



University of Pennsylvania  
**ScholarlyCommons**

---

Publicly Accessible Penn Dissertations


---

2022

## Estimation And Inference For Convex Functions And Computational Efficiency In High Dimensional Statistics

Ran Chen  
*University of Pennsylvania*

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Computer Sciences Commons](#), [Operational Research Commons](#), and the [Statistics and Probability Commons](#)

---

### Recommended Citation

Chen, Ran, "Estimation And Inference For Convex Functions And Computational Efficiency In High Dimensional Statistics" (2022). *Publicly Accessible Penn Dissertations*. 4765.  
<https://repository.upenn.edu/edissertations/4765>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/4765>  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

# Estimation And Inference For Convex Functions And Computational Efficiency In High Dimensional Statistics

## Abstract

Optimization and statistics are intrinsically intertwined with each other. Optimization has been the ends of some statistical problems, like estimation and inference for the minimizer and the minimum of convex functions, and the means for other statistical problems, like computational concerns in high dimensional statistics. In this dissertation, we consider both optimization-related problems. Estimation and inference for the minimizer and minimum of convex functions have been longstanding problems with wide application in economics and health care. But existing approaches are insufficient due to their asymptotic nature and/or incapability of characterizing function-specific difficulty. We investigate the problems under non-asymptotic frameworks that characterize function-specific difficulty and propose adaptive computational-efficient optimal methods. The first two parts of the dissertation address these problems, briefly summarized as follows. • The first part focuses on univariate convex functions. We develop computationally efficient adaptive optimal procedures under local minimax framework and discover a novel Uncertainty Principle that provides a fundamental limit on how well the minimizer and minimum can be estimated simultaneously for any convex regression function. • The second part focuses on multivariate additive convex functions. Under function-specific benchmarks, we propose computationally efficient optimal methods and establish their optimality.

Computational efficiency is another optimization-related problem of increasingly importance in statistics, especially in the AI age where the scale of data is big and the requirement on computational time is high. To achieve the balance between running time and statistical accuracy, we propose a framework that provides theoretically guaranteed optimization methods together with the analysis of interplay between running time and statistical accuracy for a class of high-dimensional problems in the third part of the dissertation. Our framework consists of three parts, statistical-optimization interplay analysis, which characterizes optimization induced statistical error in a more essential way, optimization template algorithm, and optimization convergence analysis. We showcase the power of our framework through three example problems, where we get novel results for the first two and show that our framework adapts to the degenerate case through the third example.

## Degree Type

Dissertation

## Degree Name

Doctor of Philosophy (PhD)

## Graduate Group

Statistics

## First Advisor

Tony T. Cai

## Keywords

Computational Efficiency, High-dimensional Statistics, Machine Learning, Nonparametric Statistics, Optimization, Panel Data Causal Inference

## Subject Categories

Computer Sciences | Operational Research | Statistics and Probability

This dissertation is available at ScholarlyCommons: <https://repository.upenn.edu/edissertations/4765>

ESTIMATION AND INFERENCE FOR CONVEX FUNCTIONS AND  
COMPUTATIONAL EFFICIENCY IN HIGH DIMENSIONAL STATISTICS

Ran Chen

A DISSERTATION

in

Statistics and Data Science

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

T. Tony Cai, Daniel H. Silberberg Professor, Professor of Statistics and Data Science

Graduate Group Chairperson

Nancy Zhang, Ge Li and Ning Zhao Professor, Professor of Statistics and Data Science

Dissertation Committee

T. Tony Cai, Daniel H. Silberberg Professor, Professor of Statistics and Data Science

Eric J. Tchetgen Tchetgen, Luddy Family Presidents Distinguished Professor, Professor of  
Statistics and Data Science

Zongming Ma, Associate Professor of Statistics and Data Science

ESTIMATION AND INFERENCE FOR CONVEX FUNCTIONS AND  
COMPUTATIONAL EFFICIENCY IN HIGH DIMENSIONAL STATISTICS

COPYRIGHT

2022

Ran Chen

This work is licensed under the  
Creative Commons  
NonCommercial-ShareAlike 4.0  
License

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/4.0/>

*Dedicated to my beloved mother and father*

## ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Tony Cai, for his patient guidance, mentorship and help throughout the past five years. What I have learnt from him is beyond words, and that will last.

I would like to thank Eric Tchetgen Tchetgen and Zongming Ma for kindly serving on my committee. I would also like to thank Linda Zhao, Mark Low, Nancy Zhang and Warren Ewens for their kind advices and help.

Finally, I would like to thank my parents for their unconditional love and support for the past 27 years. Everyone has a battle to fight in life, but they always prioritize mine and be there for me, and never ask anything in return. Words can not express my gratitude for this once-in-a-lifetime love, care and support, so I dedicate this milestone to them.

# ABSTRACT

## ESTIMATION AND INFERENCE FOR CONVEX FUNCTIONS AND COMPUTATIONAL EFFICIENCY IN HIGH DIMENSIONAL STATISTICS

Ran Chen

T. Tony Cai

Optimization and statistics are intrinsically intertwined with each other. Optimization has been the ends of some statistical problems, like estimation and inference for the minimizer and the minimum of convex functions, and the means for other statistical problems, like computational concerns in high dimensional statistics. In this dissertation, we consider both optimization-related problems.

Estimation and inference for the minimizer and minimum of convex functions have been longstanding problems with wide application in economics and health care. But existing approaches are insufficient due to their asymptotic nature and/or incapability of characterizing function-specific difficulty. We investigate the problems under non-asymptotic frameworks that characterize function-specific difficulty and propose adaptive computational-efficient optimal methods. The first two parts of the dissertation address these problems, briefly summarized as follows.

- The first part focuses on univariate convex functions. We develop computationally efficient adaptive optimal procedures under local minimax framework and discover a novel Uncertainty Principle that provides a fundamental limit on how well the minimizer and minimum can be estimated simultaneously for any convex regression function.
- The second part focuses on multivariate additive convex functions. Under function-specific benchmarks, we propose computationally efficient optimal methods and establish their optimality.

Computational efficiency is another optimization-related problem of increasingly importance in statistics, especially in the AI age where the scale of data is big and the requirement on computational time is high. To achieve the balance between running time and statistical accuracy, we propose a framework that provides theoretically guaranteed optimization methods together with the analysis of interplay between running time and statistical accuracy for a class of high-dimensional problems in the third part of the dissertation. Our framework consists of three parts, statistical-optimization interplay analysis, which characterizes optimization induced statistical error in a more essential way, optimization template algorithm, and optimization convergence analysis. We showcase the power of our framework through three example problems, where we get novel results for the first two and show that our framework adapts to the degenerate case through the third example.



# TABLE OF CONTENTS

ACKNOWLEDGEMENT . . . . .	iv
ABSTRACT . . . . .	v
LIST OF TABLES . . . . .	ix
LIST OF ILLUSTRATIONS . . . . .	x
CHAPTER 1 : Introduction . . . . .	1
CHAPTER 2 : Estimation and Inference for Minimizer and Minimum of Convex Functions: Optimality, Adaptivity, and Uncertainty Principles . . .	6
2.1 Introduction . . . . .	6
2.2 Benchmarks and Uncertainty Principle . . . . .	11
2.3 Adaptive Procedures and Optimality . . . . .	17
2.4 Nonparametric Regression . . . . .	25
2.5 Discussion . . . . .	37
2.6 Proofs . . . . .	39
CHAPTER 3 : Optimal Estimation and Inference for Minimizer and Minimum of Multivariate Additive Convex Functions . . . . .	47
3.1 Introduction . . . . .	47
3.2 Local Minimax Rates and Lower Bounds . . . . .	52
3.3 Projection Representation and Adaptive Optimal Procedures. . . . .	56
3.4 Nonparametric Regression . . . . .	65
CHAPTER 4 : Interplay Between Statistical Accuracy and Running Time Cost: A Framework and Examples . . . . .	78

