\mathbf{w})$ is positive definite, we have

$$\begin{aligned} \lim_{\|\mathbf{w}\| \to +\infty} L(\mathbf{w}) &\geq \lim_{\|\mathbf{w}\| \to +\infty} \lambda_{\min}\{\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top]\} \|\mathbf{w} - \mathbf{w}^0\|^2 + 1 \\ &\geq \lim_{\|\mathbf{w}\| \to +\infty} \lambda_{\min}\{\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top]\} [\|\mathbf{w}\|^2/2 - \|\mathbf{w}^0\|^2/2] + 1 = +\infty, \end{aligned} \tag{D.2}$$

where $\lambda_{\min}\{\cdot\}$ denotes the minimal eigenvalue. Note that

$$\begin{aligned} \int e^{-L(\mathbf{w})} d\mathbf{w} &= \int e^{-(\mathbf{w}-\mathbf{w}^0)^\top \mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top](\mathbf{w}-\mathbf{w}^0) - 1} \\ &\leq \int e^{-\lambda_{\min}\{\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top]\}[\|\mathbf{w}\|^2/2 - \|\mathbf{w}^0\|^2/2] - 1} < +\infty, \end{aligned}$$

This proves (A.1). To prove (A.2), we note that

$$\|\nabla L(\mathbf{w})\|^2/2 = 2(\mathbf{w} - \mathbf{w}^0)^\top \{\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top]\}^2 (\mathbf{w} - \mathbf{w}^0),$$

and

$$\nabla \cdot \nabla L(\mathbf{w}) = \text{Tr}\{\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top]\},$$

and similarly to (D.2) we can prove

$$\lim_{\|\mathbf{w}\|\to+\infty} \left\{\|\nabla L(\mathbf{w})\|^2/2 - \nabla\cdot\nabla L(\mathbf{w})\right\} = +\infty,$$

and

$$\lim_{\|\mathbf{w}\|\to+\infty} \left\{\nabla\cdot\nabla L(\mathbf{w})/\|\nabla L(\mathbf{w})\|^2\right\} = 0.$$

The assumption (A.3) can be verified straightforwardly as (A.2).

Second, we consider the mean cross entropy loss regularized with the l_2 penalty for logistic regression. Without loss of generality, we only consider the binary classification: $L(\mathbf{w}) = \mathbb{E}[-\mathbf{y}_n \log \widehat{\mathbf{y}}_n - (1-\mathbf{y}_n)\log(1-\widehat{\mathbf{y}}_n)] + \lambda\|\mathbf{w}\|^2$ with $\widehat{\mathbf{y}}_n = 1/(1+e^{-\mathbf{w}\cdot\mathbf{x}_n})$. Note that

$$\lim_{\|\mathbf{w}\|\to+\infty} L(\mathbf{w}) \geq \lambda\|\mathbf{w}\|^2 = +\infty, \quad \int e^{-L(\mathbf{w})}d\mathbf{w} \leq \int e^{-\lambda\|\mathbf{w}\|^2}d\mathbf{w} < +\infty.$$

This proves (A.1). To prove (A.2), note that $\nabla L(\mathbf{w}) = \mathbb{E}[-\mathbf{x}_n\mathbf{y}_n + \mathbf{x}_n/(1+e^{-\mathbf{w}\cdot\mathbf{x}_n})] + 2\lambda\mathbf{w}$ and $-\nabla\cdot\nabla L(\mathbf{w}) = \frac{e^{-\mathbf{w}\cdot\mathbf{x}_n}}{(1+e^{\mathbf{w}\cdot\mathbf{x}_n})^2}[2\mathbb{P}(y_n=1)-1]\mathrm{Tr}(\mathbf{x}_n\mathbf{x}_n^\top)$. Since $\lambda\|\mathbf{w}\|^2 \to \infty$, we have that $\|\nabla L(\mathbf{w})\|^2/2 - \nabla\cdot\nabla L(\mathbf{w}) \to \infty$ and $\nabla\cdot\nabla L(\mathbf{w})/\|\nabla L(\mathbf{w})\|^2 \to 0$ as $\|\mathbf{w}\| \to \infty$. Similarly, the assumption (A.3) can be verified as (A.2). This completes the proof.

D.4 Proof of Lemma 5.3

Let $\eta(t) = 2M(t)/[\gamma(t)\beta]$. By setting $\partial_t p = 0$, it can be verified that $p_\infty(\mathbf{w}) = \kappa e^{-\eta_\infty L(\mathbf{w})}$ satisfies

$$\nabla\cdot(\nabla L(\mathbf{w})p + \frac{1}{\eta(t)}\nabla p) = 0.$$

Since $\beta(\mathbf{w}) \equiv \beta$ and the assumption (A.1) ensures that $e^{-\eta_\infty L(\mathbf{w})}$ is well-defined, $p_\infty(\mathbf{w})$ is a stationary solution.

D.5 Proof of Theorem 5.4

Parallel to the notation of $p_\infty(\mathbf{w}) = \kappa e^{-\eta_\infty L(\mathbf{w})}$, we let

$$\hat{p}(\mathbf{w}, t) \equiv \kappa(t) e^{-\eta(t) L(\mathbf{w})},$$

where $\eta(t) = \frac{2M(t)}{\gamma(t)\beta(t)}$ and $\kappa(t)$ is a time-dependent normalization factor such that $\int \hat{p}(\mathbf{w}, t) d\mathbf{w} = 1$. Observe that (5.5) can be written as

$$\partial_t p = \frac{1}{\eta} \nabla_{\mathbf{w}} \cdot \left(\hat{p} \nabla_{\mathbf{w}} \left(\frac{p}{\hat{p}} \right) \right). \tag{D.3}$$

Let $\hat{p}$ be $\hat{p}(t, \mathbf{w}) = p_\infty(\mathbf{w}) \delta(t, \mathbf{w})$, where $\delta(t, \mathbf{w}) \equiv \frac{\kappa(t)}{\kappa} e^{L(\mathbf{w})(\eta_\infty - \eta(t))}$. Denote by h the scaled distance from p to p_∞:

$$h \equiv \frac{p - p_\infty}{\sqrt{p_\infty}},$$

then h satisfies the following equation,

$$\begin{aligned} \partial_t h =& \frac{1}{\eta \sqrt{p_\infty}} \nabla_{\mathbf{w}} \cdot \left[\hat{p} \, \nabla_{\mathbf{w}} \left(\frac{1}{\delta} + \frac{h}{\sqrt{p_\infty} \delta} \right) \right] \\ =& \frac{1}{\eta \sqrt{p_\infty}} \nabla_{\mathbf{w}} \cdot \left[p_\infty \left(\nabla_{\mathbf{w}} L \hat{\delta} + \nabla_{\mathbf{w}} L \hat{\delta} \left(\frac{h}{\sqrt{p_\infty}} \right) + \nabla_{\mathbf{w}} \left(\frac{h}{\sqrt{p_\infty}} \right) \right) \right], \end{aligned} \tag{D.4}$$

where $\hat{\delta}(t) = \eta(t) - \eta_\infty$. Multiplying h to the both sides of (D.4) and integrating it over x, after integration by parts, one has,

$$\begin{aligned} \frac{1}{2} \partial_t \|h\|^2 =& \frac{\hat{\delta}}{\eta} \underbrace{\int \frac{h}{\sqrt{p_\infty}} \nabla_{\mathbf{w}} \cdot (p_\infty \nabla_{\mathbf{w}} L) \, d\mathbf{w}}_{I} + \frac{\hat{\delta}}{\eta} \underbrace{\int \frac{1}{2} \left\| \frac{h}{\sqrt{p_\infty}} \right\|^2 \nabla_{\mathbf{w}} \cdot (p_\infty \nabla_{\mathbf{w}} L) \, d\mathbf{w}}_{II} \\ & - \frac{1}{\eta} \underbrace{\int p_\infty \left\| \nabla_{\mathbf{w}} \left(\frac{h}{\sqrt{p_\infty}} \right) \right\|^2 d\mathbf{w}}_{III}. \end{aligned}$$

Note that

$$\nabla_{\mathbf{w}} \cdot (p_\infty \nabla_{\mathbf{w}} L) = p_\infty \left(\nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L - \eta_\infty \|\nabla_{\mathbf{w}} L\|^2 \right),$$

so by Assumption $A3$, one has

$$|\nabla_{\mathbf{w}} \cdot (p_\infty \nabla_{\mathbf{w}} L)| \leq p_\infty^{2/3} \max\{1, \eta_\infty\} M,$$

which implies

$$I \leq \frac{\max\{1, \eta_\infty\} M}{2} \left(\|h\|^2 + \int p_\infty^{1/3} d\mathbf{w} \right).$$

For term II, first Assumption $A3$ is equivalent to

$$\lim_{\|\mathbf{w}\| \to \infty} \frac{\nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L}{2\eta_\infty \|\nabla_{\mathbf{w}} L\|^2} = 0. \tag{D.5}$$

Furthermore, Assumption $A2$ and (D.5) implies that $\lim_{\|\mathbf{w}\| \to \infty} \|\nabla_{\mathbf{w}} L\|^2 \to +\infty$, so there exists a constant R, such that

$$\nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L - 2\eta_\infty \|\nabla_{\mathbf{w}} L\|^2 \leq \eta_\infty, \quad \eta_\infty \|\nabla_{\mathbf{w}} L\|^2 \geq \eta_\infty, \quad \text{for } \forall \|\mathbf{w}\| > R.$$

Therefore one has,

$$\nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L - \eta_\infty \|\nabla_{\mathbf{w}} L\|^2 \leq 0, \quad \text{for } \forall \|\mathbf{w}\| > R.$$

By the continuity of the loss function, for $\|\mathbf{w}\| \leq R$, there exists a constant C_2, such that

$$\left| \nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L - \eta_\infty \|\nabla_{\mathbf{w}} L\|^2 \right| \leq C_2, \quad \text{for } \forall \|\mathbf{w}\| < R.$$

Combining the above two inequality gives the bound for term II,

$$|II| \leq \frac{C_2}{2} \|h\|^2.$$

Thus combine the estimates for the term I and II, one has

$$I + II \leq C_1 \|h\|^2 + C_1, \tag{D.6}$$

where $C_1 = \frac{1}{2}\max\{1, \eta_\infty\}\max\left\{\int p_\infty^{1/3} d\mathbf{w}, 1 + \frac{C_2}{2}\right\} M$.

For term III, under Assumption $A2$, one has the following Poincaré inequality (see, e.g., Pavliotis (2014)) on $p_\infty d\mathbf{w}$,

$$\int \left\| \nabla_{\mathbf{w}} \left(\frac{h}{\sqrt{p_\infty}} \right) \right\|^2 p_\infty \, d\mathbf{w} \geq C_P \int \left(\frac{h}{\sqrt{p_\infty}} - \int h\sqrt{p_\infty} d\mathbf{w} \right)^2 p_\infty \, d\mathbf{w}.$$

In addition, the fact that $\int h\sqrt{p_\infty}\, d\mathbf{w} = 0$ gives

$$III \geq C_P \left\| h \right\|^2. \tag{D.7}$$

The reason why $\int h\sqrt{p_\infty}\, d\mathbf{w} = 0$ comes from the conservation of mass. That is, if one integrates (D.3) over $\mathbf{w}$ and uses integration by parts,

$$\partial_t \left(\int p(\mathbf{w}, t) \, d\mathbf{w} \right) = 0,$$

which implies $\int h\sqrt{p_\infty}\, d\mathbf{w} = \int p \, d\mathbf{w} - \int p_\infty \, d\mathbf{w} = 0$. So combining (D.6) and (D.7) gives,

$$\frac{1}{2}\partial_t \left\| h \right\|^2 + \frac{C_P}{\eta} \left\| h \right\|^2 \leq \frac{C_1 \hat{\delta}}{\eta} \left(\left\| h \right\|^2 + 1 \right) \tag{D.8}$$

Since $\eta(t) \to \eta_\infty > 0$ as $t \to \infty$, so there exists T large enough, such that for $\forall t > T$,

$$\hat{\delta} = |\eta(t) - \eta_\infty| \leq \min\left\{ \frac{\eta_\infty}{3}, \frac{C_P}{3C_1} \right\}. \tag{D.9}$$

Plugging $\hat{\delta} \leq \frac{c}{2C_1}$ into (D.8), one has

$$\frac{1}{2}\partial_t \left\| h \right\|^2 + \frac{2C_P}{3\eta} \left\| h \right\|^2 \leq \frac{C_P}{3\eta}, \quad \text{for } \forall t > T. \tag{D.10}$$

Futhermore, (D.9) also implies $2\eta_\infty/3 \leq \eta(t) \leq 4\eta_\infty/3$, which indicates that

$$\frac{2C_P}{3\eta} \geq \frac{C_P}{2\eta_\infty}, \quad \frac{C_P}{3\eta} \leq \frac{C_P}{2\eta_\infty}.$$

Therefore, (D.10) becomes

$$\frac{1}{2}\partial_t \|h\|^2 + \frac{C_P}{2\eta_\infty}\|h\|^2 \leq \frac{C_P}{2\eta_\infty}, \quad \text{for } \forall t > T.$$

Integrate the above equation from T to $t > T$, one has,

$$\|h(t)\|^2 \leq \left(\|h(T)\|^2 + \frac{C_P}{\eta_\infty}(t-T)\right) - \frac{C_P}{\eta_\infty}\int_T^t \|h(s)\|^2 \, ds.$$

By Gronwall's Inequality, one ends up with,

$$\|h(t)\|^2 \leq \left(\frac{C_P}{\eta_\infty}(t-T) + \|h(T)\|^2\right) e^{-\frac{C_P}{\eta_\infty}(t-T)}.$$

Remark D.1. *There are some work in the literature about the convergence of the Fokker-Planck equation solution. However, most of these results focus on the convex $L(\mathbf{w})$. See, e.g., Arnold et al. (2001); Pavliotis (2014). These results are different from the case under our consideration.*

D.6 Mathematical Quantification of the Constant T in Theorem 5.4

We note that T should be large enough such that for all $t > T$,

$$|\eta(t) - \eta_\infty| \leq \min\left\{\frac{\eta_\infty}{3}, \frac{C_P}{3C_1}\right\}, \quad \text{where } C_1 = \frac{M}{2}\max\{1, \eta_\infty\}\max\left\{\int p_\infty^{1/3} d\mathbf{w}, 1 + \frac{C_2}{2}\right\}.$$

Here $C_2 > 0$ is the bound for $\left|\nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L - \eta_\infty \|\nabla_{\mathbf{w}} L\|^2\right|$ in bounded domain $\{\|\mathbf{w}\| < R\}$, such that

$$\nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L - \eta_\infty \|\nabla_{\mathbf{w}} L\|^2 \leq \begin{cases} 0, & \text{for } \forall \|\mathbf{w}\| > R, \\ C_2, & \text{for } \forall \|\mathbf{w}\| < R. \end{cases}$$

This quantification of T is based on the proof in Section D.5.

D.7 Proof of Theorem 5.5

Let

$$P_\epsilon(\check{\mathbf{w}}) = \mathbb{P}(\|\mathbf{W}_\gamma(\infty) - \check{\mathbf{w}}\| \leq \epsilon)$$

be the probability of $W(\infty)$ staying in the ϵ-neighborhood of global minimum $\check{\mathbf{w}}$, and the probability density function of $W(\infty)$ is p_∞, then

$$\begin{aligned} P_\epsilon(\check{\mathbf{w}}) &= \int_{\|\mathbf{w}-\check{\mathbf{w}}\|^2 \leq \epsilon^2} \kappa e^{-\eta_\infty L(\mathbf{w})} d\mathbf{w} \\ &= \int_{\|\mathbf{w}-\check{\mathbf{w}}\|^2 \leq \epsilon^2} \kappa e^{-\eta_\infty [L(\check{\mathbf{w}}) + (\mathbf{w}-\check{\mathbf{w}})' \Delta L(\check{\mathbf{w}})(\mathbf{w}-\check{\mathbf{w}}) + o\{(\mathbf{w}-\check{\mathbf{w}})^2\}]} d\mathbf{w} \end{aligned}$$

Since $\check{\mathbf{w}}$ is a local minimum of $L(\mathbf{w})$, so $\Delta L(\check{\mathbf{w}})$ is positive definite, then there exists an orthogonal matrix O and diagonal matrix Λ, such that $\Delta L = O'\Lambda O$. For simplicity, we assume $\Delta L = \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_d)$.

$$\begin{aligned} \lim_{\epsilon \to 0} P_\epsilon(\check{\mathbf{w}}) &= \lim_{\epsilon \to 0} \left[\kappa e^{-\eta_\infty L(\check{\mathbf{w}})} \int_{\|\mathbf{w}\|^2 \leq \epsilon^2} \prod_{j=1}^{d} e^{-\eta_\infty \lambda_j w_j} d\mathbf{w} \right] e^{\eta_\infty \epsilon^2} \\ &= \lim_{\epsilon \to 0} \left[\kappa e^{-\eta_\infty L(\check{\mathbf{w}})} \prod_{j=1}^{d} \frac{1}{\sqrt{\eta_\infty \lambda_j}} \int_{-\epsilon\sqrt{\eta_\infty \lambda_j}}^{\epsilon\sqrt{\eta_\infty \lambda_j}} e^{-w^2} dw \right] e^{\eta_\infty \epsilon^2} \\ &= \lim_{\epsilon \to 0} \left[\frac{\kappa e^{-\eta_\infty L(\check{\mathbf{w}})}}{\eta_\infty^{d/2}} \prod_{j=1}^{d} \frac{1}{\sqrt{\lambda_j}} \left(\Phi\left(\epsilon\sqrt{\eta_\infty \lambda_j}\right) - \Phi\left(-\epsilon\sqrt{\eta_\infty \lambda_j}\right) \right) \right] e^{\eta_\infty \epsilon^2}, \end{aligned}$$

where the first equality comes from change of variable $\mathbf{w} - \check{\mathbf{w}} \to \mathbf{w}$, and the second one comes from $\eta_\infty \lambda_j \mathbf{w}_j \to \mathbf{w}_j$. Here $\phi(\cdot)$ in the last equality is the cumulative density function for standard normal distribution. Using the approximation of the cumulative density function in Polya (1945), one can simplify the above equation by

$$\begin{aligned} \lim_{\epsilon \to 0} P_\epsilon(\check{\mathbf{w}}) &= \lim_{\epsilon \to 0} \left[\frac{\kappa e^{-2\eta_\infty L(\check{\mathbf{w}})}}{\eta_\infty^{d/2}} \prod_{j=1}^{d} \sqrt{\frac{1 - e^{-\epsilon^2 \eta_\infty \lambda_j / \pi}}{\lambda_j}} \right] e^{\eta_\infty \epsilon^2} \\ &= \lim_{\epsilon \to 0} \frac{\kappa e^{-2\eta_\infty L(\check{\mathbf{w}})}}{\eta_\infty^{d/2} \det(\Delta L(\check{\mathbf{w}}))} \left[e^{\eta_\infty \epsilon^2} \prod_{j=1}^{d} \sqrt{1 - e^{-\epsilon^2 \eta_\infty \lambda_j / \pi}} \right]. \end{aligned}$$

This completes the proof.

D.8 Numerical Illustrations of Theorem 5.5

To illustrate Theorem 5.5, we explore three different examples showing how the probability changes with respect to M/γ, $\Delta L(\breve{\mathbf{w}})$, and the variance $\sigma^2(\breve{\mathbf{w}}) = \beta(\breve{\mathbf{w}})$:

- **Example 1**: Consider the risk function $L(\mathbf{w})$ has three different global minima $\mathbf{w}_i$, $i = 1, 2, 3$, with different Hessians 4.5, 12.5, and 28.125, respectively. We are interested in the probability of the mini-batch SGD $\lim_{k\to\infty} \mathbf{w}_k$ staying in the ϵ-neighborhood of global minima with respect to the ratio M/γ, where M is the batch size and γ is the learning rate. The results are shown in Figure D.1.

- **Example 2**: Consider the variance of SGD has four different levels: $5, 10, 50, 100$. We are interested in the probability of the mini-batch SGD $\lim_{k\to\infty} \mathbf{w}_k$ staying in the ϵ-neighborhood of a same global minimum of $L(\mathbf{w})$ with respect to the ratio M/γ. The results are shown in Figure D.2.

- **Example 3**: Consider two-dimensional cases. We are interested in the risk function $L(\mathbf{w})$ has two different global minima and furthermore, $L(\mathbf{w})$ has two different global minima. For two minima case, we consider $L(\mathbf{w})$ has two different Hessians $(2.42, 0.022)$ and $(2.22, 0.222)$. For three minima case, we consider $L(\mathbf{w})$ has three different Hessians $(15, 20)$, $(14.22, 42.66)$, and $(102.13, 25.53)$.

Results. The results of Example 1 are given in Figure D.1. We draw the following conclusions.

- First, if the batch size M and learning rate γ are the same, then $\mathbf{W}(\infty)$ is more likely to stay near the flat minimum whose Hessian is smaller.

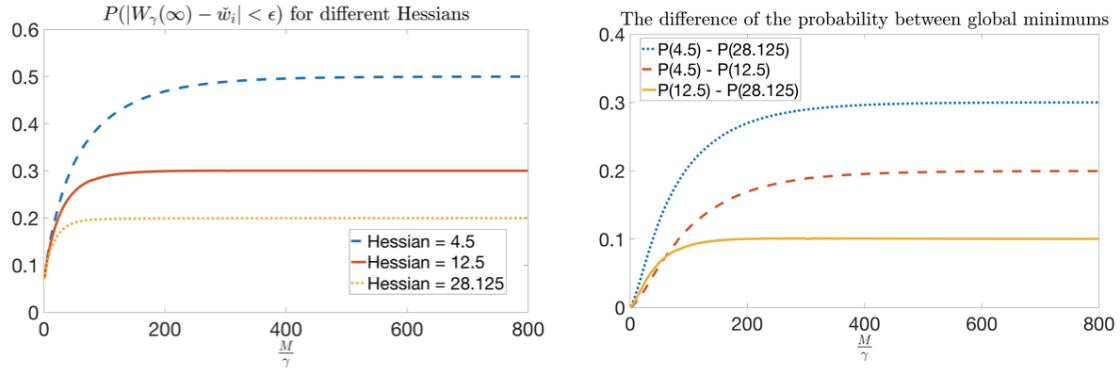

Figure D.1: Illustration of Example 1 with $\epsilon = 0.1$. The left panel shows the probability of $W(\infty)$ staying in the ϵ-neighborhood of different global minima. The right panel compares the differences of probabilities that $W(\infty)$ staying in the ϵ-neighborhood of different global minima.

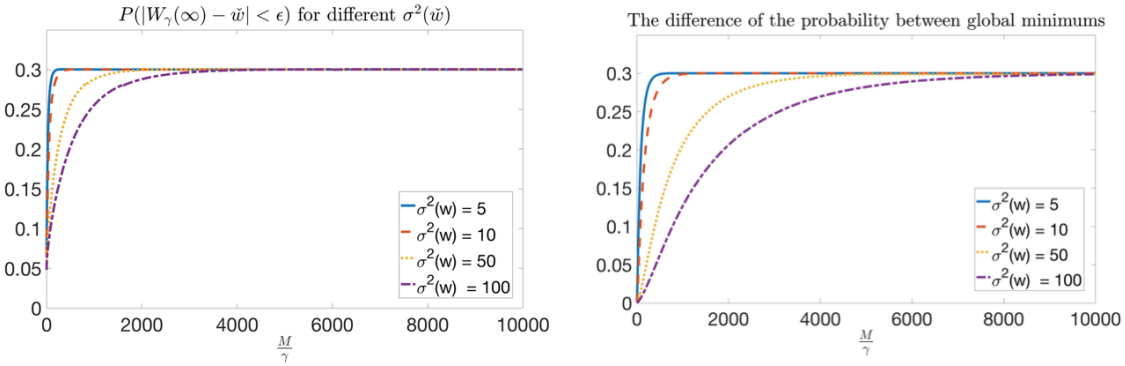

Figure D.2: Illustration of Example 1 with $\epsilon = 0.1$. The left panel shows the probability of $W(\infty)$ staying in the ϵ-neighborhood under different SGD variances $\sigma^2(\check{\mathbf{w}})$. The right panel compares the differences of probabilities that $W(\infty)$ staying in the ϵ-neighborhood of different $\sigma(\check{\mathbf{w}})$.

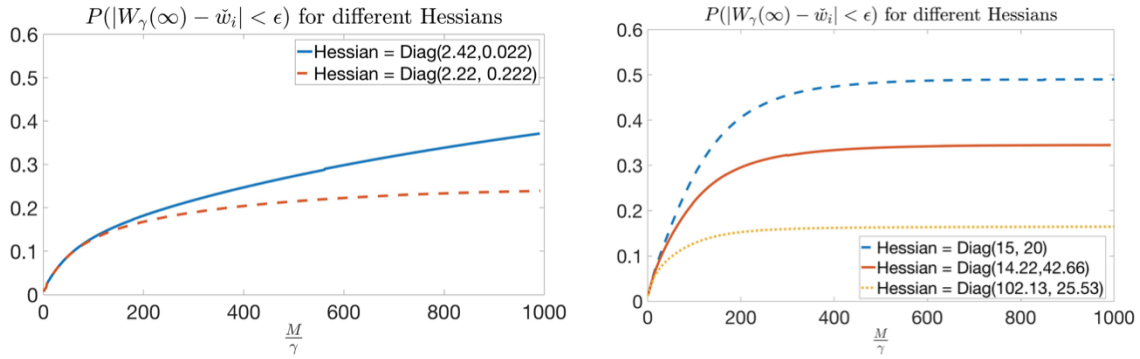

Figure D.3: Illustration of Example 3 with $\epsilon = 0.1$. The left panel shows the probability of the limiting mini-batch SGD $\lim_{k\to\infty} \mathbf{w}_k$ staying in the ϵ-neighborhood of two different global minima. The right panel shows the probability of three different global minima.

- Second, as the ratio M/γ increases, the probability of $\mathbf{W}(\infty)$ converging to a flatter minimum will increase faster than that of a sharper minimum.

- Third, if the ratio of Hessians (the Hessian at a sharp minima divides the Hessian at a flat minima) increases, the difference of probabilities would increase as illustrated in the right panel of Figure D.1. Moreover, if we increase the ratio M/γ, the difference of probabilities becomes more distinct.

The results of Example 2 are given in Figure D.2. We draw the following conclusion.

- If the variance $\sigma(\check{\mathbf{w}})$ increases, the effect of the ratio M/γ for the probability that converging to the global minimum will decrease. That implies as $\sigma(\check{\mathbf{w}})$ increases, the probability of SGD converging to a flat minimum will increase slower.

The results of Example 3 are given in Figure D.3. We draw the following conclusion.

- For a same ratio M/γ, if the product of the eigenvalues of the Hessian increases, then $\mathbf{W}(\infty)$ will be more likely to stay near the minimum. For a same sharpness of the minimum, if one increase the batch size or decrease the learning rate, $\mathbf{W}(\infty)$ will be more likely to stay near the minimum.

- We conclude that the product of eigenvalues of the Hessian matrix will affect the probability of $W(\infty)$ staying in the ϵ-neighborhood of the minimum, which is different from the sum of eigenvalues, the smallest eigenvalue, or the largest eigenvalue for multi-dimensional cases.

Appendix E

Appendix For: Another Look at Statistical Calibration: A Non-Asymptotic Theory and Prediction-Oriented Optimality

This section consists of three parts. In Section E.1, we give proofs for main results of Section 6.3. In Section E.2, we prove results of Section 6.4. In Section E.3, we present a key lemma.