4.1	Introduction . . . . .	78
4.2	General Framework . . . . .	89
4.3	Application to 1 Bit Matrix Completion . . . . .	98
4.4	Application to Causal Inference for Panel Data . . . . .	104
4.5	Application to Linear Regression (LASSO) . . . . .	113
4.6	Discussion . . . . .	121
APPENDIX . . . . .		124
A.1	Proofs of the Results in Chapter 2 . . . . .	124
A.2	Proofs of Supporting Technical Lemmas for Chapter 2 . . . . .	200
A.3	Comparison with CLS Methods and Connections with Classical Minimax Framework for Chapter 2 . . . . .	270
A.4	Simulation Results for Chapter 2 . . . . .	280
A.5	Proofs of the Results in Chapter 3 . . . . .	314
A.6	Proofs of the Results in Chapter 4 . . . . .	399
BIBLIOGRAPHY . . . . .		447

## LIST OF TABLES

TABLE A.1	List of the methods to be compared and their applicable scenario. .	282
-----------	---	-----

## LIST OF ILLUSTRATIONS

FIGURE 2.1	Water filling process. . . . .	13
FIGURE 2.2	Illustration of the localization step . . . . .	18
FIGURE 2.3	Illustration of the stopping rule . . . . .	18
FIGURE 4.1	Illustration of geometry of dual variable . . . . .	96
FIGURE A.1	Illustration of construction of $g_\delta$ , colored red in the plot . . . . .	143
FIGURE A.2	Illustration of upper bound proof . . . . .	145
FIGURE A.3	Plot of true functions . . . . .	283
FIGURE A.4	Plots for $f_1(x) = 100 2x - 1 $ . . . . .	288
FIGURE A.5	Tables for $f_1(x) = 100 2x - 1 $ . . . . .	289
FIGURE A.6	Plots for $f_2(x) = 100 2x - 1 \mathbb{1}\{x < 0.5\} + 100 2x - 1 ^2\mathbb{1}\{x \geq 0.5\}$ . . .	290
FIGURE A.7	Tables for $f_2(x) = 100 2x - 1 \mathbb{1}\{x < 0.5\} + 100 2x - 1 ^2\mathbb{1}\{x \geq 0.5\}$ . . .	291
FIGURE A.8	Plots for $f_3(x) = 100 2x - 1 \mathbb{1}\{x < 0.5\} + 100 \exp(2 - \frac{1}{ x-0.5 })\mathbb{1}\{x \geq 0.5\}$	292
FIGURE A.9	Tables for $f_3(x) = 100 2x - 1 \mathbb{1}\{x < 0.5\} + 100 \exp(2 - \frac{1}{ x-0.5 })\mathbb{1}\{x \geq 0.5\}$	293
FIGURE A.10	Plots for $f_4(x) = 100 10x - 1 ^2\mathbb{1}\{x < 0.1\} + 100 10 \cdot x/9 - 1/9 ^2\mathbb{1}\{x \geq 0.1\}$	294
FIGURE A.11	Tables for $f_4(x) = 100 10x - 1 ^2\mathbb{1}\{x < 0.1\} + 100 10 \cdot x/9 - 1/9 ^2\mathbb{1}\{x \geq 0.1\}$	295
FIGURE A.12	Plots for $f_5(x) = 100 10x - 1 ^4\mathbb{1}\{x < 0.1\} + 100 10 \cdot x/9 - 1/9 ^4\mathbb{1}\{x \geq 0.1\}$	296
FIGURE A.13	Tables for $f_5(x) = 100 10x - 1 ^4\mathbb{1}\{x < 0.1\} + 100 10 \cdot x/9 - 1/9 ^4\mathbb{1}\{x \geq 0.1\}$	297
FIGURE A.14	Plots for $f_6(x) = 100( 2x - 1 )^2$ . . . . .	298
FIGURE A.15	Tables for $f_6(x) = 100( 2x - 1 )^2$ . . . . .	299
FIGURE A.16	Plots for $f_7(x) = 100 2x - 1 ^2\mathbb{1}\{x < 0.5\} + 100 \exp(2 - \frac{1}{ x-0.5 })\mathbb{1}\{x \geq 0.5\}$	300
FIGURE A.17	Tables for $f_7(x) = 100 2x - 1 ^2\mathbb{1}\{x < 0.5\} + 100 \exp(2 - \frac{1}{ x-0.5 })\mathbb{1}\{x \geq 0.5\}$	301
FIGURE A.18	Plots for $f_8(x) = 100( 2x - 1 )^4$ . . . . .	302
FIGURE A.19	Tables for $f_8(x) = 100( 2x - 1 )^4$ . . . . .	303
FIGURE A.20	Plots for $f_9(x) = 100 \exp(2 - \frac{1}{ x-0.5 })$ . . . . .	304
FIGURE A.21	Tables for $f_9(x) = 100 \exp(2 - \frac{1}{ x-0.5 })$ . . . . .	305
FIGURE A.22	Plots for $f_{10}(x) = 100 \exp(2 - \frac{1}{ x-0.1 })\mathbb{1}\{x < 0.1\} + 100 10 \cdot x/9 - 1/9 ^2\mathbb{1}\{x \geq$ $0.1\}$ . . . . .	306
FIGURE A.23	Tables for $f_{10}(x) = 100 \exp(2 - \frac{1}{ x-0.1 })\mathbb{1}\{x < 0.1\} + 100 10 \cdot x/9 -$ $1/9 ^2\mathbb{1}\{x \geq 0.1\}$ . . . . .	307
FIGURE A.24	Empirical risks for minimizer . . . . .	310
FIGURE A.25	Empirical risks for minimum . . . . .	311
FIGURE A.26	Empirical lengths for minimizer . . . . .	312
FIGURE A.27	Empirical lengths for minimum . . . . .	313

# CHAPTER 1

## Introduction

Optimization and statistics has been increasingly intertwined in the AI age in the sense that optimization has been both a means and an ends for many statistical problems.

Estimation of and inference for the location and size of the extremum of a nonparametric regression function has been one of the longstanding problems in statistics with wide applications where optimization is the ends of the problem. See, for example, Kiefer and Wolfowitz (1952); Blum (1954); Chen (1988).

The problem has been investigated in different settings. For fixed design, upper bounds for estimating the minimum over various smoothness classes have been obtained (Muller, 1989; Facer and Müller, 2003; Shoung et al., 2001). Belitser et al. (2012) establishes the minimax rate of convergence over a given smoothness class for estimating both the minimizer and minimum. For sequential design, the minimax rate for estimation of the location has been established; see Chen et al. (1996); Polyak and Tsybakov (1990); Dippon (2003). Mokkadem and Pelletier (2007) introduces a companion for the Kiefer–Wolfowitz–Blum algorithm in sequential design for estimating both the minimizer and minimum.

Another related line of research is the stochastic continuum-armed bandits, which have been used to model online decision problems under uncertainty. Applications include online auctions, web advertising and adaptive routing. Stochastic continuum-armed bandits can be viewed as aiming to find the maximum of a nonparametric regression function through a sequence of actions. The objective is to minimize the expected total regret, which requires the trade-off between exploration of new information and exploitation of historical information. See, for example, Kleinberg (2004); Auer et al. (2007); Kleinberg et al. (2019).

The first two parts of this dissertation consider optimal estimation and confidence intervals for the minimizer and minimum of convex functions under both the white noise and

nonparametric regression models in a non-asymptotic minimax framework that evaluates the performance of any procedure at individual functions. We consider univariate convex functions in Chapter 2 and multivariate additive convex functions in Chapter 3.

In Chapter 2, we investigate the problem for univariate convex functions under the non-asymptotic local minimax framework that evaluates the performance of any procedure at individual functions, instead of the conventional minimax framework, which evaluates the performance of the estimators and confidence intervals in the worst case over a large collection of functions. Non-asymptotic local minimax framework enables a much more precise analysis than the conventional minimax framework, and brings out new phenomena in simultaneous estimation and inference for the minimizer and minimum. We establish a novel Uncertainty Principle that provides a fundamental limit on how well the minimizer and minimum can be estimated simultaneously for any convex regression function. A similar result holds for the expected length of the confidence intervals for the minimizer and minimum. Under this stricter framework, we propose fully adaptive computationally efficient optimal procedures and establish their optimality, i.e. we establish sharp minimax lower bounds (under local minimax framework) for the estimation accuracy and expected length of the confidence intervals for the minimizer and minimum and we establish matching statistical upper bounds for our procedures.