E.1 Proofs for Section 6.3

E.1.1 Upper Bound Result: Theorem 6.1

We define a new norm $\|\cdot\|$ in $\mathcal{H}$ by

$$\|g\|^2 = \|g\|^2_{L_2(\Pi)} + \|g\|^2_{\mathcal{H}}, \quad \forall g \in \mathcal{H}.$$

Note that $\|\cdot\|$ is a norm because that $\|\cdot\|^2$ defined above is a quadratic form and it equals to zero if and only if $g = 0$. Since the density function of Π is bounded away from zero

and infinity, there exists some constant $c > 0$ such that $\|g\|^2_{L_2(\Pi)} \leq c\|g\|^2_{\mathcal{H}}$. Thus, $\|g\|^2 \leq (c+1)\|g\|^2_{\mathcal{H}}$. This together with the fact $\|g\|^2_{\mathcal{H}} \leq \|g\|^2$ imply that $\|\cdot\|$ and $\|\cdot\|_{\mathcal{H}}$ are equivalent. In particular, $\|g\| < \infty$ if and only if $\|g\|_{\mathcal{H}} < \infty$. Let $\langle \cdot, \cdot \rangle$ be the inner product associated with $\|\cdot\|$, which can be constructed as follows:

$$\langle g_1, g_2 \rangle = \frac{1}{4}(\|g_1 + g_2\|^2 - \|g_1 - g_2\|^2), \quad \forall g_1, g_2 \in \mathcal{H}.$$

Then $\langle g_1, g_2 \rangle = \langle g_1, g_2 \rangle_{L_2(\Pi)} + \langle g_1, g_2 \rangle_{\mathcal{H}}$. Denote by $R(\cdot, \cdot)$ the reproducing kernel associated with $(\mathcal{H}, \|\cdot\|)$. By Mercer's theorem, we have the following eigenvalue decomposition:

$$R(x, x') = \sum_{\nu \geq 1} (1 + \lambda_\nu^{-1})^{-1} \phi_\nu(x) \phi_\nu(x').$$

Let $g_\nu = \langle g, \phi_\nu \rangle_{L_2(\Pi)}$ for any $g \in \mathcal{H}$. Then

$$\|g\|^2 = \sum_{\nu=1}^{\infty} (1 + \lambda_\nu^{-1}) g_\nu^2, \quad \|g\|^2_{L_2(\Pi)} = \sum_{\nu=1}^{\infty} g_\nu^2, \quad \|g\|^2_{\mathcal{H}} = \sum_{\nu=1}^{\infty} \lambda_\nu^{-1} g_\nu^2.$$

Now, we define a norm $\|\cdot\|_a$ for any $0 \leq a \leq 1$ by

$$\|g\|_a^2 = \sum_{\nu=1}^{\infty} (1 + \lambda_\nu^{-a}) g_\nu^2.$$

It is clear that $\|g\|_0 = \sqrt{2}\|g\|_{L_2(\Pi)}$ and $\|g\|_1 = \|g\|$. Let $\langle \cdot, \cdot \rangle_a$ be the inner product associated with $\|\cdot\|_a$ for any $0 \leq a \leq 1$ (Cox, 1984).

Write

$$l_n(g) = \frac{1}{2n} \sum_{i=1}^{n} [Y_i - g(X_i)]^2,$$

and $l_{n\lambda}(g) = l_n(g) + \frac{1}{2}\lambda\|g\|^2_{\mathcal{H}}$. Then, $\widehat{\zeta}_{n\lambda} = \arg\min_{g \in \mathcal{H}} l_{n\lambda}(g)$. Denote by $l_\infty(g) = \mathbb{E}[l_n(g)]$, then $l_\infty(g) = \frac{1}{2}(\sigma^2 + \|\zeta - g\|^2_{L_2(\Pi)})$. Write

$$\bar{\zeta}_{\infty\lambda} = \arg\min_{g \in \mathcal{H}} l_{\infty\lambda}(g), \text{ where } l_{\infty\lambda}(g) = l_\infty(g) + \frac{1}{2}\lambda\|g\|^2_{\mathcal{H}}.$$

We now can decompose the estimation error of $\widehat{\zeta}_{n\lambda}$ as follows:

$$\widehat{\zeta}_{n\lambda} - \zeta = (\widehat{\zeta}_{n\lambda} - \bar{\zeta}_{\infty\lambda}) + (\bar{\zeta}_{\infty\lambda} - \zeta).$$

Here, the two terms on the right-hand side are referred to as the stochastic error and the deterministic error, respectively. We study these two terms separately in the following.

Step 1: Deterministic error Denote by $\zeta_\nu = \langle \zeta, \phi_\nu \rangle_{L_2(\Pi)}$. Then $\zeta(\cdot) = \sum_{\nu=1}^{\infty} \zeta_\nu \phi_\nu(\cdot)$. It is clear that

$$\bar{\zeta}_{\infty\lambda} = \sum_{\nu=1}^{\infty} \frac{\zeta_\nu}{1 + \lambda \cdot \lambda_\nu^{-1}} \phi_\nu(\cdot).$$

Denote $\bar{\zeta}_\nu = \zeta_\nu / (1 + \lambda \cdot \lambda_\nu^{-1})$. The following lemma gives a non-asymptotic result for the deterministic error.

Lemma E.1. *For any $n \geq 1$,*

$$\|\bar{\zeta}_{\infty\lambda} - \zeta\|_a \leq \begin{cases} \frac{1}{2}(1+a)^{(1+a)/2}(1-a)^{(1-a)/2}\lambda^{(1-a)/2}\|\zeta\|_{\mathcal{H}}, & \text{if } 0 \leq a < 1, \\ \|\zeta\|_{\mathcal{H}}, & \text{if } a = 1. \end{cases}$$

Proof. For any $0 \leq a \leq 1$, we have

$$\begin{aligned} \|\bar{\zeta}_{\infty\lambda} - \zeta\|_a^2 &= \sum_{\nu=1}^{\infty} (1 + \lambda_\nu^{-a})(\bar{\zeta}_\nu - \zeta_\nu)^2 \\ &\leq \lambda^2 \sup_{\nu \geq 1} \frac{(1 + \lambda_\nu^{-a})\lambda_\nu^{-1}}{(1 + \lambda \cdot \lambda_\nu^{-1})^2} \sum_{\nu=1}^{\infty} \lambda_\nu^{-1} \zeta_\nu^2 \\ &= \lambda^2 \|\zeta\|_{\mathcal{H}}^2 \sup_{\nu \geq 1} \frac{(1 + \lambda_\nu^{-a})\lambda_\nu^{-1}}{(1 + \lambda \cdot \lambda_\nu^{-1})^2}. \end{aligned}$$

Observe that

$$\sup_{\nu \geq 1} \frac{(1 + \lambda_\nu^{-a})\lambda_\nu^{-1}}{(1 + \lambda \cdot \lambda_\nu^{-1})^2} \leq \sup_{x > 0} \frac{(1 + x^{-a})x^{-1}}{(1 + \lambda x^{-1})^2} \leq \begin{cases} \frac{(1+a)^{1+a}(1-a)^{1-a}}{2\lambda^{a+1}}, & \text{if } 0 \leq a < 1, \\ \lambda^{-2}, & \text{if } a = 1. \end{cases}$$

This completes the proof. □

Step 2: Stochastic error For any $g, g_1, g_2 \in \mathcal{H}$, we have the first- and second-order Fréchet derivatives as follows:

$$Dl_n(g)g_1 = -\frac{1}{n}\sum_{i=1}^{n}[Y_i - g(X_i)]g_1(X_i), \quad Dl_\infty(g)g_1 = -\langle \zeta - g, g_1\rangle_{L_2(\Pi)}$$
$$D^2 l_n(g)g_1g_2 = \frac{1}{n}\sum_{i=1}^{n} g_1(X_i)g_2(X_i), \quad D^2 l_\infty(g)g_1g_2 = \langle g_1, g_2\rangle_{L_2(\Pi)}.$$

Here, the Fréchet derivatives are defined in $\|\cdot\|$-norm. Since $D^2 l_{\infty\lambda}(\bar{\zeta})\phi_\nu\phi_\mu = \langle D^2 l_{\infty\lambda}(\bar{\zeta})\phi_\nu, \phi_\mu\rangle = (1 + \lambda\cdot\lambda_\nu^{-1})\delta_{\nu\mu}$ and $\|\phi_\nu\| = (1+\lambda_\nu^{-1})^{1/2}$, we have

$$\left[D^2 l_{\infty\lambda}(\bar{\zeta})\right]^{-1}\phi_\nu = (1+\lambda\cdot\lambda_\nu^{-1})^{-1}(1+\lambda_\nu^{-1})\phi_\nu. \tag{E.1}$$

Define that

$$\tilde{\zeta}^\dagger \stackrel{\text{def}}{=} \bar{\zeta}_{\infty\lambda} - \left[D^2 l_{\infty\lambda}(\bar{\zeta}_{\infty\lambda})\right]^{-1} Dl_{n\lambda}(\bar{\zeta}_{\infty\lambda}),$$

then we can decompose the stochastic error as follows:

$$\widehat{\zeta}_{n\lambda} - \bar{\zeta}_{\infty\lambda} = (\tilde{\zeta}^\dagger - \bar{\zeta}_{\infty\lambda}) + (\widehat{\zeta}_{n\lambda} - \tilde{\zeta}^\dagger). \tag{E.2}$$

We study the two terms on the right-hand side of (E.2) separately. For simplicity of the notations, we abbreviate the subscripts of $\widehat{\zeta}_{n\lambda}$ and $\bar{\zeta}_{\infty\lambda}$ in the rest of this section.

For any $0 \le a \le 1$ and $\lambda > 0$, we define $\Delta(a,\lambda)$ by satisfying

$$\Delta(a,\lambda) \ge \lambda^{a+d/2m}\sum_{\nu=1}^{\infty}(1+\lambda_\nu^{-a})(1+\lambda\cdot\lambda_\nu^{-1})^{-2}. \tag{E.3}$$

Under Assumption 2, there exists a $\Delta(a,\lambda) < \infty$. The following lemma gives a non-asymptotic bound for $(\tilde{\zeta}^\dagger - \bar{\zeta})$ in (E.2).

Lemma E.2. *For any $0 \le a \le 1$ and $n \ge 1$, we have that with probability at least $1 - 3\exp(-\alpha^2)$,*

$$\|\tilde{\zeta}^{\dagger} - \bar{\zeta}\|_a^2 \le \left\{\alpha A \|\zeta\|_{\mathcal{H}} + c_\phi \sigma(1+\sqrt{2}\alpha)\right\}^2 \Delta(a,\lambda) n^{-1} \lambda^{-(a+d/2m)}.$$

Here, c_ϕ is define in Assumption 2, $\Delta(a,\lambda)$ is defined in (E.3), and A is a constant to be given in (E.29).

Proof. Observe that

$$Dl_{n\lambda}(\bar{\zeta}) = Dl_{n\lambda}(\bar{\zeta}) - Dl_{\infty\lambda}(\bar{\zeta}) = Dl_n(\bar{\zeta}) - Dl_\infty(\bar{\zeta}).$$

Thus,

$$\begin{aligned} &\sup_{\nu\ge 1} |Dl_{n\lambda}(\bar{\zeta})\phi_\nu| \\ &\le \sup_{\nu\ge 1} \left| \frac{1}{n}\sum_{i=1}^n \left\{(\zeta-\bar{\zeta})(X_i)\phi_\nu(X_i) - \mathbb{E}[(\zeta-\bar{\zeta})(X)\phi_\nu(X)]\right\}\right| \\ &\quad + \sup_{\nu\ge 1}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \phi_\nu(X_i)\right|. \end{aligned} \tag{E.4}$$

We consider the two terms on the right-hand side of (E.4) seperately. For the first term, since Lemma E.1 implies $\|\zeta - \bar{\zeta}\| \le \|\zeta\|_{\mathcal{H}}$, we can apply Lemma E.4 in Section E.3 by letting $g = \zeta - \bar{\zeta}$ and $t = \alpha A \|\zeta\|_{\mathcal{H}}$. Then, with probability at least $1 - 2\exp(-\alpha^2)$,

$$\sup_{\nu\ge 1}\left|\frac{1}{n}\sum_{i=1}^n \left[(\zeta-\bar{\zeta})(X_i)\phi_\nu(X_i) - \mathbb{E}\{(\zeta-\bar{\zeta})(X)\phi_\nu(X)\}\right]\right| \le \frac{\alpha A\|\zeta\|_{\mathcal{H}}}{\sqrt{n}}.$$

Then we consider the second term in (E.4). Let $\Sigma_{\phi_\nu} = [\phi_\nu(X_i)\phi_\nu(X_j)]_{1\le i,j\le n}$ and $\overrightarrow{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^\top$. The uniform version of the Hanson-Wright inequality for suprema of quadratic forms (Talagrand, 1996) gives

$$\begin{aligned} &\mathbb{P}\left(\sup_{\nu\ge 1}\sigma^{-2}\overrightarrow{\varepsilon}^\top \Sigma_{\phi_\nu}\overrightarrow{\varepsilon} > \sup_{\nu\ge 1}\left\{\mathrm{tr}(\Sigma_{\phi_\nu}) + 2\sqrt{\mathrm{tr}(\Sigma_{\phi_\nu}^2)}\alpha + 2\|\Sigma_{\phi_\nu}\|\alpha^2\right\}\right) \\ &\le \exp(-\alpha^2). \end{aligned}$$

Observe that $\mathrm{tr}(\Sigma_{\phi_\nu}) = \sum_{i=1}^n \phi_\nu^2(X_i) \le nc_\phi^2$ and

$$\|\Sigma_{\phi_\nu}\| \le \sqrt{\mathrm{tr}(\Sigma_{\phi_\nu}^2)} = \sqrt{\sum_{i,j=1}^n \phi_\nu^2(X_i)\phi_\nu^2(X_j)} \le nc_\phi^2,$$

we have probability at least $1 - \exp(-\alpha^2)$ such that

$$\sup_{\nu\ge 1}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\phi_\nu(X_i)\right| \le \frac{c_\phi\sigma(1+\sqrt{2}\alpha)}{\sqrt{n}}.$$

Therefore, (E.4) implies that with probability at least $1 - 3\exp(-\alpha^2)$,

$$\sup_{\nu\ge 1}|Dl_{n\lambda}(\bar\zeta)\phi_\nu| \le \alpha A\|\zeta\|_{\mathcal{H}} n^{-1/2} + c_\phi\sigma(1+\sqrt{2}\alpha)n^{-1/2}. \tag{E.5}$$

By the definition of $\tilde\zeta^\dagger$ and (E.1), with probability at least $1 - 3\exp(-\alpha^2)$,

$$\begin{aligned}
\|\tilde\zeta^\dagger - \bar\zeta\|_a^2 &= \|[D^2 l_{\infty\lambda}(\bar\zeta)]^{-1} Dl_{n\lambda}(\bar\zeta)\|_a^2 \\
&= \sum_{\nu=1}^\infty (1+\lambda_\nu^{-a})(1+\lambda\cdot\lambda_\nu^{-1})^{-2}[Dl_{n\lambda}(\bar\zeta)\phi_\nu]^2 \\
&\le \left\{\sup_\nu [Dl_{n\lambda}(\bar\zeta)\phi_\nu]^2\right\}\sum_{\nu=1}^\infty (1+\lambda_\nu^{-a})(1+\lambda\cdot\lambda_\nu^{-1})^{-2} \\
&\le \frac{[\alpha A\|\zeta\|_{\mathcal{H}} + c_\phi\sigma(1+\sqrt{2}\alpha)]^2}{n}\sum_{\nu=1}^\infty (1+\lambda_\nu^{-a})(1+\lambda\cdot\lambda_\nu^{-1})^{-2} \\
&\le \frac{[\alpha A\|\zeta\|_{\mathcal{H}} + c_\phi\sigma(1+\sqrt{2}\alpha)]^2}{n}\Delta(a,\lambda)\lambda^{-(a+d/2m)},
\end{aligned}$$

This completes the proof for Lemma E.2. □

For any $0 < b \le 1$, we define c_b as

$$c_b \stackrel{\text{def}}{=} \sum_{\nu=1}^\infty (1+\lambda_\nu^{-b})^{-1}. \tag{E.6}$$

Then, $c_b < \infty$ by Assumption 2. Now we give a non-asymptotic bound for $(\widehat\zeta - \tilde\zeta^\dagger)$ in (E.2).

Lemma E.3. *For any $\rho > 0$ and $\alpha > 0$, if there exists some $b \in (d/2m, 1]$ such that*

$$n^{-1}\lambda^{-(b+d/2m)} < \frac{\rho^2}{\Delta(b,\lambda)\{2\alpha^2 c_\phi^4 c_b + [\alpha A\|\zeta\|_{\mathcal{H}} + c_\phi\sigma(1+\sqrt{2}\alpha)]^2\}}, \tag{E.7}$$

then with probability at least $1 - 5\exp(-\alpha^2)$,

$$\|\widehat{\zeta} - \tilde{\zeta}^\dagger\|_a^2 < \frac{2\alpha^2 c_\phi^4 c_b \Delta(a,\lambda)\rho^2}{(1-\rho)^2} n^{-1}\lambda^{-(a+d/2m)}, \ \ \forall 0 \le a \le 1.$$

Here, c_ϕ is define in Assumption 2, $\Delta(a,\lambda)$ is defined in (E.3), c_b is defined in (E.6), and A is a constant to be given in (E.29).

Proof. By the definition of $\tilde{\zeta}^\dagger$, we have that

$$Dl_{n\lambda}(\bar{\zeta}) = D^2 l_{\infty\lambda}(\bar{\zeta})(\bar{\zeta} - \tilde{\zeta}^\dagger).$$

Since $l_{n\lambda}$ is quadratic,

$$Dl_{n\lambda}(\widehat{\zeta}) = Dl_{n\lambda}(\bar{\zeta}) + D^2 l_{n\lambda}(\bar{\zeta})(\widehat{\zeta} - \bar{\zeta}) = 0.$$

Thus,

$$D^2 l_{\infty\lambda}(\bar{\zeta})(\widehat{\zeta} - \tilde{\zeta}^\dagger) = D^2 l_\infty(\bar{\zeta})(\widehat{\zeta} - \bar{\zeta}) - D^2 l_n(\bar{\zeta})(\widehat{\zeta} - \bar{\zeta}),$$

and this implies

$$\widehat{\zeta} - \tilde{\zeta}^\dagger = [D^2 l_{\infty\lambda}(\bar{\zeta})]^{-1}(D^2 l_\infty(\bar{f}) - D^2 l_n(\bar{\zeta}))(\widehat{\zeta} - \bar{\zeta}).$$

Recall that $\bar{\zeta} = \sum_{\nu=1}^{\infty} \bar{\zeta}_\nu \phi_\nu$. Let $\widehat{\zeta} = \sum_{\nu=1}^{\infty} \widehat{\zeta}_\nu \phi_\nu$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\|\widehat{\zeta} - \tilde{\zeta}^\dagger\|_a^2 &= \sum_{\nu=1}^{\infty} (1+\lambda_\nu^{-a})(1+\lambda\cdot\lambda_\nu^{-1})^{-2} \\
&\qquad \times \left[\sum_{j=1}^{\infty} (\widehat{\zeta}_j - \bar{\zeta}_j)\left(\frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\phi_\nu(X_i) - \langle \phi_j, \phi_\nu\rangle_{L_2(\Pi)}\right)\right]^2 \\
&\le \sum_{\nu=1}^{\infty} (1+\lambda_\nu^{-a})(1+\lambda\cdot\lambda_\nu^{-1})^{-2} \left[\sum_{j=1}^{\infty} (\widehat{\zeta}_j - \bar{\zeta}_j)^2 (1+\lambda_j^{-b})\right] \\
&\qquad \times \left(\sum_{j=1}^{\infty} (1+\lambda_j^{-b})^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\phi_\nu(X_i) - \langle \phi_j, \phi_\nu\rangle_{L_2(\Pi)}\right]^2\right).
\end{aligned}$$

Since $|\phi_j(X_i)\phi_\nu(X_i)| \le c_\phi^2$ and $\phi_j(\cdot)\phi_\nu(\cdot)$ is measurable, by applying the McDiarmid's inequality, we have that with probability at least $1 - 2\exp(-\alpha^2)$,

$$\sup_{j,\nu\ge 1} \left[\frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\phi_\nu(X_i) - \langle \phi_j, \phi_\nu\rangle_{L_2(\Pi)}\right]^2 \le \frac{2\alpha^2 c_\phi^4}{n}. \tag{E.8}$$

Combining (E.3), (E.6), and (E.8), for any $d/2m < b \le 1$, we have probability at least $1 - 2\exp(-\alpha^2)$,

$$\|\widehat{\zeta} - \tilde{\zeta}^\dagger\|_a^2 \le \frac{2\alpha^2 c_\phi^4 \Delta(a,\lambda) c_b}{n} \lambda^{-(a+d/2m)} \|\widehat{\zeta} - \bar{\zeta}\|_b^2. \tag{E.9}$$

Take $a = b$, then with probability at least $1 - 2\exp(-\alpha^2)$,

$$\|\widehat{\zeta} - \tilde{\zeta}^\dagger\|_b^2 \le \frac{2\alpha^2 c_\phi^4 \Delta(b,\lambda) c_b}{n} \lambda^{-(b+d/2m)} \|\widehat{\zeta} - \bar{\zeta}\|_b^2.$$

If (E.7) holds, then $\|\widehat{\zeta} - \tilde{\zeta}\|_b < \rho \|\widehat{\zeta} - \bar{\zeta}\|_b$ and $\|\tilde{\zeta} - \bar{\zeta}\|_b \ge \|\widehat{\zeta} - \bar{\zeta}\|_b - \|\widehat{\zeta} - \tilde{\zeta}\|_b > (1-\rho)\|\widehat{\zeta} - \bar{\zeta}\|_b$. By Lemma E.2, with probability at least $1 - 5\exp(-\alpha^2)$,

$$\|\widehat{\zeta} - \bar{\zeta}\|_b^2 < \frac{[\alpha A \|\zeta\|_{\mathcal{H}} + c_\phi \sigma(1+\sqrt{2}\alpha)]^2 \Delta(b)}{(1-\delta)^2 n} \lambda^{-(b+d/2m)} < \frac{\rho^2}{(1-\rho)^2}.$$

where the second inequality is from (E.7). We complete the proof by plugging the above inequality to (E.9). □

Step 3: Putting it together We consider for $a = 0$. Let

$$\Delta(0, \lambda) = \frac{4m}{4m - d} C_\lambda^{d/2m},$$

which satisfies the definition (E.3) and does not depend on λ. Let

$$\begin{aligned} \lambda &= n^{-\frac{2m}{2m+d}} \left\{ 2\sqrt{\Delta(0,\lambda)} \left[\alpha A + \frac{\sigma c_\phi (1+\sqrt{2}\alpha)}{\|\zeta\|_{\mathcal{H}}} \right] \right\}^{\frac{4m}{2m+d}} \\ &= n^{-\frac{2m}{2m+d}} \left\{ 4\sqrt{\frac{m}{4m-d}} C_\lambda^{\frac{d}{4m}} \left[\alpha A + \frac{\sigma c_\phi (1+\sqrt{2}\alpha)}{\|\zeta\|_{\mathcal{H}}} \right] \right\}^{\frac{4m}{2m+d}}. \end{aligned} \tag{E.10}$$

Here, A is a constant not depending on $n, \sigma, \|\zeta\|_{\mathcal{H}}$. Then for any $b \in (d/2m, 1)$ and ρ satisfying

$$\rho > c_\rho n^{-\frac{m(1-b)}{2m+d}} \alpha^{\frac{2m(1-b)}{2m+d}} \left(1 + \frac{\sigma}{\|\zeta\|_{\mathcal{H}}} \right)^{-\frac{2mb+d}{2m+d}} (\|\zeta\|_{\mathcal{H}} + \sigma),$$

the λ defined in (E.10) satisfies the condition (E.7) in Lemma E.3. Thus, Lemma E.3 implies

$$\|\widehat{\zeta} - \tilde{\zeta}^\dagger\|_{L_2(\Pi)} \le \tilde{c} \alpha^{\frac{4m-d}{2m+d}} n^{-\frac{4m-d}{4m+2d}} (\|\zeta\|_{\mathcal{H}} + \sigma) \left(1 + \frac{\sigma}{\|\zeta\|_{\mathcal{H}}} \right)^{-\frac{3d}{2m+d}}, \tag{E.11}$$

for some constant $\tilde{c}$ not depending on $n, \sigma, m, d, \|\zeta\|_{\mathcal{H}}$. Combining Lemma E.1, E.2 and E.3, with probability at least $1 - 8\exp(-\alpha^2)$,

$$\begin{aligned} \|\widehat{\zeta} - \zeta\|_{L_2(\Pi)} \le C_* &\left[1 + \alpha^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{\|\zeta\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right] \\ &\cdot \alpha^{\frac{2m}{2m+d}} n^{-\frac{m}{2m+d}} (\|\zeta\|_{\mathcal{H}} + \sigma) \left(1 + \frac{\sigma}{\|\zeta\|_{\mathcal{H}}} \right)^{-\frac{d}{2m+d}}, \end{aligned}$$

for some constant C_* not depending on $n, \sigma, m, d, \|\zeta\|_{\mathcal{H}}$, and λ given by (E.10). In particular, let $\alpha = \alpha_0 = 3.36$, and we have $1 - 8\exp(-\alpha_0^2) = 99.99\%$, and which completes the proof of Theorem 6.1.

E.1.2 Lower Bound Result: Theorem 6.2

Let N be a natural number to be defined later and $b = \{b_\nu : \nu = 1, \ldots, N\} \in \{0,1\}^N$ be a length-N binary sequence. Recall that λ_νs and $\phi_\nu(\cdot)$s are the eigenvalues and eigenfunctions of $K(\cdot,\cdot)$, respectively, and they satisfy Assumption 2. We define a set of functions $\zeta_b(\cdot)$ indexed by b:

$$\zeta_b(\cdot) = c^\dagger N^{-\frac{1}{2}} \sum_{\nu=1}^{N} b_\nu \lambda_{\nu+N}^{\frac{1}{2}} \phi_{\nu+N}(\cdot),$$

where the constant $c^\dagger$ is defined as a positive root of the following equation of z:

$$\begin{aligned} z - \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{z}\right)^{-\frac{2d}{2m+d}}\right] \\ \cdot \left(2 + 2\alpha_0^{\frac{2m-d}{2m+d}}\right)^{-1} (z+\sigma) \left(1 + \frac{\sigma}{z}\right)^{-\frac{d}{2m+d}} = 0. \end{aligned} \tag{E.12}$$

A positive root exists since left-hand side of equation is smaller than 0 when $z > 0$ and $z \to 0$ and greater than 0 when $z \to +\infty$, and it is continuous in z. By definition, $\zeta_b(\cdot)$ is a finite linear combination of kernel eigenfunctions. Moreover,

$$\|\zeta_b(\cdot)\|_{\mathcal{H}}^2 = (c^\dagger)^2 N^{-1} \sum_{\nu=1}^{N} b_\nu^2 \leq (c^\dagger)^2, \tag{E.13}$$

which is finite.

By the Varshamov-Gilbert bound, there exists a collection of binary sequences $\{b^{(1)}, \ldots, b^{(M)}\} \subset \{0,1\}^N$ such that $M \geq e^{N/8}$ with the pairwise Hamming distance satisfying

$$H(b^{(l)}, b^{(q)}) \geq \frac{N}{8}, \quad \forall 1 \leq l < q \leq M.$$

From Assumption 2, we have that for any $b^{(l)}, b^{(q)} \in \{0,1\}^N$,

$$\begin{aligned}\|\zeta_{b^{(l)}} - \zeta_{b^{(q)}}\|^2_{L_2(\Pi)} &\geq c_\lambda (c^\dagger)^2 N^{-1} (2N)^{-\frac{2m}{d}} \sum_{\nu=1}^{N} \left[b^{(l)}_{\boldsymbol{\nu}} - b^{(q)}_{\boldsymbol{\nu}}\right]^2 \\ &\geq c_\lambda (c^\dagger)^2 N^{-1} (2N)^{-\frac{2m}{d}} \frac{N}{8} \\ &= 2^{-3-\frac{2m}{d}} c_\lambda (c^\dagger)^2 N^{-\frac{2m}{d}}.\end{aligned}$$

On the other hand by Assumption 2, we have that for any $b^{(l)} \in \{b^{(1)}, \ldots, b^{(M)}\}$,

$$\begin{aligned}\|\zeta_{b^{(l)}}\|^2_{L_2(\Pi)} &\leq C_\lambda (c^\dagger)^2 N^{-1} N^{-\frac{2m}{d}} \sum_{\nu=1}^{N} \left[b^{(l)}_{\boldsymbol{\nu}}\right]^2 \\ &\leq C_\lambda (c^\dagger)^2 N^{-\frac{2m}{d}}.\end{aligned}$$