Chapter 2 is based on the joint work with T.Tony Cai and Yuancheng Zhu.

In Chapter 3, we focus on multivariate additive convex functions. We study estimation of the minimizer and both estimation and inference for the minimum under non-asymptotic local minimax framework. For the inference of the minimizer, we use a benchmark better characterizes the best performance any procedure can achieve at individual functions. We establish minimax lower bounds for the estimation accuracy and expected length(volume) for the minimizer and minimum, we propose computationally efficient optimal procedures, and we establish optimality by showing that the statistical upper bounds match the corresponding lower bounds up to a constant depending on the dimension and the pre-specified

probability coverage.

Optimization, in addition to being an ends for aforementioned statistical problem, is also very much involved in finding a desired statistical estimator. Many statistical estimators are formulated as an optimizer of a certain optimization problem, where an exact solution is hard or unable to compute. With the increasing scale of data nowadays, this computational cost issue becomes an increasingly important concern, especially for high-dimensional data.

Current efforts in investigating computational cost can be roughly categorized into three kinds. One is the computational-theoretical approach, where people investigate the computational cost by categorizing problems: people want to tell whether a problem is polynomial time computable. There are also attempts on categorizing problems in slightly different ways (Chandrasekaran and Jordan, 2013), i.e. adding additional consideration on statistical accuracy for categorizing. The second line is dealing with exploding sample size (Shender and Lafferty, 2013; Horev et al., 2015; Sussman et al., 2015; Kpotufe and Verma, 2017), mostly by reducing the effective sample size. The third line considers either or both statistical accuracy and optimization running time, but separately. The style of this line is having a statistically good estimator that is an optimizer of an optimization problem first, and then investigating theoretically guaranteed iterative optimization method for computing this estimator. This separation causes many problems. Conventional optimization convergence rate in terms of the distant to an optimizer is not the best way for characterizing the statistical behavior of the computed estimator, especially for over-parameterized cases where multiple solution is possible. The separation of the optimization problem from the statistical problem omits many statistically important considerations, e.g. the dependence of the convergence on dimension, the assumptions that statistical setting admits. The separation also places many statistically important optimization problem into an inferior position due to a different taste in optimization community.

Our approach is building a framework that provides theoretically guaranteed iterative optimization algorithm and precise quantification of how iteration number affects the statistical

accuracy for a class of problems that admits estimators of a certain general form without imposing artificial or hard-to-verify conditions.

Our framework consists of three parts solving two major problems, i.e. investigating statistical-optimization interplay and developing theoretically guaranteed optimization procedure, which leads to achievement of our goal. The first part incorporates the optimization induced error into the statistical analysis through an approximate optimization problem rather than an approximate solution. This is a more to-the-essence way of characterizing how optimization induced error affects statistical accuracy. This frees statistical analysis of the computed estimator from any optimization procedure, and makes it possible to get rid of hard-to-verify conditions facilitating optimization. The second part provides a template algorithm. The third part provides theoretical convergence analysis of our template algorithm in terms of converging to the approximate problem. In this part, our convergence analysis considers the dependence on both iteration number and statistically important quantities, e.g. dimension.

We apply our framework to three examples, 1-bit matrix completion (Davenport et al., 2014), causal inference for panel data (Athey et al., 2021) and (high dimensional sparse) linear regression with LASSO. In first two examples, we get interesting new results. For (high dimensional sparse) linear regression with LASSO, which is a degenerate case for our framework, we show that our framework adapts to degenerate case and gives stronger results when stronger assumptions are valid. For causal inference of panel data, we also sharpen the statistical analysis under the case that computational resource is unlimited, which is the case considered in the literature (Athey et al., 2021).

### **1.0.1. Notation**

Now we give a list of notation that we will be using through out the dissertation. We will remind the readers of relevant notation and additional notation for each chapter in each chapter again.



The cdf of the standard normal distribution is denoted by  $\Phi$ . For  $0 < \alpha < 1$ ,  $z_\alpha = \Phi^{-1}(1 - \alpha)$ . For  $\alpha = 0$ ,  $z_\alpha = \infty$ . For two real numbers  $a$  and  $b$ ,  $a \wedge b = \min\{a, b\}$ ,  $a \vee b = \max\{a, b\}$ . For  $f \in L_2[0, 1]$  and  $r > 0$ ,  $\mathcal{B}_r(f) = \{g \in L_2[0, 1] : \|g - f\|_2 \leq r\}$  and  $\partial\mathcal{B}_r(f) = \{g \in L_2[0, 1] : \|g - f\|_2 = r\}$ .

We use  $\|\cdot\|$  to denote the  $L_2$  norm for vectors, matrices (where  $L_2$  norm is Frobenius norm), real numbers (where  $L_2$  norm is absolute value), univariate functions and multivariate functions, depending on the setting. We use  $|\cdot|$  to denote the length of an interval, absolute value for a number, and cardinal for a discrete set. We use  $\mathbb{1}\{A\}$  to denote indicator function that takes 1 when event  $A$  happens and 0 otherwise. We use bold symbols to denote multivariate functions, e.g.  $\mathbf{f}$ ,  $\mathbf{g}$ ,  $\mathbf{h}$ .

We also use  $\|\cdot\|_F$  in addition to  $\|\cdot\|$  for Frobenius norm for matrices.  $\|\cdot\|_F$  is to give special emphasis for matrices when there might be confusion.  $\|\cdot\|_*$  stands for nuclear norm. We use  $D(A\|B) = \frac{1}{d_1 d_2} \sum_{i,j} D(A_{i,j}\|B_{i,j})$  to denote average KL divergence between  $d_1$  by  $d_2$  probability matrix  $A$  and  $B$  for 1-bit matrix completion, where  $D(a\|b) = a \log(\frac{a}{b}) + (1 - a) \log(\frac{1-a}{1-b})$ . We use  $\mathfrak{T}\{A\}$  to denote the function where it takes 0 if  $A$  holds and  $\infty$  if  $A$  does not hold. We use  $\mathcal{R}(\varepsilon, C)$  to denote the  $\varepsilon$  neighborhood of convex set  $C$ :  $\mathcal{R}(\varepsilon, C) = \{X : \inf_{Z \in C} \|X - Z\| \leq \varepsilon\}$ . We use  $B_d(x)$  to denote a ball in matrices space centered at  $x$  with radius  $d$  under Frobenius norm. We use  $\text{Proj}_C(P)$  to denote the projection point of  $P$  on convex set  $C$ , the projection is in terms of Euclidean distance.

## CHAPTER 2

### Estimation and Inference for Minimizer and Minimum of Convex Functions: Optimality, Adaptivity, and Uncertainty Principles

#### 2.1. Introduction

In this chapter, we focus on estimation and inference of the minimizer and minimum of univariate convex functions.

We first focus on the white noise model, which is given by

$$dY(t) = f(t)dt + \varepsilon dW(t), \quad 0 \leq t \leq 1,$$

where  $W(t)$  is a standard Brownian motion, and  $\varepsilon > 0$  is the noise level. The drift function  $f$  is assumed to be in  $\mathcal{F}$ , the collection of convex functions defined on  $[0, 1]$  with a unique minimizer  $Z(f) = \arg \min_{0 \leq t \leq 1} f(t)$ . The minimum value of the function  $f$  is denoted by  $M(f)$ , i.e.,  $M(f) = \min_{0 \leq t \leq 1} f(t) = f(Z(f))$ . The goal is to optimally estimate  $Z(f)$  and  $M(f)$ , as well as construct optimal confidence intervals for  $Z(f)$  and  $M(f)$ . Estimation and inference for the minimizer  $Z(f)$  and minimum  $M(f)$  under the nonparametric regression model will be discussed later in Section 2.4.