Following a standard argument, the lower bound of estimating the true function $\zeta(\cdot)$ can be reduced to the error probability in a multi-way hypothesis test (Chapter 2 of Tsybakov (2009)). Let Θ be a random variable uniformly distributed on the discrete set $\{1, \ldots, M\}$ and let the true function $\zeta = \zeta_{b^{(\Theta)}}$. Let $\widehat{\Theta}$ be an estimator of Θ based on the data $\{(X_i, Y_i) : i = 1, \ldots, n\}$ that are generated by (6.1). Following Chapter 2 of Tsybakov (2009),

$$\begin{aligned}&\infty_{\widetilde{\zeta}_n} \sup_{\zeta \in \mathcal{H}} \mathbb{P}\left\{\|\widetilde{\zeta_n} - \zeta\|^2_{L_2(\Pi)} \geq \frac{1}{4} \min_{b^{(l)} \neq b^{(q)}} \|\zeta_{b^{(l)}} - \zeta_{b^{(q)}}\|^2_{L_2(\Pi)}\right\} \\ &\geq \infty_{\widehat{\Theta} \in \{1,\ldots,M\}} \mathbb{P}\{\widehat{\Theta} \neq \Theta\} \\ &= \infty_{\widehat{\Theta} \in \{1,\ldots,M\}} \mathbb{E}_{X_1,\ldots,X_n} \mathbb{P}\left\{\widehat{\Theta} \neq \Theta | X_1, \ldots, X_n\right\},\end{aligned} \tag{E.14}$$

where the infimum is taken over all estimators $\widetilde{\zeta}_n$ that are measurable functions of the data $\{(X_i, Y_i) : i = 1, \ldots, n\}$. By Fano's lemma (Chapter 2 of Tsybakov (2009)),

$$\mathbb{P}\left\{\widehat{\Theta} \neq \Theta | X_1, \ldots, X_n\right\} \geq 1 - \frac{1_{X_1,\ldots,X_n}(Y_1, \ldots, Y_n; \Theta) + \log 2}{\log M}, \tag{E.15}$$

where $1_{X_1,\ldots,X_n}(Y_1, \ldots, Y_n)$ is the mutual information between $\{Y_1, \ldots, Y_n\}$ and Θ with $\{X_1, \ldots, X_n\}$ being held fixed. Let $\mathcal{K}(\cdot|\cdot)$ be the Kullback-Leibler distance and $\mathbf{P}_\zeta$ be the

conditional distribution of Y_is given $\{X_1, \ldots, X_n\}$. Thus,

$$
\begin{aligned}
&\mathbb{E}_{X_1,\ldots,X_n}\left[1_{X_1,\ldots,X_n}\left(Y_1,\ldots,Y_n;\Theta\right)\right] \\
&\leq \binom{M}{2}^{-1} \sum_{b^{(l)}\neq b^{(q)}} \mathbb{E}_{X_1,\ldots,X_n} \mathcal{K}\left(\mathbf{P}_{\zeta_{b^{(l)}}} | \mathbf{P}_{\zeta_{b^{(q)}}}\right) \\
&\leq \frac{n}{2}\binom{M}{2}^{-1} \sum_{b^{(l)}\neq b^{(q)}} \mathbb{E}_{X_1,\ldots,X_n}\left[\frac{1}{n\sigma^2}\sum_{i=1}^{n}\left(\zeta_{b^{(l)}}(X_i)-\zeta_{b^{(q)}}(X_i)\right)^2\right] \\
&\leq \frac{n}{2\sigma^2}\binom{M}{2}^{-1} \sum_{b^{(l)}\neq b^{(q)}} \|\zeta_{b^{(l)}}-\zeta_{b^{(q)}}\|_{L_2(\Pi)}^2 \\
&\leq \frac{n}{2\sigma^2} \max_{b^{(l)}\neq b^{(q)}} \|\zeta_{b^{(l)}}-\zeta_{b^{(q)}}\|_{L_2(\Pi)}^2 \\
&\leq \frac{2n}{\sigma^2} \max_{b^{(l)}\in\{b^{(1)},\ldots,b^{(M)}\}} \|\zeta_{b^{(l)}}\|_{L_2(\Pi)}^2 \\
&\leq \frac{2n}{\sigma^2} C_\lambda (c^\dagger)^2 N^{-\frac{2m}{d}}.
\end{aligned}
$$

Combining (E.14) and (E.15) yields that

$$
\begin{aligned}
&\infty_{\widetilde{\zeta}_n} \sup_{\zeta\in\mathcal{H}} \mathbb{P}\left\{\|\widetilde{\zeta}_n-\zeta\|_{L_2(\Pi)}^2 \geq 2^{-5-\frac{2m}{d}} c_\lambda (c^\dagger)^2 N^{-\frac{2m}{d}}\right\} \\
&\geq 1-\frac{1}{\log M}\left[\mathbb{E}1_{X_1,\ldots,X_n}(Y_1,\ldots,Y_n;\Theta)+\log 2\right] \\
&\geq 1-\frac{16C_\lambda(c^\dagger)^2 n}{N^{1+\frac{2m}{d}}\sigma^2}-\frac{8\log 2}{N}.
\end{aligned}
$$

Define $N = c^* n^{d/(2m+d)}$, where

$$
c^* \stackrel{\text{def}}{=} \alpha_0^{-\frac{2d}{2m+d}} \left(2+2\alpha_0^{\frac{2m-d}{2m+d}}\right)^{-\frac{d}{m}} C_*^{-\frac{d}{2m}} c_\lambda^{\frac{d}{2m}} 2^{-\frac{5d}{2m}-1}.
$$

By the definition of $c^\dagger$, and by the fact that the right-hand side of (E.12) is monotone increasing

in $z > 0$, and by the inequality (E.13), when σ is held fixed,

$$\infty_{\widetilde{\zeta}_n} \sup_{\zeta \in \mathcal{H}} \mathbb{P} \left\{ \|\widetilde{\zeta}_n - \zeta\|_{L_2(\Pi)}^2 \geq \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{\|\zeta\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right]^2 \right.$$
$$\left. \cdot C_* n^{-\frac{2m}{2m+d}} \alpha_0^{\frac{4m}{2m+d}} (\|\zeta\|_{\mathcal{H}} + \sigma)^2 \left(1 + \frac{\sigma}{\|\zeta\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right\}$$
$$\geq 1 - \frac{16 C_\lambda \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{\|\zeta\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right]^2 (\|\zeta\|_{\mathcal{H}} + \sigma)^2}{\left(2 + 2\alpha_0^{\frac{2m-d}{2m+d}} \right)^2 \left(1 + \frac{\sigma}{\|\zeta\|_{\mathcal{H}}} \right)^{\frac{2d}{2m+d}} (c^*)^{1+\frac{2m}{d}} \sigma^2} - \frac{8 \log 2}{c^* n^{\frac{d}{2m+d}}}.$$

The right-hand side of above inequality is monotone decreasing in σ and n. Hence, there exists some constants $0 < \sigma_0, n_0 < \infty$ such that for any $\sigma \geq \sigma_0$ and $n \geq n_0$, the right-hand side of above inequality is positive. This completes the proof.

E.1.3 The Optimal Calibration Result: Proposition 6.3

In order to derive an equivalent form for $\theta^{\text{opt-pred}}$ that holds uniformly for $\zeta \in \mathcal{H}$, it is necessary to use the minimax optimal risk for estimating $\delta(\cdot, \theta)$ for all $\delta(\cdot, \theta) \in \mathcal{H}$. This is because $\delta(\cdot, \theta) = \zeta(\cdot) - \eta(\cdot, \theta)$, and $\delta(\cdot, \theta) \in \mathcal{H}$ holds for any $\theta \in \Theta$ by Assumption 1, and computer models $\eta(\cdot, \theta)$ are given functions. By replacing $\zeta(\cdot)$ with $\delta(\cdot, \theta)$ in Theorems 6.1 and 6.2, we have $\forall \theta \in \mathcal{H}$,

$$\begin{aligned} &\infty_{\widetilde{\delta}_n} \sup_{\delta \in \mathcal{H}} \|\widetilde{\delta}_n(\cdot, \theta) - \delta(\cdot, \theta)\|_{L_2(\Pi)}^2 \\ &= C_* \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{\|\delta(\cdot, \theta)\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right]^2 \\ &\quad \cdot \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m}{2m+d}} (\|\delta(\cdot, \theta)\|_{\mathcal{H}} + \sigma)^2 \left(1 + \frac{\sigma}{\|\delta(\cdot, \theta)\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}}. \end{aligned} \tag{E.16}$$

By the definition of $\theta^{\text{opt-pred}}$ in (6.4),

$$\begin{aligned}\theta^{\text{opt-pred}} &= \arg\min_{\theta\in\Theta}\left\{\infty_{\widetilde{\delta}_n}\|\delta(\cdot,\theta)-\widetilde{\delta}_n(\cdot,\theta)\|_{L_2(\Pi)}\right\}, \quad \forall\zeta\in\mathcal{H}\\ &= \arg\min_{\theta\in\Theta}\left\{\infty_{\widetilde{\delta}_n}\sup_{\delta\in\mathcal{H}}\|\delta(\cdot,\theta)-\widetilde{\delta}_n(\cdot,\theta)\|_{L_2(\Pi)}\right\}\\ &= \arg\min_{\theta\in\Theta}\left\{\|\delta(\cdot,\theta)\|_{\mathcal{H}}\right\} = \arg\min_{\theta\in\Theta}\left\{\|\zeta(\cdot)-\eta(\cdot,\theta)\|_{\mathcal{H}}\right\},\end{aligned}$$

where the third step is by the fact that parameters n, σ, m, d are fixed in the setting of computer model calibrations and the right-hand side of (E.16) is monotonically increasing as the RKHS norm $\|\delta(\cdot,\theta)\|_{\mathcal{H}}$ increases. The last step above is by the definition of $\delta(\cdot,\theta)$.

E.1.4 The Optimal Prediction Result: Corollaries 6.4 and 6.5

By the same arguments as the proof of Theorem 6.1 and replace $\zeta(\cdot)$ with $\delta(\cdot,\theta)$, we have that for any $\theta\in\Theta$, $n\geq 1$, with probability at least 99.99%,

$$\begin{aligned}&\min_{\lambda>0}\|\widehat{\delta}_{n\lambda}(\cdot,\theta)-\delta(\cdot,\theta)\|^2_{L_2(\Pi)}\\ &\leq C_*\left[1+\alpha_0^{\frac{2m-d}{2m+d}}n^{-\frac{2m-d}{4m+2d}}\left(1+\frac{\sigma}{\|\delta(\cdot,\theta)\|_{\mathcal{H}}}\right)^{-\frac{2d}{2m+d}}\right]^2\\ &\quad\cdot\alpha_0^{\frac{4m}{2m+d}}n^{-\frac{2m}{2m+d}}\left(\|\delta(\cdot,\theta)\|_{\mathcal{H}}+\sigma\right)^2\left(1+\frac{\sigma}{\|\delta(\cdot,\theta)\|_{\mathcal{H}}}\right)^{-\frac{2d}{2m+d}},\end{aligned}$$

where the $C_*>0$ is the same as in Theorem 6.1 and it does not depend on $n, \sigma, \|\delta(\cdot,\theta)\|_{\mathcal{H}}$. The optimal λ in the above inequality is

$$\lambda = n^{-\frac{2m}{2m+d}}\left\{4\sqrt{\frac{m}{4m-d}}C_\lambda^{\frac{d}{4m}}\left[\alpha A+\frac{\sigma c_\phi(1+\sqrt{2}\alpha)}{\|\zeta(\cdot)-\eta(\cdot,\theta^{\text{opt-pred}})\|_{\mathcal{H}}}\right]\right\}^{\frac{4m}{2m+d}},$$

where A is defined in the proof of Theorem 6.1 and it does not depend on quantities $n, \sigma, \|\zeta-\eta(\cdot,\theta^{\text{opt-pred}})\|_{\mathcal{H}}$. By Proposition 6.3, we prove the Corollary 6.4.

Now we show Corollary 6.5. Similar to Theorem 6.2 and replacing $\zeta(\cdot)$ with $\delta(\cdot,\theta)$, we

have that for any $\theta \in \Theta$,

$$\begin{aligned}
&\infty_{\widetilde{\delta}_n} \sup_{\delta(\cdot,\theta)\in\mathcal{H}} \mathbb{P}\left\{ \|\widetilde{\delta}_n(\cdot,\theta) - \delta(\cdot,\theta)\|^2_{L_2(\Pi)} \right. \\
&\geq C_* \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{\|\delta(\cdot,\theta)\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right]^2 \\
&\quad \left. \cdot \alpha_0^{\frac{4m}{2m+d}} n^{-\frac{2m}{2m+d}} \left(\|\delta(\cdot,\theta)\|_{\mathcal{H}} + \sigma \right)^2 \left(1 + \frac{\sigma}{\|\delta(\cdot,\theta)\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right\} > 0.
\end{aligned}$$

By Proposition 6.3, we complete the proof.

E.1.5 Improvement of Prediction by Computer Models: Theorem 6.6

By comparing the finite-sample minimax risks of $\zeta_{n\lambda}^{\text{opt-pred}}$ and $\widehat{\zeta}_{n\lambda}(\cdot)$, if (6.9) holds,

$$\min_{\lambda>0} \sup_{\zeta\in\mathcal{H}} \|\zeta_{n\lambda}^{\text{opt-pred}}(\cdot) - \zeta(\cdot)\|^2_{L_2(\Pi)} < \min_{\lambda>0} \sup_{\zeta\in\mathcal{H}} \|\widehat{\zeta}_{n\lambda}(\cdot) - \zeta(\cdot)\|^2_{L_2(\Pi)},$$

where $\zeta_{n\lambda}^{\text{opt-pred}}(\cdot)$ and $\widehat{\zeta}_{n\lambda}(\cdot)$ are defined by (6.8) and (6.5), respectively. This completes the proof.

E.2 Proofs for Section 6.4

E.2.1 Consistency Result: Proposition 6.7

Recall the model discrepancy is defined by $\delta(\cdot,\theta) = \zeta(\cdot) - \eta(\cdot,\theta)$ for any $\theta \in \Theta$. We have data given by $\{(X_i, Y_i - \eta(X_i,\theta)) : i = 1,\ldots,n\}$. The GCV estimate, denoted by $\lambda_{\text{G}}(\theta)$, is consistent for minimizing the predictive mean squared errors (MSE) (Li, 1986; Wahba, 1990). Note that $\lambda_{\text{G}}(\theta)$ depends on θ with the given data. Following the same argument in Section E.1.1, the optimal λ^{opt} for minimizing the predictive MSE of $\delta(\cdot,\theta)$ is given by (E.10) with $\zeta(\cdot)$ replaced by $\delta(\cdot,\theta)$. That is, as $n \to \infty$,

$$\lambda_{\text{G}}(\theta) \to_{\mathbb{P}} \lambda^{\text{opt}} = n^{-\frac{2m}{2m+d}} \left\{ 2\sqrt{\frac{4m}{4m-d}} C_\lambda^{d/4m} \left[\alpha A + \frac{\sigma c_\phi (1+\sqrt{2}\alpha)}{\|\delta(\cdot,\theta)\|_{\mathcal{H}}} \right] \right\}^{\frac{4m}{2m+d}}.$$

Here, the constants $C_\lambda, \alpha, A, c_\phi$ are specified in Section E.1.1. It is obvious that λ^{opt} in the above equation is monotone decreasing as $\|\delta(\cdot,\theta)\|_{\mathcal{H}}$ increases.

Now consider the objective function in (6.16) when λ is selected by GCV. Denote the vector of random noises $\overrightarrow{\varepsilon} \stackrel{\text{def}}{=} (\varepsilon_1, \ldots, \varepsilon_n)^\top$. Then, if θ is at convergence,

$$\begin{aligned}
&(\overrightarrow{Y} - \eta(\overrightarrow{X}, \theta))^\top (\Sigma + n\lambda_{\text{G}}(\theta)\mathbf{I})^{-1} (\overrightarrow{Y} - \eta(\overrightarrow{X}, \theta)) \\
&= \overrightarrow{\varepsilon}^\top (\Sigma + n\lambda_{\text{G}}(\theta)\mathbf{I})^{-1} \overrightarrow{\varepsilon} + \delta(\overrightarrow{X}, \theta)^\top (\Sigma + n\lambda_{\text{G}}(\theta)\mathbf{I})^{-1} \delta(\overrightarrow{X}, \theta) \\
&\quad + 2\overrightarrow{\varepsilon}^\top (\Sigma + n\lambda_{\text{G}}(\theta)\mathbf{I})^{-1} \delta(\overrightarrow{X}, \theta).
\end{aligned} \tag{E.17}$$

Let the eigenvector eigenvalue decomposition of Σ be $UDU^\top$, where U is orthogonal and D is diagonal with the νth element $\lambda_{\nu n} > 0$. By Assumption 2, it is known that $\lambda_{\nu n} \asymp n\nu^{-2m/d}$ when n is large (Koltchinskii and Giné, 2000). Here, we write for two positive sequences a_n and b_n, $a_n \asymp b_n$ if a_n/b_n is bounded away from zero and infinity. Denote that $U^\top \delta(\overrightarrow{X}, \theta) = (\delta_{1n}, \ldots, \delta_{nn})^\top$ and $U^\top \overrightarrow{\varepsilon} = (\varepsilon_{1n}, \ldots, \varepsilon_{nn})^\top \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. We study three terms on the right-hand side of (E.17) separately.

The first term can be written as

$$\overrightarrow{\varepsilon}^\top (\Sigma + n\lambda_{\text{G}}(\theta)\mathbf{I})^{-1} \overrightarrow{\varepsilon} = \sum_{\nu=1}^{n} \frac{\varepsilon_{\nu n}^2}{\lambda_{\nu n} + n\lambda_{\text{G}}(\theta)},$$

which is a sum of weighted chi-square random variables. By standard concentration inequalities (Boucheron et al., 2013), for any $\gamma \in (0,1)$,

$$\begin{aligned}
(1-\gamma) \sum_{\nu=1}^{n} \frac{\sigma^2}{\lambda_{\nu n} + n\lambda_{\text{G}}(\theta)} &\le \sum_{\nu=1}^{n} \frac{\varepsilon_{\nu n}^2}{\lambda_{\nu n} + n\lambda_{\text{G}}(\theta)} \\
&\le \sum_{\nu=1}^{n} \frac{\sigma^2}{\lambda_{\nu n} + n\lambda_{\text{G}}(\theta)} + 2\gamma \sum_{\nu=1}^{n} \left(\frac{\sigma^2}{\lambda_{\nu n} + n\lambda_{\text{G}}(\theta)} \right)^2
\end{aligned}$$

holds with probability at least $1-\exp(-\gamma^2 n/4)-\exp(-2\gamma^2 n^{(2m-d)/(2m+d)})$. Note that

$$\begin{aligned}\sum_{\nu=1}^{n}\frac{\sigma^2}{\lambda_{\nu n}+n\lambda_{\mathrm{G}}(\theta)} &\asymp \frac{\sigma^2}{n}\sum_{\nu=1}^{n}\frac{1}{\nu^{-2m/d}+\lambda_{\mathrm{G}}(\theta)}\\ &\asymp \frac{\sigma^2}{n}\int_{x=1}^{n}\frac{1}{x^{-2m/d}+\lambda_{\mathrm{G}}(\theta)} \asymp \sigma^2\lambda_{\mathrm{G}}^{-1}(\theta),\end{aligned}$$

and similarly,

$$\left(\sum_{\nu=1}^{n}\frac{\sigma^2}{\lambda_{\nu n}+n\lambda_{\mathrm{G}}(\theta)}\right)^2 \asymp \frac{\sigma^4}{n^2}\sum_{\nu=1}^{n}\left(\frac{1}{\nu^{-2m/d}+\lambda_{\mathrm{G}}(\theta)}\right)^2 \asymp \sigma^4 n^{-1}\lambda_{\mathrm{G}}^{-2}(\theta),$$

Since $\lambda_{\mathrm{G}}(\theta)\asymp n^{-2m/(2m+d)}$, we have that

$$\left(\sum_{\nu=1}^{n}\frac{\sigma^2}{\lambda_{\nu n}+n\lambda_{\mathrm{G}}(\theta)}\right)^2 = o\left\{\sum_{\nu=1}^{n}\frac{\sigma^2}{\lambda_{\nu n}+n\lambda_{\mathrm{G}}(\theta)}\right\}.$$

Thus when n is large, by letting $\gamma = n^{-(2m-d)/(6m+3d)}$,

$$\overrightarrow{\varepsilon}^{\top}(\Sigma+n\lambda_{\mathrm{G}}(\theta)\mathbf{I})^{-1}\overrightarrow{\varepsilon} = \sum_{\nu=1}^{n}\frac{\sigma^2}{\lambda_{\nu n}+n\lambda_{\mathrm{G}}(\theta)} \tag{E.18}$$

holds with probability approaching one.

The second term on the right-hand side of (E.17) satisfies

$$\begin{aligned}&\delta(\overrightarrow{X},\theta)^{\top}(\Sigma+n\lambda_{\mathrm{G}}(\theta)\mathbf{I})^{-1}\delta(\overrightarrow{X},\theta)\\ &=\sum_{\nu=1}^{n}\frac{\delta_{\nu n}^2}{\lambda_{\nu n}+n\lambda_{\mathrm{G}}(\theta)} \le \sum_{\nu=1}^{n}\frac{\delta_{\nu n}^2}{\lambda_{\nu n}} \le \|\delta(\cdot,\theta)\|_{\mathcal{H}}^2,\end{aligned} \tag{E.19}$$

where the last step can be proved as follows. Let $h\in\mathcal{H}$ minimize $\|h\|_{\mathcal{H}}^2$ subject to $h(X_i)=\delta(X_i,\theta)$. Then $h(\cdot)=\sum_{i=1}^{n}c_i^h K(\cdot,x_i)$ and $c^h=\Sigma^{-1}\delta(\overrightarrow{X},\theta)$. Then

$$\|\delta(\cdot,\theta)\|_{\mathcal{H}}^2 \ge \|h(\cdot)\|_{\mathcal{H}}^2 = \left(c^h\right)^{\top}\Sigma c^h = \sum_{\nu=1}^{n}\delta_{\nu n}^2/\lambda_{\nu n}.$$

Thus, comparing (E.19) with (E.18), when n is large,

$$\delta(\overrightarrow{X},\theta)^\top(\Sigma+n\lambda_{\mathrm{G}}(\theta)\mathbf{I})^{-1}\delta(\overrightarrow{X},\theta)=o\left\{\overrightarrow{\varepsilon}^\top(\Sigma+n\lambda_{\mathrm{G}}(\theta)\mathbf{I})^{-1}\overrightarrow{\varepsilon}\right\}.$$

The third term on the right-hand side of (E.17), by the Cauchy-Schwarz inequality, satisfies

$$\overrightarrow{\varepsilon}^\top(\Sigma+n\lambda_{\mathrm{G}}(\theta)\mathbf{I})^{-1}\delta(\overrightarrow{X},\theta)=o\left\{\overrightarrow{\varepsilon}^\top(\Sigma+n\lambda_{\mathrm{G}}(\theta)\mathbf{I})^{-1}\overrightarrow{\varepsilon}\right\}. \tag{E.20}$$

Combining (E.18), (E.19) and (E.20), we know that the left-hand side of (E.17) indeed satisfies, when n is large,

$$\begin{aligned}&(\overrightarrow{Y}-\eta(\overrightarrow{X},\theta))^\top(\Sigma+n\lambda_{\mathrm{G}}(\theta)\mathbf{I})^{-1}(\overrightarrow{Y}-\eta(\overrightarrow{X},\theta))\\&=\overrightarrow{\varepsilon}^\top(\Sigma+n\lambda_{\mathrm{G}}(\theta)\mathbf{I})^{-1}\overrightarrow{\varepsilon}=\sum_{\nu=1}^{n}\frac{\sigma^2}{\lambda_{\nu n}+n\lambda_{\mathrm{G}}(\theta)},\end{aligned}$$

which is decreasing as $\lambda_{\mathrm{G}}(\theta)$ increases. Recall that $\lambda_{\mathrm{G}}(\theta)$ is monotone decreasing as $\|\delta(\cdot,\theta)\|_{\mathcal{H}}=\|\zeta(\cdot)-\eta(\cdot,\theta)\|_{\mathcal{H}}$ increases. Therefore, minimizing over $\theta\in\Theta$ for

$$(\overrightarrow{Y}-\eta(\overrightarrow{X},\theta))^\top(\Sigma+n\lambda_{\mathrm{G}}(\theta)\mathbf{I})^{-1}(\overrightarrow{Y}-\eta(\overrightarrow{X},\theta))$$

leads to the minimizer $\widehat{\theta}_n^{\text{opt-pred}}$ that also minimizes $\|\zeta(\cdot)-\eta(\cdot,\theta)\|_{\mathcal{H}}$. By Proposition 6.3, we complete the proof.

E.2.2 Comparison with Frequentist Calibrations: Remark 6.8

For the calibration part, Proposition 6.7 establishes $\widehat{\theta}_n^{\text{opt-pred}}\to_{\mathbb{P}}\theta^{\text{opt-pred}}$ and Tuo and Wu (2015) has shown that $\widehat{\theta}_n^{L_2}\to_{\mathbb{P}}\theta^{L_2}$. On the other hand, it is known that $\widehat{\theta}_n^{l_2}$ converges to the minimizer of Kullback-Leibler distance between $\zeta(\cdot)$ and $\eta(\cdot,\theta)$ (see, e.g., White (1982)), that is, $\widehat{\theta}_n^{l_2}\to_{a.s.}\theta^{L_2}$, which also implies that $\widehat{\theta}_n^{l_2}\to_{\mathbb{P}}\theta^{L_2}$.

Then we consider the prediction part. Both the predictors, $\eta(\cdot,\widehat{\theta}_n^{L_2})$ in Tuo and Wu (2015) and $\eta(\cdot,\widehat{\theta}_n^{l_2})+\widehat{\delta}_{n\lambda}(\cdot,\widehat{\theta}_n^{l_2})$ in Wong et al. (2017), are based on calibrations around θ^{L_2}. Note

that as $\lambda \to \infty$, the regularized estimator defined in (6.7) satisfies $\widehat{\delta}_{n\lambda}(\cdot, \theta^{L_2}) \to 0$ where $\theta = \theta^{L_2}$ (see, e.g., Wahba (1990)). Then

$$\|\eta(\cdot, \theta^{L_2}) - \zeta(\cdot)\|^2_{L_2(\Pi)} \geq \min_{\lambda > 0} \|\eta(\cdot, \theta^{L_2}) + \widehat{\delta}_{n\lambda}(\cdot, \theta^{L_2}) - \zeta(\cdot)\|^2_{L_2(\Pi)}. \tag{E.21}$$

That is, the predictor with discrepancy estimator in Wong et al. (2017) would achieve smaller predictive MSE than the predictor without discrepancy estimator in Tuo and Wu (2015). Similar to Corollary 6.4 and 6.5, we can show that the right-hand side of (E.21) achieves the minimax optimal risk with $\theta = \theta^{L_2}$ and

$$\begin{aligned}
&\min_{\lambda > 0} \|\eta(\cdot, \theta^{L_2}) + \widehat{\delta}_{n\lambda}(\cdot, \theta^{L_2}) - \zeta(\cdot)\|^2_{L_2(\Pi)} \\
&= C_* \left[1 + \alpha_0^{\frac{2m-d}{2m+d}} n^{-\frac{2m-d}{4m+2d}} \left(1 + \frac{\sigma}{\|\delta(\cdot, \theta^{L_2})\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}} \right]^2 \alpha_0^{\frac{4m}{2m+d}} \\
&\quad \cdot n^{-\frac{2m}{2m+d}} \left(\|\delta(\cdot, \theta^{L_2})\|_{\mathcal{H}} + \sigma \right)^2 \left(1 + \frac{\sigma}{\|\delta(\cdot, \theta^{L_2})\|_{\mathcal{H}}} \right)^{-\frac{2d}{2m+d}}
\end{aligned}$$

holds with probability at least 99.99%, where the constants C_*, α_0 are defined in Theorem 6.1. Comparing this result with the minimax optimal risk when $\theta = \theta^{\text{opt-pred}}$ as shown in Section 6.3.2, we have that the predictors based on $\theta^{\text{opt-pred}}$ achieve a smaller minimax predictive mean squared error compared to predictors based on θ^{L_2}, since

$$\|\delta(\cdot, \theta^{L_2})\|_{\mathcal{H}} \geq \|\delta(\cdot, \theta^{\text{opt-pred}})\|_{\mathcal{H}} = \min_{\theta \in \Theta} \|\delta(\cdot, \theta)\|_{\mathcal{H}}.$$

This completes the proof.