##### 2.1.1. Function-specific Benchmarks and Uncertainty Principle

As the first step toward evaluating the performance of a procedure at individual convex functions in  $\mathcal{F}$ , we define the function-specific benchmarks for estimation of the minimizer and minimum respectively by

$$R_z(\varepsilon; f) = \sup_{g \in \mathcal{F}} \inf_{\hat{Z}} \max_{h \in \{f, g\}} \mathbb{E}_h |\hat{Z} - Z(h)|, \quad (2.1.1)$$

$$R_m(\varepsilon; f) = \sup_{g \in \mathcal{F}} \inf_{\hat{M}} \max_{h \in \{f, g\}} \mathbb{E}_h |\hat{M} - M(h)|. \quad (2.1.2)$$

As in (2.1.1) and (2.1.2), we use subscript ‘ $z$ ’ to denote quantities related to the minimizer and ‘ $m$ ’ for the minimum throughout the paper. For any given  $f \in \mathcal{F}$ , the benchmarks  $R_z(\varepsilon; f)$  and  $R_m(\varepsilon; f)$  quantify the estimation accuracy at  $f$  of the minimizer  $Z(f)$  and minimum  $M(f)$  against the hardest alternative to  $f$  within the function class  $\mathcal{F}$ .

We show that  $R_z(\varepsilon; f)$  and  $R_m(\varepsilon; f)$  are the right benchmarks for capturing the estimation accuracy at individual functions in  $\mathcal{F}$  and will construct adaptive procedures that simultaneously perform within a constant factor of  $R_z(\varepsilon; f)$  and  $R_m(\varepsilon; f)$  for all  $f \in \mathcal{F}$ . In addition, it is also shown that any estimator  $\hat{Z}$  for the minimizer that is “super-efficient” at some  $f_0 \in \mathcal{F}$ , i.e., it significantly outperforms the benchmark  $R_z(\varepsilon; f_0)$ , must pay a penalty at another function  $f_1 \in \mathcal{F}$  and thus no procedure can uniformly outperform the benchmark. Same holds for the estimation of the minimum.

More interestingly, the non-asymptotic local minimax framework enables us to establish a novel Uncertainty Principle for estimating the minimizer and minimum of a convex function. The Uncertainty Principle reveals an intrinsic tension between the task of estimating the minimizer and that of estimating the minimum. That is, there is a fundamental limit to the estimation accuracy of the minimizer and minimum for all functions in  $\mathcal{F}$  and consequently the minimizer and minimum of a convex function cannot be estimated accurately at the same time. More specifically, it is shown that

$$\inf_{f \in \mathcal{F}} R_z(\varepsilon; f) \cdot R_m(\varepsilon; f)^2 \geq \frac{\Phi(-0.5)^3}{2} \varepsilon^2, \quad (2.1.3)$$

where  $\Phi(\cdot)$  is the cumulative distribution function (cdf) of the standard normal distribution. This is akin to the Heisenberg Uncertainty Principle in physics, which states that the velocity and the location of a particle can not be measured precisely at the same time. The connection will be discussed in more detail in Section 2.2.

For confidence intervals with a pre-specified coverage probability, the hardness of the problem is naturally characterized by the expected length. Let  $\mathcal{I}_{z,\alpha}(\mathcal{F})$  and  $\mathcal{I}_{m,\alpha}(\mathcal{F})$  be, respec-

tively, the collection of confidence intervals for the minimizer  $Z(f)$  and the minimum  $M(f)$  with guaranteed coverage probability  $1 - \alpha$  for all  $f \in \mathcal{F}$ . Let  $L(CI)$  be the length of a confidence interval  $CI$ . The minimum expected lengths at  $f$  of all confidence intervals in  $\mathcal{I}_{z,\alpha}(\{f, g\})$  and  $\mathcal{I}_{m,\alpha}(\{f, g\})$  with the hardest alternative  $g \in \mathcal{F}$  for  $f$  are given by

$$L_{z,\alpha}(\varepsilon; f) = \sup_{g \in \mathcal{F}} \inf_{CI \in \mathcal{I}_{z,\alpha}(\{f, g\})} \mathbb{E}_f L(CI), \quad (2.1.4)$$

$$L_{m,\alpha}(\varepsilon; f) = \sup_{g \in \mathcal{F}} \inf_{CI \in \mathcal{I}_{m,\alpha}(\{f, g\})} \mathbb{E}_f L(CI). \quad (2.1.5)$$

As in the case of estimation, we will first evaluate these benchmarks for the performance of confidence intervals in terms of the local moduli of continuity and then construct data-driven and computationally efficient confidence interval procedures. Furthermore, we also establish the Uncertainty Principle for the confidence intervals,

$$\inf_{f \in \mathcal{F}} L_{z,\alpha}(\varepsilon; f) \cdot L_{m,\alpha}(\varepsilon; f)^2 \geq C_\alpha \varepsilon^2. \quad (2.1.6)$$

where  $C_\alpha$  is a positive constant depending on  $\alpha$  only. The Uncertainty Principle shows a fundamental limit for the accuracy of simultaneous inference for the minimizer  $Z(f)$  and minimum  $M(f)$  for any  $f \in \mathcal{F}$ .

### 2.1.2. Adaptive Procedures

Another major step in our analysis is developing data-driven and computationally efficient algorithms for the construction of adaptive estimators and adaptive confidence intervals as well as establishing the optimality of these procedures at each  $f \in \mathcal{F}$ .

The key idea behind the construction of the adaptive procedures is to iteratively localize the minimizer by computing the integrals over the relevant subintervals together with a carefully constructed stopping rule. For estimation of the minimum and minimizer, additional estimation procedures are added after the localization steps. For the construction of the

confidence intervals, another important idea is to look back a few steps before the stopping time.

The resulting estimators,  $\hat{Z}$  for the minimizer  $Z(f)$  and  $\hat{M}$  for the minimum  $M(f)$ , are shown to attain within a constant factor of the benchmarks  $R_z(\varepsilon; f)$  and  $R_m(\varepsilon; f)$  simultaneously for all  $f \in \mathcal{F}$ ,

$$\mathbb{E}_f |\hat{Z} - Z(f)| \leq C_z R_z(\varepsilon; f) \quad \text{and} \quad \mathbb{E}_f |\hat{M} - M(f)| \leq C_m R_m(\varepsilon; f),$$

for some absolute constants  $C_z$  and  $C_m$  not depending on  $f$ . The confidence intervals,  $CI_{z,\alpha}$  for the minimizer  $Z(f)$  and  $CI_{m,\alpha}$  for the minimum  $M(f)$ , are constructed and shown to be adaptive to individual functions  $f \in \mathcal{F}$ , while having guaranteed coverage probability  $1 - \alpha$ . That is,  $CI_{z,\alpha} \in \mathcal{I}_{z,\alpha}(\mathcal{F})$  and  $CI_{m,\alpha} \in \mathcal{I}_{m,\alpha}(\mathcal{F})$  and for all  $f \in \mathcal{F}$ ,

$$\begin{aligned} \mathbb{E}_f L(CI_{z,\alpha}) &\leq C_z(\alpha) L_{z,\alpha}(\varepsilon; f) \\ \mathbb{E}_f L(CI_{m,\alpha}) &\leq C_m(\alpha) L_{m,\alpha}(\varepsilon; f), \end{aligned}$$

where  $C_z(\alpha)$  and  $C_m(\alpha)$  are constants depending on  $\alpha$  only.

### 2.1.3. Related Literature

In addition to estimation and inference for the location and size of the extremum of a nonparametric regression function mentioned at the beginning of this dissertation, the problems considered in this dissertation are also connected to nonparametric estimation and inference under shape constraints, which have also been well studied in the literature.

Nonparametric convex regression has been investigated in various settings, ranging from estimation and confidence bands for the whole function (Birge, 1989; Guntuboyina et al., 2018; Hengartner and Stark, 1995; Dumbgen, 1998), to estimation and inference at a fixed point (Kiefer, 1982; Cai et al., 2013; Cai and Low, 2015; Ghosal and Sen, 2017). Deng et al. (2020) established limiting distributions for some local parameters of a convex regression

function including the minimizer based on the convexity-constrained least squares (CLS) estimator and constructed a confidence interval for the minimizer. As seen in Section 2.4.4 and further discussions in the appendix Section A.3.1, this confidence interval is suboptimal in terms of the expected length. It is also much more computationally intensive as it requires solving the CLS problem.

The local minimax framework characterized by the benchmarks (2.1.1)-(2.1.2) and (2.1.4)-(2.1.5) was first developed in Cai et al. (2013) for estimation and Cai and Low (2015) for inference for the value of a convex function at a fixed point, which is a linear functional. The objects of interest in this dissertation, the minimizer and minimum, are nonlinear functionals. Due to the nonlinear nature of the minimizer and minimum, the analysis is much more challenging than for the function value at a fixed point.

Another related line of research is stochastic numerical optimization of convex functions. Agarwal et al. (2011) studies stochastic convex optimization with bandit feedback and proposes an algorithm that is shown to be nearly minimax optimal. Chatterjee et al. (2016) uses the framework introduced in Cai and Low (2015) to study the local minimax complexity of stochastic convex optimization based on queries to a first-order oracle that produces unbiased subgradient in a rather restrictive setting.

#### **2.1.4. Organization of this Chapter**

In Section 2.2, we analyze individual minimax risks, relating them to appropriate local moduli of continuity and more explicit alternative expression, and explain the uncertainty principle with a discussion of the connections with the classical minimax framework. Superefficiency is also considered. In Section 2.3, we introduce the adaptive procedures for the white noise model and show that they are optimal. In Section 2.4, we consider the nonparametric regression model. Adaptive procedures are proposed and their optimality is established. In addition, a summary of the numerical results is given. Section 2.5 discusses some future directions. Two main theorems are proved in Section 2.6. To avoid interrupting

the logic flow of main chapters, the proofs of other results are given in the Appendix Section A.1 and Section A.2.

### 2.1.5. Notation

We finish this section with some notation that will be used in this chapter. The cdf of the standard normal distribution is denoted by  $\Phi$ . For  $0 < \alpha < 1$ ,  $z_\alpha = \Phi^{-1}(1 - \alpha)$ . For two real numbers  $a$  and  $b$ ,  $a \wedge b = \min\{a, b\}$ ,  $a \vee b = \max\{a, b\}$ .  $\|\cdot\|_2$  denotes the  $L_2$  norm. For  $f \in L_2[0, 1]$  and  $r > 0$ ,  $\mathcal{B}_r(f) = \{g \in L_2[0, 1] : \|g - f\|_2 \leq r\}$  and  $\partial\mathcal{B}_r(f) = \{g \in L_2[0, 1] : \|g - f\|_2 = r\}$ .

## 2.2. Benchmarks and Uncertainty Principle

In this section, we first introduce the local moduli of continuity and use them to characterize the four benchmarks for estimation and confidence intervals introduced in Section 2.1.1, which are summarized in the following table:

	Estimation	Inference
Minimizer $Z(f)$	$R_z(\varepsilon; f)$	$L_{z,\alpha}(\varepsilon; f)$
Minimum $M(f)$	$R_m(\varepsilon; f)$	$L_{m,\alpha}(\varepsilon; f)$

We provide an alternative expression for the local moduli of continuity that are easier to evaluate. The results are used to establish a novel Uncertainty Principle, which shows an intrinsic tension between the estimation/inference accuracy for the minimizer and the minimum for all functions in  $\mathcal{F}$ .

### 2.2.1. Local Moduli of Continuity

For any given convex function  $f \in \mathcal{F}$ , we define the following local moduli of continuity, one for the minimizer, and the other for the minimum,

$$\omega_z(\varepsilon; f) = \sup \{ |Z(f) - Z(g)| : \|f - g\|_2 \leq \varepsilon, g \in \mathcal{F} \}, \quad (2.2.1)$$

$$\omega_m(\varepsilon; f) = \sup \{ |M(f) - M(g)| : \|f - g\|_2 \leq \varepsilon, g \in \mathcal{F} \}, \quad (2.2.2)$$

As in the case of a linear functional, the local moduli  $\omega_z(\varepsilon; f)$  and  $\omega_m(\varepsilon; f)$  clearly depend on the function  $f$  and can be regarded as an analogue of the inverse Fisher Information in regular parametric models.

The following theorem characterizes the four benchmarks for estimation and inference in terms of the corresponding local modulus of continuity.

**Theorem 2.2.1.** *Let  $0 < \alpha < 0.3$ . Then*

$$a_1 \omega_z(\varepsilon; f) \leq R_z(\varepsilon; f) \leq A_1 \omega_z(\varepsilon; f), \quad (2.2.3)$$

$$a_1 \omega_m(\varepsilon; f) \leq R_m(\varepsilon; f) \leq A_1 \omega_m(\varepsilon; f), \quad (2.2.4)$$

$$b_\alpha \omega_z(\varepsilon/3; f) \leq L_{z,\alpha}(\varepsilon; f) \leq B_\alpha \omega_z(\varepsilon; f), \quad (2.2.5)$$

$$b_\alpha \omega_m(\varepsilon/3; f) \leq L_{m,\alpha}(\varepsilon; f) \leq B_\alpha \omega_m(\varepsilon; f), \quad (2.2.6)$$

where the constants  $a_1, A_1, b_\alpha, B_\alpha$  can be taken as  $a_1 = \Phi(-0.5) \approx 0.309$ ,  $A_1 = 1.5$ ,  $b_\alpha = 0.6 - 2\alpha$ , and  $B_\alpha = 3(1 - 2\alpha)z_\alpha$ .

Theorem 2.2.1 shows that the four benchmarks can be characterized in terms of the local moduli of continuity. However, these local moduli of continuity are not easy to compute. We now introduce two geometric quantities to facilitate further understanding of these



benchmarks. For  $f \in \mathcal{F}$ ,  $u \in \mathbb{R}$  and  $\varepsilon > 0$ , let  $f_u(t) = \max\{f(t), u\}$  and define

$$\rho_m(\varepsilon; f) = \sup\{u - M(f) : \|f - f_u\|_2 \leq \varepsilon\}, \quad (2.2.7)$$

$$\rho_z(\varepsilon; f) = \sup\{|t - Z(f)| : f(t) \leq \rho_m(\varepsilon; f) + M(f), t \in [0, 1]\}. \quad (2.2.8)$$

Obtaining  $\rho_m(\varepsilon; f)$  and  $\rho_z(\varepsilon; f)$  can be viewed as a *water-filling process*. One adds water into the epigraph defined by the convex function  $f$  until the “volume” (measured by  $\|\cdot\|_2$ ) is equal to  $\varepsilon$ . As illustrated in Figure 2.1,  $\rho_m(\varepsilon; f)$  measures the depth of the water (CD), and  $\rho_z(\varepsilon; f)$  captures the width of the water surface (FC).  $\rho_m(\varepsilon; f)$  and  $\rho_z(\varepsilon; f)$  essentially quantify the flatness of the function  $f$  near its minimizer  $Z(f)$ .

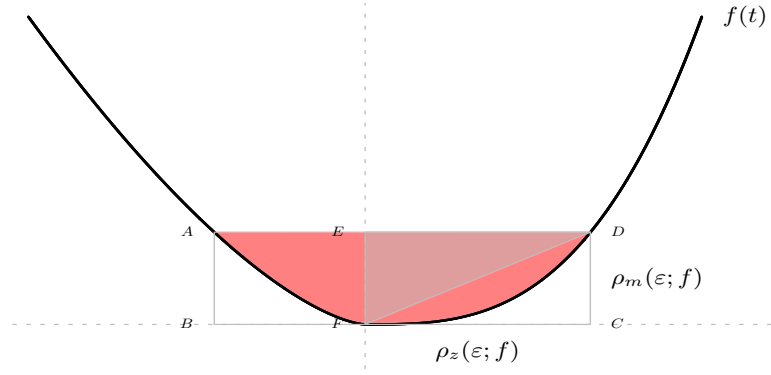


Figure 2.1: Water filling process.

The geometric quantities  $\rho_m(\varepsilon; f)$  and  $\rho_z(\varepsilon; f)$  defined in (2.2.7) and (2.2.8) have the following properties.