E.2.3 Posterior Mean of Calibration Parameters

In this section, we give details on the classical result that under Gaussian process priors, the Bayesian posterior mean of θ and $\delta(\cdot)$ agrees with (6.12) given a special choice of λ.

We consider the following Gaussian process priors:

$$\begin{aligned} &\zeta(x) = \eta(x,\theta) + \delta(x), \text{ where} \\ &\eta(x,\theta) = \sum_{j=1}^{p} \theta_j h_j(x),\ \theta \sim \mathcal{N}(0, \alpha I),\ \delta(x) \sim \text{GP}\{0, \beta K(\cdot,\cdot)\}, \end{aligned} \tag{E.22}$$

where $h_j(x)$s are deterministic functions, GP stands for Gaussian stochastic process, and α, β are positive hyperparameters. The kernel $K(\cdot,\cdot)$ is associated with the RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$. Although the GP assumption in (E.22) implies that $\delta(\cdot) \notin \mathcal{H}$ with probability one (see, Driscoll (1973)), the well-known duality between RKHS and Hilbert space spanned by a family of Gaussian variables (see, e.g., Wahba (1990)) ensures that Bayesian estimates under GP priors are RKHS regularized estimates.

The proof here is adapted from Wahba (1990). Let T be the $n \times n$ matrix with ijth entry $h_j(X_i)$. Under the priors in (E.22), we have that

$$\mathbb{E}[\zeta(x)|\overrightarrow{Y}] = (h_1(x), \ldots, h_M(x))\mathbb{E}[\theta|\overrightarrow{Y}] + \mathbb{E}[\delta(x)|\overrightarrow{Y}]. \tag{E.23}$$

On the other hand,

$$\begin{aligned} &\mathbb{E}[\overrightarrow{Y}\overrightarrow{Y}^\top] = \alpha T T^\top + \beta\Sigma + \sigma^2 I, \\ &\mathbb{E}[\zeta(x)\overrightarrow{Y}] = \alpha T \begin{pmatrix} h_1(x,\theta) \\ \vdots \\ h_M(x,\theta) \end{pmatrix} + \beta \begin{pmatrix} K(X_1,x) \\ \vdots \\ K(X_n,x) \end{pmatrix}. \end{aligned}$$

This yields the posterior mean given by

$$\begin{aligned} &\mathbb{E}[\zeta(x)|\overrightarrow{Y}] \\ &= \{\mathbb{E}[\zeta(x)\overrightarrow{Y}]\}^\top \{\mathbb{E}[\overrightarrow{Y}\overrightarrow{Y}^\top]\}^{-1}\overrightarrow{Y} \\ &= (h_1(x), \ldots, h_M(x))\, \alpha\beta^{-1}T^\top \left(\alpha\beta^{-1}TT^\top + \Sigma + \sigma^2\beta^{-1}I\right)^{-1}\overrightarrow{Y} \\ &\quad + (K(X_1,x), \ldots, K(X_n,x)) \left(\alpha\beta^{-1}TT^\top + \Sigma + \sigma^2\beta^{-1}I\right)^{-1}\overrightarrow{Y}. \end{aligned} \tag{E.24}$$

Note that

$$\begin{aligned}\hat{\theta}^{\dagger} &\stackrel{\text{def}}{=} \lim_{\alpha\to\infty} \alpha\beta^{-1}T^{\top}(\alpha\beta^{-1}TT^{\top} + \Sigma + \sigma^2\beta^{-1}I)^{-1} \\ &= \{T^{\top}(\Sigma + \sigma^2\beta^{-1}I)^{-1}T\}^{-1}T^{\top}(\Sigma + \sigma^2\beta^{-1}I)^{-1},\end{aligned}$$

and

$$\begin{aligned}\hat{\delta}^{\dagger} &\stackrel{\text{def}}{=} \lim_{\alpha\to\infty} (\alpha\beta^{-1}TT^{\top} + \Sigma + \sigma^2\beta^{-1}I)^{-1} \\ &= (\Sigma + \sigma^2\beta^{-1}I)^{-1}\{I - T(T^{\top}(\Sigma + \sigma^2\beta^{-1}I)^{-1}T)^{-1}T^{\top}(\Sigma + \sigma^2\beta^{-1}I)^{-1}\}.\end{aligned}$$

From Chapter 1 of Wahba (1990), $(\hat{\theta}^{\dagger}, \hat{\delta}^{\dagger})$ is the solution to the smoothing splines,

$$(\widehat{\theta}^{\dagger}, \widehat{\delta}^{\dagger}) = \underset{\theta\in\Theta, \delta\in\mathcal{H}}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p}\theta_j h_j(X_i) - \delta(X_i)\right]^2 + \frac{\sigma^2}{n\beta}\|\delta\|_{\mathcal{H}}^2 \right\}.$$

Note that the above objective function is same as (6.12) by letting $\lambda = \sigma^2/n\beta$.

Compare (E.23) with (E.24), we conclude that under the prior (E.22) with improper $\alpha \to \infty$, the posterior mean of θ and $\delta(\cdot)$ agrees with the minimizer of objective function (6.12) if $\lambda = \sigma^2/n\beta$.

E.3 Key Lemma

Lemma E.4. *Recall that $R(\cdot,\cdot)$ is the reproducing kernel associated with $(\mathcal{H}, \|\cdot\|)$. Suppose that $c_R = \sup_{x\in\Omega}\sqrt{R(x,x)}$ is finite. Then for any $t \geq 0$ and $\nu \geq 1$, we have*

$$\begin{aligned}&\mathbb{P}\left(\sup_{\substack{\nu\geq 1 \\ g:\|g\|\leq\|\zeta\|_{\mathcal{H}}}} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{g(X_i)\phi_\nu(X_i) - \mathbb{E}[g(X)\phi_\nu(X)]\}\right| \geq t\right) \\ &\leq 2\exp\left(-\frac{t^2}{A^2\|\zeta\|_{\mathcal{H}}^2}\right).\end{aligned}$$

Here, A is a constant given in (E.29) and it does not depend on $n, \sigma, \|\zeta\|_{\mathcal{H}}$.

Proof. For any g_1, g_2 satisfying $\|g_1\| \leq \|\zeta\|_{\mathcal{H}}, \|g_2\| \leq \|\zeta\|_{\mathcal{H}}$, we have that for any $\nu \geq 1$,

$$|g_1(X_i)\phi_\nu(X_i) - g_2(X_i)\phi_\nu(X_i)| \leq \max_{x\in\Omega}|g_1(x) - g_2(x)|c_\phi,$$

where c_ϕ is defined in Assumption 2. Let

$$Z_n(g, \phi_\nu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[g(X_i)\phi_\nu(X_i) - \mathbb{E}\{g(X)\phi_\nu(X)\} \right].$$

Since $\Pi(\Omega) = 1$, by the Azuma-Hoeffding inequality, we have for any $t > 0$,

$$\begin{aligned} &\mathbb{P}(|Z_n(g_1, \phi_\nu) - Z_n(g_2, \phi_\nu)| \geq t) \\ &\quad \leq 2 \exp\left(-\frac{t^2}{8c_\phi^2 \max_{x\in\Omega} |g_1(x) - g_2(x)|^2} \right). \end{aligned} \tag{E.25}$$

In the following, we apply the maximal inequalities for empirical process (Kosorok (2008)). Recall that the Orlics norm $\|Z\|_{\psi_2}$ for any random variable Z is

$$\|Z\|_{\psi_2} \stackrel{\text{def}}{=} \infty_{c>0} \left\{ \mathbb{E}\psi_2(|Z|/c) \leq 1 \right\},$$

where $\psi_2(x) \stackrel{\text{def}}{=} \exp(x^2) - 1$. By (E.25) and Lemma 8.1 in Kosorok (2008), we obtain that

$$\| |Z_n(g_1, \phi_\nu) - Z_n(g_2, \phi_\nu)| \|_{\psi_2} \leq \sqrt{24} c_\phi \|g_1 - g_2\|_{L_\infty(\Omega)}. \tag{E.26}$$

Let $\tau = \sqrt{\log \frac{3}{2}}$ and $\psi(x) = \psi_2(\tau x)$. Then, $\psi(\cdot)$ is convex, nondecreasing with $\psi(0) = 0$ and $\psi(1) \leq \frac{1}{2}$. Moreover, since $\forall x, y \geq 1$, $\tau^{x^2}(\tau^{x^2(y^2-1)} + 1 - \tau^{y^2}) \geq \tau(\tau^{y^2-1} + 1 - \tau^{y^2}) > 2 - \tau^{y^2}$, we have $\psi(x)\psi(y) \leq \psi(xy)$. From the proof of Lemma 8.2 in Kosorok (2008), for any random variables $Z_1, \ldots, Z_k$,

$$\left\| \max_{1\leq i\leq k} Z_i \right\|_{\psi_2} \leq \frac{2}{\tau} \psi_2^{-1}(k) \max_{1\leq i\leq k} \|Z_i\|_{\psi_2}. \tag{E.27}$$

We define a ball $\mathcal{B}_{\|\zeta\|_\mathcal{H}}(\|\cdot\|) = \{g \in \mathcal{H} : \|g\| \leq \|\zeta\|_\mathcal{H}\}$. It is known the covering number of $\mathcal{B}_{\|\zeta\|_\mathcal{H}}(\|\cdot\|)$ denoted by

$$\mathcal{N}\left\{ \kappa, \mathcal{B}_{\|\zeta\|_\mathcal{H}}(\|\cdot\|), \|\cdot\|_{L_\infty(\Omega)} \right\}, \quad \text{for any } \kappa > 0,$$

has the following bound (Edmunds and Triebel, 1996):

$$\log \mathcal{N}\left\{\kappa, \mathcal{B}_{\|\zeta\|_{\mathcal{H}}}(\|\cdot\|), \|\cdot\|_{L_\infty}(\Omega)\right\} \leq c_0 \left(\frac{\|\zeta\|_{\mathcal{H}}}{\kappa}\right)^{d/m}. \tag{E.28}$$

Here, c_0 is independent of $\|\zeta\|_{\mathcal{H}}$ and κ. Note for any $g \in \mathcal{B}_{\|\zeta\|_{\mathcal{H}}}(\|\cdot\|)$ and $x \in \Omega$, $|g(x)| = |\langle g(\cdot), R(x,\cdot)\rangle| \leq \|g\|\sqrt{R(x,x)}$. Hence, $\|g\|_{L_\infty(\Omega)} = \max_{x\in\Omega}|g(x)| \leq \|\zeta\|_{\mathcal{H}} \cdot c_R$. By the general maximal inequality (see, Theorem 8.4 in Kosorok (2008)), (E.26), (E.27) and (E.28), we obtain that

$$\left\| \sup_{\substack{\nu\geq 1 \\ g_1,g_2\in\mathcal{B}_{\|\zeta\|_{\mathcal{H}}}(\|\cdot\|), \|g_1-g_2\|_{L_\infty(\Omega)}\leq\|\zeta\|_{\mathcal{H}}\cdot c_R}} |Z_n(g_1,\phi_\nu) - Z_n(g_2,\phi_\nu)| \right\|_{\psi_2} \leq A\|\zeta\|_{\mathcal{H}},$$

where

$$\begin{aligned} A &= A(c_\phi, c_R, d, m) \\ &= \frac{32\sqrt{6}c_\phi c_0^{m/d}}{\sqrt{\log 1.5}} \int_0^{c_0^{-m/d}c_R} \sqrt{\log[1+\exp(x^{-d/m})]}dx \\ &\quad + \frac{40\sqrt{6}c_\phi c_R}{\sqrt{\log 1.5}} \sqrt{\log[1+\exp(2^{1-d/m}c_0 c_R^{-d/m})]}. \end{aligned} \tag{E.29}$$

By letting $g_2 = 0$ in (E.29) and the Lemma 8.1 in Kosorok (2008), we complete the proof.

□

Bibliography

.

Aguilar, C., E. Westman, J. S. Muehlboeck, P. Mecocci, B. Vellas, M. Tsolaki, I. Kloszewska, and et al.

2013. Different multivariate techniques for automated classification of MRI data in Alzheimer's disease and mild cognitive impairment. *Psychiatry Research: Neuroimaging*, 212(2):89–98.

Antoniadis, A.

1997. Wavelets in statistics: a review. *Journal of the Italian Statistical Society*, 6:97–144.

Arnold, A., P. Markowich, G. Toscani, and A. Unterreiter

2001. On convex sobolev inequalities and the rate of convergence to equilibrium for fokker-planck type equations. *Communication in Partial Differential Equations*, 26(1–2):43–100.

Aronszajn, N.

1950. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Bayati, M. and A. Montanari

2012. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

Benjamini, Y. and Y. Hochberg

1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300.

Berglund, N.

2013. Kramers' law: Validity, derivations and generalisations. *Markov Processes And Related Fields*, 19:459–490.

Bickel, P.

1981. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *The Annals of Statistics*, 9(6):1301–1309.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov

2009. Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bogdan, M., E. van den Berg, C. Sabatti, W. Su, and E. J. Candès

2015. Slope-adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103–1140.

Boucheron, S., G. Lugosi, and P. Massart

2013. *Concentration Inequalities. A Nonasymptotic Theory of Independence*. Oxford, UK: Oxford University Press.

Bovier, A., M. Eckhoff, V. Gayrard, and M. Klein

2004. Metastability in reversible diffusion processes i: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424.

Bovier, A., V. Gayrard, and M. Klein

2005. Metastability in reversible diffusion processes ii: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7(1):69–99.

Breheny, P. and J. Huang

2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–253.

Bull, M. J. and N. T. Plummer

2014. Part 1: The human gut microbiome in health and disease. *Integrative Medicine: A Clinician's Journal*, 13(6):17–22.

Candès, E. J., J. K. Romberg, and T. Tao

2006. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223.

Candès, E. J. and T. Tao

2005. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215.

Cartan, H. P.

1971. *Differential Calculus*. Paris, France: Éditions Hermann.

Chaudhari, P., A. Oberman, S. Osher, S. Soatto, and G. Carlier

2018. Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):1–30.

Chaudhari, P. and S. Soatto

2018. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *2018 Information Theory and Applications Workshop (ITA), IEEE*, Pp. 1–10.

Chen, X., , B. E. Ankenman, and B. L. Nelson

2013. Enhancing stochastic kriging metamodels with gradient estimators. *Operations Research*, 61(2):512–528.

Chincarini, A., F. Sensi, L. Rei, G. Gemme, S. Squarcia, R. Longo, F. Brun, and et al.