**Proposition 2.2.1.** For  $0 < c < 1$ ,  $f \in \mathcal{F}$ ,

$$c \leq \frac{\rho_m(c\varepsilon; f)}{\rho_m(\varepsilon; f)} \leq c^{\frac{2}{3}} \quad \text{and} \quad \max\left\{\left(\frac{c}{2}\right)^{\frac{2}{3}}, c\right\} \leq \frac{\rho_z(c\varepsilon; f)}{\rho_z(\varepsilon; f)} \leq 1. \quad (2.2.9)$$

The following result connects the local moduli of continuity to these two geometric quantities.

**Proposition 2.2.2.** Let  $\rho_m(\varepsilon; f)$  and  $\rho_z(\varepsilon; f)$  be defined in (2.2.7) and (2.2.8), respectively.

Then

$$\rho_m(\varepsilon; f) \leq \omega_m(\varepsilon; f) \leq 3\rho_m(\varepsilon; f), \quad (2.2.10)$$

$$\rho_z(\varepsilon; f) \leq \omega_z(\varepsilon; f) \leq 3\rho_z(\varepsilon; f). \quad (2.2.11)$$

Therefore, through the local moduli of continuity, the hardness of the estimation and inference tasks are tied to the geometry of the convex function near its minimizer. Note that as the function gets flatter near its minimizer,  $\rho_m(\varepsilon; f)$  decreases while  $\rho_z(\varepsilon; f)$  increases. It is useful to calculate  $\rho_m(\varepsilon; f)$  and  $\rho_z(\varepsilon; f)$  in a concrete example.

*Example 2.2.1.* Consider the function  $f(t) = |t - \frac{1}{2}|^k$  where  $k \geq 1$  is a constant. We will calculate  $\rho_m(\varepsilon; f)$  and then obtain  $\rho_z(\varepsilon; f)$  by first computing  $\|f - f_u\|_2^2$  and then setting it to  $\varepsilon^2$  to solve for  $\rho_m(\varepsilon; f)$ .

It is easy to see that in this case  $\|f - f_u\|_2^2 = \frac{4k^2}{(2k+1)(k+1)} \cdot u^{\frac{2k+1}{k}}$ . Setting  $\|f - f_u\|_2^2 = \varepsilon^2$  yields  $u = \left(\frac{(2k+1)(k+1)}{4k^2}\right)^{\frac{k}{2k+1}} \varepsilon^{\frac{2k}{2k+1}}$ . Hence,

$$\rho_m(\varepsilon; f) = \left(\frac{(2k+1)(k+1)}{4k^2}\right)^{\frac{k}{2k+1}} \varepsilon^{\frac{2k}{2k+1}}.$$

To compute  $\rho_z(\varepsilon; f)$ , note that  $f^{-1}(u) = \frac{1}{2} \pm u^{\frac{1}{k}} = \frac{1}{2} \pm \left(\frac{(2k+1)(k+1)}{4k^2}\right)^{\frac{1}{2k+1}} \varepsilon^{\frac{2}{2k+1}}$ . Hence

$$\rho_z(\varepsilon; f) = \min \left\{ \left(\frac{(2k+1)(k+1)}{4k^2}\right)^{\frac{1}{2k+1}} \varepsilon^{\frac{2}{2k+1}}, \frac{1}{2} \right\}.$$

Proposition 2.2.2 then yields tight bounds for the local moduli of continuity  $\omega_m(\varepsilon; f)$  and  $\omega_z(\varepsilon; f)$ .

*Remark 2.2.1.* Note that the results obtained in Example 2.2.1 can be extended to a class of convex functions. For  $f \in \mathcal{F}$  satisfying

$$0 < \liminf_{t \rightarrow Z(f)} \frac{f(t) - M(f)}{|t - Z(f)|^k} \leq \limsup_{t \rightarrow Z(f)} \frac{f(t) - M(f)}{|t - Z(f)|^k} < \infty$$

for some  $k \geq 1$ , it is easy to show that

$$\omega_m(\varepsilon; f) \sim \varepsilon^{\frac{2k}{2k+1}}, \quad \omega_z(\varepsilon; f) \sim \varepsilon^{\frac{2}{2k+1}}, \quad \text{as } \varepsilon \rightarrow 0^+.$$

### 2.2.2. Uncertainty Principle

Section 2.2.1 provides a precise characterization of the four benchmarks under the non-asymptotic local minimax framework in terms of the local moduli of continuity and the geometric quantities  $\rho_m(\varepsilon; f)$  and  $\rho_z(\varepsilon; f)$ . These results yield a novel Uncertainty Principle.

**Theorem 2.2.2** (Uncertainty Principle). *Let  $R_z(\varepsilon; f)$ ,  $R_m(\varepsilon; f)$ ,  $L_{z,\alpha}(\varepsilon; f)$ , and  $L_{m,\alpha}(\varepsilon; f)$  be defined as in (2.1.1)–(2.1.5). Let  $0 < \alpha < 0.3$ . Then for any  $f \in \mathcal{F}$ ,*

$$274\varepsilon^2 > R_z(\varepsilon; f) \cdot R_m(\varepsilon; f)^2 \geq \frac{\Phi(-0.5)^3}{2}\varepsilon^2, \quad (2.2.12)$$

$$3^7 \cdot (1 - 2\alpha)^3 \varepsilon^2 > L_{z,\alpha}(\varepsilon; f) \cdot L_{m,\alpha}(\varepsilon; f)^2 \geq \frac{(0.6 - 2\alpha)^3}{18}\varepsilon^2. \quad (2.2.13)$$

Note that the bounds in (2.2.12) and (2.2.13) are universal for all  $f \in \mathcal{F}$  and show that there is a fundamental limit to the accuracy of estimation and inference for the minimizer and minimum of a convex function. The Uncertainty Principle in Theorem 2.2.2 is akin to the well-known Heisenberg Uncertainty Principle in physics, which states that a particle's location and velocity cannot be determined precisely at the same time. The underlying reason for the Heisenberg Uncertainty Principle is that the momentum operator for the velocity and displacement operator for the location are non-commutative. More precisely, the degree of uncertainty depends on the extent these two operators are related through the Lie bracket, which can be viewed as a measure of non-commutativity. For details on the Heisenberg Uncertainty Principle; see, for example, Griffiths and Schroeter (2018).

Our finding here states that the minimizer and the minimum of a convex function cannot be estimated accurately at the same time. This statistical uncertainty principle comes from an intrinsic relationship between the two operators  $Z(\cdot)$  and  $M(\cdot)$ : For any convex function

$f \in \mathcal{F}$  and any  $r > 0$ , there exists  $g \in \partial\mathcal{B}_r(f) \cap \mathcal{F}$  such that

$$|Z(g) - Z(f)| \cdot |M(g) - M(f)|^2 \geq \frac{1}{2} \left(\frac{r}{\varepsilon}\right)^2 \cdot \varepsilon^2, \quad (2.2.14)$$

where  $r/\varepsilon = \|(f - g)/\varepsilon\|_2$  characterizes the probabilistic distance between the two convex functions  $f$  and  $g$  under the white noise model. The  $L_2$  norm of the difference plays a similar role to the Lie bracket in the Heisenberg Uncertainty Principle. In both settings, there is a quantity determining the “entanglement” of two functionals/operators. The difference is that the “entanglement” for quantum physics is extracted and viewed in quantum sense while ours is extracted and viewed in probability sense.

*Remark 2.2.2.* To the best of our knowledge, the uncertainty principles established in this paper are the first of their kind in nonparametric statistics in that they reveal the fundamental tensions between estimation/inference of different quantities. It is shown in the appendix Section A.3.3 that similar uncertainty principles also hold for certain subclasses of the convex functions. Note that it is not possible to establish such results using the conventional minimax analysis where the performance is measured in the worst case over a large parameter space.

### 2.2.3. Penalty for Super-efficiency

We have shown that the estimation benchmarks  $R_z(\varepsilon; f)$  and  $R_m(\varepsilon; f)$  defined in (2.1.1) and (2.1.2) can be characterized by the local moduli of continuity. Before we show in Section 2.3 that these benchmarks are indeed achievable by adaptive procedures, we first prove that they cannot be essentially outperformed by any estimator uniformly over  $\mathcal{F}$ . The benchmarks  $R_z(\varepsilon; f)$  and  $R_m(\varepsilon; f)$  play a role analogous to the information lower bound in the classical statistics.

**Theorem 2.2.3** (Penalty for super-efficiency). *For any estimator  $\hat{Z}$ , if  $\mathbb{E}_{f_0}|\hat{Z} - Z(f_0)| \leq$*

$\gamma R_z(\varepsilon; f_0)$  for some  $f_0 \in \mathcal{F}$  and  $\gamma < 0.1$ , then there exists  $f_1 \in \mathcal{F}$  such that

$$\mathbb{E}_{f_1}(|\hat{Z} - Z(f_1)|) \geq \frac{1}{40} \left( \log \frac{1}{\gamma} \right)^{1/3} R_z(\varepsilon; f_1). \quad (2.2.15)$$

Similarly, for any estimator  $\hat{M}$ , if  $\mathbb{E}_{f_0}|\hat{M} - M(f_0)| \leq \gamma R_m(\varepsilon; f_0)$  for some  $f_0 \in \mathcal{F}$  and  $\gamma < 0.1$ , then there exists  $f_1 \in \mathcal{F}$  such that

$$\mathbb{E}_{f_1}|\hat{M} - M(f_1)| \geq \frac{1}{8} \left( \log \frac{1}{\gamma} \right)^{1/3} R_m(\varepsilon; f_1). \quad (2.2.16)$$

*Remark 2.2.3.* Theorem 2.2.3 shows that if an estimator of  $Z(f)$  or  $M(f)$  is super-efficient at some  $f_0 \in \mathcal{F}$  in the sense of outperforming the benchmark by a factor of  $\gamma$  for some small  $\gamma > 0$ , then it must be sub-efficient at some  $f_1 \in \mathcal{F}$  by underperforming the benchmark by at least a factor of  $\left( \log \frac{1}{\gamma} \right)^{\frac{1}{3}}$ .

## 2.3. Adaptive Procedures and Optimality

We now turn to the construction of data-driven and computationally efficient algorithms for estimation and confidence intervals for the minimizer  $Z(f)$  and minimum  $M(f)$  under the white noise model. The procedures are shown to be adaptive to each individual function  $f \in \mathcal{F}$  in the sense that they simultaneously achieve, up to a universal constant, the corresponding benchmarks  $R_z(\varepsilon; f)$ ,  $R_m(\varepsilon; f)$ ,  $L_{z,\alpha}(\varepsilon; f)$ , and  $L_{m,\alpha}(\varepsilon; f)$  for all  $f \in \mathcal{F}$ . These results are much stronger than what can be obtained from a conventional minimax analysis.

### 2.3.1. The Construction

There are three main building blocks in the construction of the estimators and confidence intervals: Localization, stopping, and estimation/inference.

In the localization step, we begin with the initial interval  $[0, 1]$ . Then, iteratively, we halve

the intervals and select one halved interval. The candidate-halved-intervals for selection are the two resulting sub-intervals of the previously selected interval and one neighboring halved interval, when such an interval exists, on each side. The selection rule is to choose the one with the smallest integral of the white noise process over it. See Figure 2.2 for an illustration of the localization step.

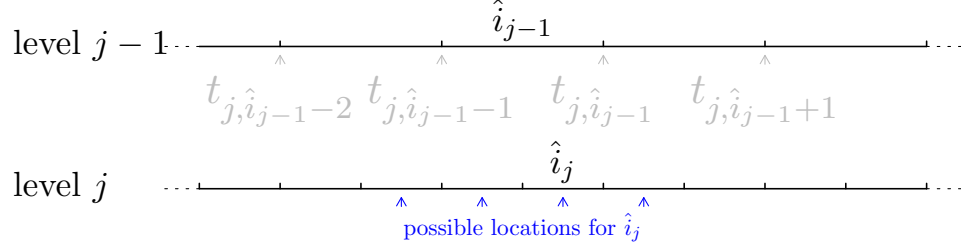


Figure 2.2: Illustration of the localization step. At level  $j$ , the middle two intervals are the two subintervals of the selected interval at level  $j - 1$ . One adjacent interval of the same length on each side is added and the interval at level  $j$  is selected among these four intervals.

The second step of the construction is the stopping rule. The localization step is iterative, so one needs to determine when there is no further gain and stop the iteration. The integral over each selected interval is a random variable and can be viewed as an estimate of the minimum times the length of the interval. The intuition is that, as the iteration progresses, the bias decreases and the variance increases. As shown in Figure 2.3, the basic idea is to use the differences of the integrals over the two neighboring intervals 5 blocks away from the current designated interval, when such intervals exist, on both sides. If either of the differences is smaller than 2 standard deviations, then the iteration stops.



Figure 2.3: Illustration of the stopping rule.

After selecting the final subinterval, the last step in the construction is the estimation/inference for both the minimum and minimizer, which will be described separately later. The detailed construction is given as follows.

## Sample Splitting

For technical reasons, we split the data into three independent pieces to ensure independence of the data used in the three steps of the construction. This is done as follows.

Let  $B_1(t)$  and  $B_2(t)$  be two independent standard Brownian motions, and both be independent of the observed data  $Y$ . Let

$$\begin{aligned} Y_l(t) &= Y(t) + \frac{\sqrt{2}}{2}\varepsilon B_1(t) + \frac{\sqrt{6}}{2}\varepsilon B_2(t), \\ Y_s(t) &= Y(t) + \frac{\sqrt{2}}{2}\varepsilon B_1(t) - \frac{\sqrt{6}}{2}\varepsilon B_2(t), \\ Y_e(t) &= Y(t) - \sqrt{2}\varepsilon B_1(t). \end{aligned} \tag{2.3.1}$$

Then  $Y_l(\cdot)$ ,  $Y_s(\cdot)$  and  $Y_e(\cdot)$  are independent and can be written as

$$\begin{aligned} dY_l(t) &= f(t)dt + \sqrt{3}\varepsilon dW_1(t), \\ dY_s(t) &= f(t)dt + \sqrt{3}\varepsilon dW_2(t), \\ dY_e(t) &= f(t)dt + \sqrt{3}\varepsilon dW_3(t), \end{aligned} \tag{2.3.2}$$

where  $W_1$ ,  $W_2$  and  $W_3$  are independent standard Brownian motions.

We now have three independent copies:  $Y_l$  is used for localization,  $Y_s$  for stopping, and  $Y_e$  for the construction of the final estimator and confidence interval for the minimum.

*Remark 2.3.1.* If one is only interested in estimation and inference for the minimizer, the copy  $Y_e$  is not needed, and it suffices to split into two independent copies with smaller variance and thus leads to slightly better performance. Another point is that, although here the three processes  $Y_l$ ,  $Y_s$ , and  $Y_e$  are made to have the same noise level, it is not necessary for the noise levels to be the same. For the simplicity and ease of presentation, we split the original sample into three independent and homoskedastic copies for estimation and inference for both the minimizer and minimum.

## Localization

For  $j = 0, 1, \dots$ , and  $i = 0, 1, \dots, 2^j$ , let

$$m_j = 2^{-j}, \quad t_{j,i} = i \cdot m_j, \quad \text{and} \quad i_j^* = \max\{i : Z(f) \in [t_{j,i-1}, t_{j,i}]\}. \quad (2.3.3)$$

That is, at level  $j$  for  $j = 0, 1, \dots$ , the  $i_j^*$ -th subinterval is the one containing the minimizer  $Z(f)$ . For  $j = 0, 1, \dots$ , and  $i = 1, 2, \dots, 2^j$ , define

$$X_{j,i} = \int_{t_{j,i-1}}^{t_{j,i}} dY_l(t),$$

where  $Y_l$  is one of the three independent copies constructed above through sample splitting.

For convenience, we define  $X_{j,i} = +\infty$  for  $j = 0, 1, \dots$ , and  $i \in \mathbb{Z} \setminus \{1, 2, \dots, 2^j\}$ .