2016. Integrating longitudinal information in hippocampal volume measurements for the early detection of alzheimer's disease. *NeuroImage*, 125:834–847.

Choromanska, A., M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun

2015. The loss surfaces of multilayer networks. *Artificial Intelligence and Statistics (AISTATS)*, Pp. 192–204.

Cover, T. M. and J. A. Thomas

2006. *Elements of Information Theory*. New York: John Wiley & Sons.

Cox, D. D.

1984. Multivariate smoothing spline functions. *SIAM Journal on Numerical Analysis*, 21(4):789–813.

Cox, D. D.

1988. Approximation of method of regularization estimators. *The Annals of Statistics*, 16(2):694–712.

Cox, D. D. and F. O'Sullivan

1990. Asymptotic analysis of penalized likelihood and related estimators. *The Annals of Statistics*, 18(4):1676–1695.

Craven, P. and G. Wahba

1978. Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403.

Dauphin, Y. N., R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio

2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in Neural Information Processing Systems (NIPS)*, Pp. 2933–2941.

Dinh, L., R. Pascanu, S. Bengio, and Y. Bengio

2017. Sharp minima can generalize for deep nets. *International Conference on Machine Learning (ICML)*, Pp. 1019–1028.

Donoho, D. L. and I. M. Johnstone

1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Donoho, D. L., R. C. Liu, and B. MacGibbon

1990. Minimax risk over hyperrectangles, and implications. *The Annals of Statistics*, 18(3):1416–1437.

Donoho, D. L. and J. Tanner

2010. Exponential bounds implying construction of compressed sensing matrices, error-correcting codes, and neighborly polytopes by random sampling. *IEEE Transactions on Information Theory*, 56(4):2002–2016.

Driscoll, M. F.

1973. The reproducing kernel hilbert space structure of the sample paths of a gaussian process. *Probability Theory and Related Fields*, 26(4):309–316.

Dziugaite, G. K. and D. M. Roy

2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*.

Edmunds, D. E. and H. Triebel

1996. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge, UK: Cambridge University Press.

Efron, B. and R. J. Tibshirani

1993. *An Introduction to the Bootstrap*. London, UK: Chapman and Hall.

Fan, J.

1997. Comments on wavelets in statistics: a review by A. Antoniadis. *Journal of the Italian Statistical Society*, 6(2):131–138.

Fan, J. and R. Li

2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Foster, D. P. and E. I. George

1994. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975.

Frees, E. W. and E. A. Valdez

1998. Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25.

Friedman, J., T. Hastie, and R. Tibshirani

2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Ge, R., F. Huang, C. Jin, and Y. Yuan

2015. Escaping from saddle points–online stochastic gradient for tensor decomposition. *Conference on Learning Theory (COLT)*, Pp. 797–842.

Geman, S. and D. Geman

1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Glasserman, P.

2013. *Monte Carlo Methods in Financial Engineering*. New York: Springer Science & Business Media.

Gramacy, R. B., D. Bingham, J. P. Holloway, M. J. Grosskopf, C. C. Kuranz, E. Rutter, M. Trantham, and P. R. Drake

2015. Calibrating a large computer experiment simulating radiative shock hydrodynamics. *The Annals of Applied Statistics*, 9(3):1141–1168.

Gu, C.

2013. *Smoothing Spline ANOVA Models*. New York: Springer Science & Business Media.

Hall, P., J. W. Kay, and D. M. Titterington

1990. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528.

Hall, P. and A. Yatchew

2007. Nonparametric estimation when data on derivatives are available. *The Annals of Statistics*, 35(1):300–323.

Hall, P. and A. Yatchew

2010. Nonparametric least squares estimation in derivative families. *Journal of Econometrics*, 157(2):362–374.

Hastie, T. and R. Tibshirani

1990. *Generalized Additive Models*. London, UK: Chapman & Hall/CRC.

Hastie, T. and R. Tibshirani

1993. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

Higdon, D., J. Gattiker, B. Williams, and M. Rightley

2008. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482):570–583.

Higdon, D., M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne

2004. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466.

Hoffer, E., I. Hubara, and D. Soudry

2017. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in Neural Information Processing Systems*, Pp. 1731–1741.

Huang, J. Z.

1998. Projection estimation in multiple regression with application to functional anova models. *The Annals of Statistics*, 26(1):242–272.

Jack Jr, C. R., M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, and et al.

2008. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691.

Jack Jr, C. R., D. S. Knopman, W. J. Jagust, R. C. Petersen, M. W. Weiner, P. S. Aisen, L. M. Shaw, and et al

2013. Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207–216.

Jack Jr, C. R., D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski

2010. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128.

Jack Jr, C. R., V. J. Lowe, S. D. Weigand, H. J. Wiste, M. L. Senjem, D. S. Knopman, and M. M. a. a. Shiung

2009. Serial pib and mri in normal, mild cognitive impairment and alzheimer's disease: implications for sequence of pathological events in alzheimer's disease. *Brain*, 132(5):1355–1365.

Jastrzebski, S., Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey

2017. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*.

Johnstone, I. M.

2017. Gaussian estimation: Sequence and wavelet models.

Jones, B. L. and J. A. Mereu

2002. A critique of fractional age assumptions. *Insurance: Mathematics and Economics*, 30(3):363–370.

Joseph, V. R. and S. N. Melkote

2009. Statistical adjustments to engineering models. *Journal of Quality Technology*, 41(4):362–375.

Joseph, V. R. and H. Yan

2015. Engineering-driven statistical adjustment and calibration. *Technometrics*, 57(2):257–267.

Kennedy, M. C. and A. O'Hagan

2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464.

Keskar, N. S., D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang

2016. On large-batch training for deep learning: Generalization gap and sharp minima. *International Conference on Learning Representations (ICLR)*.

Kimeldorf, G. and G. Wahba

1971. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95.

Kolpas, A., J. Moehlis, and I. G. Kevrekidis

2007. Coarse-grained analysis of stochasticity-induced switching between collective motion states. *Proceedings of the National Academy of Sciences*, 104(14):5931–5935.

Koltchinskii, V. and E. Giné

2000. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167.

Kosorok, M. R.

2008. *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Lan, H., M. Chen, J. B. Flowers, B. S. Yandell, D. S. Stapleton, C. M. Mata, E. T.-K. Mui, M. T. Flowers, K. L. Schueler, K. F. Manly, et al.

2006. Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*, 2(1):e6.

L'Ecuyer, P.

1990. A unified view of the ipa, sf, and lr gradient estimation techniques. *Management Science*, 36(11):1293–1416.

Li, K.-C.

1986. Asymptotic optimality of c_l and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14(3):1101–1112.

Li, Q., C. Tai, and W. E

2017. Stochastic modified equations and adaptive stochastic gradient algorithms. *International Conference on Machine Learning (ICML)*, 70:2101–2110.

Lin, Y.

1998. Tensor product space anova models in multivariate function estimation. *Thesis (Ph.D.)–University of Pennsylvania*.

Lin, Y.

2000. Tensor product space anova models. *The Annals of Statistics*, 28(3):734–755.

Liu, M., D. Zhang, and D. Shen

2016. Relationship induced multi-template learning for diagnosis of alzheimer's disease and mild cognitive impairment. *IEEE Transactions on Medical Imaging*, 35(6):1463–1474.

Mah, L., M. A. Binns, D. C. Steffens, and A. D. N. Initiative

2015. Anxiety symptoms in amnestic mild cognitive impairment are associated with

medial temporal atrophy and predict conversion to alzheimer disease. *The American Journal of Geriatric Psychiatry*, 23(5):466–476.

Mandt, S., M. D. Hoffman, and D. M. Blei

2017. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(1):4873–4907.

Martin, G. M., A. Bergman, and N. Barzilai

2007. Genetic determinants of human health span and life span: progress and new opportunities. *PLoS Genetics*, 3(7):e125.

Martins-Filho, C., S. Mishra, and A. Ullah

2008. A class of improved parametrically guided nonparametric regression estimators. *Econometric Reviews*, 27(4):542–573.

Mazumder, R., J. H. Friedman, and T. Hastie

2011. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138.

McKhann, G., D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan

1984. Clinical diagnosis of alzheimer's disease report of the nincds-adrda work group* under the auspices of department of health and human services task force on alzheimer's disease. *Neurology*, 34(7):939–939.

Meinshausen, N. and B. Yu

2009. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270.

Nadeau, J. H. and E. J. Topol

2006. The genetics of health. *Nature Genetics*, 38(10):1095.

Oakley, J. E. and A. O'Hagan

2004. Probabilistic sensitivity analysis of complex models: a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769.

Oden, J. T. and J. N. Reddy

2012. *An Introduction to the Mathematical Theory of Finite Elements*. New York: John Wiley & Sons.

Patrascu, A. and I. Necoara

2015. Random coordinate descent methods for ℓ_0 regularized convex optimization. *IEEE Transactions on Automatic Control*, 60(7):1811–1824.

Pavliotis, G. A.

2014. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Springer.

Plumlee, M.

2017. Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, 112(519):1274–1285.

Poggio, T., K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, and H. Mhaskar

2017. Theory of deep learning iii: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*.

Polya, G.

1945. Remarks on computing the probability integral in one and two dimensions. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Pp. 63–78.

Prince, M., R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri

2013. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's & Dementia*, 9(1):63–75.

Raginsky, M., A. Rakhlin, and M. Telgarsky

2017. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*.

Raskutti, G., M. J. Wainwright, and B. Yu

2011. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994.

Reuter, M., H. D. Rosas, and B. Fischl

2010. Highly accurate inverse consistent registration: a robust approach. *Neuroimage*, 53(4):1181–1196.

Reuter, M., N. J. Schmansky, H. D. Rosas, and B. Fischl

2012. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418.

Riesz, F. and B. Sz.-Nagy

1955. *Functional Analysis*. New York: Dover Publications.

Sabuncu, M. R., R. S. Desikan, J. Sepulcre, B. T. T. Yeo, H. Liu, N. J. Schmansky, M. Reuter, and et al.

2011. The dynamics of cortical and hippocampal atrophy in alzheimer disease. *Archives of Neurology*, 68(8):1040–1048.

Santner, T. J., B. J. Williams, and W. I. Notz

2003. *The Design and Analysis of Computer Experiments*. New York: Springer.

Schuff, N., D. Tosun, P. S. Insel, G. C. Chiang, D. Truran, P. S. Aisen, C. R. Jack Jr, M. W. Weiner, and A. D. N. Initiative

2012. Nonlinear time course of brain volume loss in cognitively normal and impaired elders. *Neurobiology of Aging*, 33(5):845–855.

Shreiner, A. B., J. Y. Kao, and V. B. Young

2015. The gut microbiome in health and in disease. *Current Opinion in Gastroenterology*, 31(1):69–75.

Song, Q. and F. Liang

2015. High-dimensional variable selection with reciprocal ℓ1-regularization. *Journal of the American Statistical Association*, 110(512):1607–1620.

Stone, C. J.

1980. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360.

Stone, C. J.

1982. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

Stone, C. J.

1985. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705.

Stone, C. J.

1994. The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *The Annals of Statistics*, 22(1):118–171.

Su, W., M. Bogdan, and E. Candès

2017. False discoveries occur early on the lasso path. *The Annals of Statistics*, 45(5):2133–2150.

Su, W. and E. Candès

2016. Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068.

Suri, R. and Y. T. Leung

1987. Single run optimization of a siman model for closed loop flexible assembly systems. *Proceedings of the 19th Conference on Winter Simulation*, Pp. 738–748.

Talagrand, M.

1996. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126(3):505–563.

Tibshirani, R.

1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R.

2011. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Toledo, J. B., X. Da, M. W. Weiner, D. A. Wolk, S. X. Xie, S. E. Arnold, C. Davatzikos, L. M. Shaw, J. Q. Trojanowski, and A. D. N. Initiative

2014. Csf apo-e levels associate with cognitive decline and mri changes. *Acta Neuropathologica*, 127(5):621–632.

Tosun, D., N. Schuff, L. M. Shaw, J. Q. Trojanowski, M. W. Weiner, and A. D. N. Initiative

2011. Relationship between csf biomarkers of alzheimer's disease and rates of regional cortical thinning in adni data. *Journal of Alzheimer's Disease*, 26(s3):77–90.

Tsybakov, A. B.

2009. *Introduction to Nonparametric Estimation*. New York: Springer.

Tuo, R. and C. F. J. Wu

2015. Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43(6):2331–2352.

Tuo, R. and C. F. J. Wu

2016. A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):767–795.

Tuo, R. and C. F. J. Wu

2018. Prediction based on the kennedy-o'hagan calibration model: asymptotic consistency and other properties. *Statistica Sinica*, 28:743–759.

Tzourio-Mazoyer, N., B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot

2002. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289.

Van de Geer, S. A. and P. Bühlmann

2009. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Varian, H. R.

1992. *Microeconomic Analysis*. New York: W. W. Norton & Company.

Vershynin, R.

2012. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications*, Pp. 210–268.

Wahba, G.

1990. *Spline Models for Observational Data*. Philadelphia, PA: SIAM.

Wahba, G., Y. Wang, C. Gu, R. Klein, and B. Klein

1995. Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. *The Annals of Statistics*, 23(6):1865–1895.

Wainwright, M. J.

2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

Weinberger, H. F.

1974. *Variational Methods for Eigenvalue Approximation*. Philadelphia, PA: SIAM.

White, H.

1982. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 50(1):1–25.

Wong, R. K. W., C. B. Storlie, and T. C. M. Lee

2017. A frequentist approach to computer model calibration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):635–648.

Wright, S. J.

2015. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34.

Wu, L. and Z. Zhu

2017. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*.

Yau, W.-Y. W., D. L. Tudorascu, E. M. McDade, S. Ikonomovic, J. A. James, D. Minhas, W. Mowrey, and et al.

2015. Longitudinal assessment of neuroimaging and clinical markers in autosomal dominant alzheimer's disease: a prospective cohort study. *The Lancet Neurology*, 14(8):804–813.

Ye, F. and C.-H. Zhang

2010. Rate minimaxity of the lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11(Dec):3519–3540.

Yuan, M. and Y. Lin

2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals

2017. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations (ICLR)*.

Zhang, C.-H.

2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, D., D. Shen, and A. D. N. Initiative

2012. Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. *PloS One*, 7(3):e33182.

Zou, H.

2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and R. Li

2008. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533.