Let  $\hat{i}_0 = 1$  and for  $j = 1, 2, \dots$ , let

$$\hat{i}_j = \arg \min_{2\hat{i}_{j-1}-2 \leq i \leq 2\hat{i}_{j-1}+1} X_{j,i}.$$

Note that given the value of  $\hat{i}_{j-1}$  at level  $j-1$ , in the next iteration the procedure halves the interval  $[t_{\hat{i}_{j-1}-1}, t_{\hat{i}_{j-1}}]$  into two subintervals and selects the interval  $[t_{\hat{i}_j-1}, t_{\hat{i}_j}]$  at level  $j$  from these and their immediate neighboring subintervals. So  $i$  only ranges over 4 possible values at level  $j$ . See Figure 2.2 for an illustration.

## Stopping Rule

It is necessary to have a stopping rule to select a final subinterval constructed in the localization iterations. We use another independent copy  $Y_s$  constructed in the sample splitting step to devise a stopping rule. For  $j = 0, 1, \dots$ , and  $i = 1, 2, \dots, 2^j$ , let

$$\tilde{X}_{j,i} = \int_{t_{j,i-1}}^{t_{j,i}} dY_s(t).$$



Again, for convenience, we define  $\tilde{X}_{j,i} = +\infty$  for  $j = 0, 1, \dots$ , and  $i \in \mathbb{Z} \setminus \{1, 2, \dots, 2^j\}$ . Let the statistic  $T_j$  be defined as

$$T_j = \min\{\tilde{X}_{j,\hat{i}_j+6} - \tilde{X}_{j,\hat{i}_j+5}, \tilde{X}_{j,\hat{i}_j-6} - \tilde{X}_{j,\hat{i}_j-5}\},$$

where we use the convention  $+\infty - x = +\infty$  and  $\min\{+\infty, x\} = x$ , for any  $-\infty \leq x \leq \infty$ .

The stopping rule is based on the value of  $T_j$ . It is helpful to provide some intuition before formally defining the stopping rule. Intuitively, the algorithm should stop at a place where the signal to noise ratio of  $T_j$  is small or where the signal is negative. Let  $\sigma_j^2 = 6m_j\varepsilon^2$ . It is easy to see that, when  $\tilde{X}_{j,\hat{i}_j+6} - \tilde{X}_{j,\hat{i}_j+5} < \infty$ ,

$$\tilde{X}_{j,\hat{i}_j+6} - \tilde{X}_{j,\hat{i}_j+5} | \hat{i}_j \sim N \left( \int_{t_{j,\hat{i}_j+5}}^{t_{j,\hat{i}_j+6}} (f(t+m_j) - f(t)) dt, \sigma_j^2 \right). \quad (2.3.4)$$

Note that the standard deviation  $\sigma_j$  decreases at the rate  $\frac{1}{\sqrt{2}}$  as  $j$  increases. We now turn to the mean of  $\tilde{X}_{j,\hat{i}_j+6} - \tilde{X}_{j,\hat{i}_j+5} | \hat{i}_j$ . Recall the notation introduced in (2.3.3). It is easy to see that the algorithm should stop as soon as  $\int_{t_{j,\hat{i}_j+5}}^{t_{j,\hat{i}_j+6}} (f(t+m_j) - f(t)) dt$  turns negative, since for any  $\hat{i}_j$ , if  $\int_{t_{j,\hat{i}_j+5}}^{t_{j,\hat{i}_j+6}} (f(t+m_j) - f(t)) dt < 0$ , then  $|\hat{i}_j - i_j^*| \geq 5$  and consequently  $|\hat{i}_{j_1} - i_{j_1}^*| \geq 5$  for any  $j_1 \geq j$ . When  $\int_{t_{j,\hat{i}_j+5}}^{t_{j,\hat{i}_j+6}} (f(t+m_j) - f(t)) dt$  is positive, a careful analysis in the proof shows that it shrinks at a rate faster than or equal to  $\frac{1}{4}$  as  $j$  increases. Analogous results hold for  $\tilde{X}_{j,\hat{i}_j-6} - \tilde{X}_{j,\hat{i}_j-5} | \hat{i}_j$ .

Finally, the iterations stop at level  $\hat{j}$  where

$$\hat{j} = \min\{j : \frac{T_j}{\sigma_j} \leq 2\}.$$

The subinterval containing the minimizer  $Z(f)$  is localized to be  $[t_{\hat{j},\hat{i}_{\hat{j}}-1}, t_{\hat{j},\hat{i}_{\hat{j}}}]$ .

## Estimation and Inference

After the final subinterval  $[t_{\hat{j}, \hat{i}_{\hat{j}}-1}, t_{\hat{j}, \hat{i}_{\hat{j}}}]$  is obtained, we then use it to construct estimators and confidence intervals for  $Z(f)$  and  $M(f)$ . We begin with the minimizer  $Z(f)$ . The estimator of  $Z(f)$  is given by the midpoint of the interval  $[t_{\hat{j}, \hat{i}_{\hat{j}}-1}, t_{\hat{j}, \hat{i}_{\hat{j}}}]$ , i.e.,

$$\hat{Z} = \frac{t_{\hat{j}, \hat{i}_{\hat{j}}} + t_{\hat{j}, \hat{i}_{\hat{j}}-1}}{2}. \quad (2.3.5)$$

To construct the confidence interval for  $Z(f)$ , one needs to take a few steps to the left and to the right at level  $\hat{j}$ . Let  $K_\alpha = \lceil \frac{\log \alpha}{\log \Phi(-2)} \rceil$  and define

$$L = \max\{0, \hat{i}_{\hat{j}} - 12 \times 2^{K_\alpha} + 1\}, \quad U = \min\{2^{\hat{j}}, \hat{i}_{\hat{j}} + 12 \times 2^{K_\alpha} - 2\}.$$

The  $1 - \alpha$  confidence interval for  $Z(f)$  is given by

$$CI_{Z, \alpha} = [t_{\hat{j}, L}, t_{\hat{j}, U}]. \quad (2.3.6)$$

For estimation of and confidence interval for the minimum  $M(f)$ , define

$$\bar{X}_{j, i} = \int_{t_{j, i-1}}^{t_{j, i}} Y_e(t) dt.$$

Let  $\tilde{i}_{\hat{j}} = \hat{i}_{\hat{j}} + 2 \left( \mathbb{1}\{\tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}+6} - \tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}+5} \leq 2\sigma_j\} - \mathbb{1}\{\tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}-6} - \tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}-5} \leq 2\sigma_j\} \right)$  and define the final estimator of the minimum  $M(f)$  by

$$\hat{M} = \frac{1}{m_{\hat{j}}} \bar{X}_{\hat{j}, \tilde{i}_{\hat{j}}}. \quad (2.3.7)$$

We now turn to the inference for  $M(f)$ . Recall that  $K_\alpha = \lceil \frac{\log \alpha}{\log \Phi(-2)} \rceil$ . Compared with the confidence interval for the minimizer, we take four more blocks on each side at the level

$(\hat{j} - K_{\frac{\alpha}{4}} - 1)_+$ . More specifically, we define

$$t_L = t_{(\hat{j} - K_{\frac{\alpha}{4}} - 1)_+, \hat{i}_{(\hat{j} - K_{\frac{\alpha}{4}} - 1)_+} - 5}, \quad t_R = t_{(\hat{j} - K_{\frac{\alpha}{4}} - 1)_+, \hat{i}_{(\hat{j} - K_{\frac{\alpha}{4}} - 1)_+} + 4}.$$

Set

$$\tilde{K}_\alpha = \max\{4, 2 + \lceil \log_2(2 + z_{\alpha/3}) \rceil\}. \quad (2.3.8)$$

Note that the indices of the intervals with  $t_L$  and  $t_R$  being the right end points at level  $\hat{j} + \tilde{K}_{\frac{\alpha}{4}}$  are, respectively,

$$i_L = t_L \cdot 2^{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}} \quad \text{and} \quad i_R = t_R \cdot 2^{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}.$$

Note also that  $i_R - i_L = 9 \times 2^{1 + \tilde{K}_{\frac{\alpha}{4}} + K_{\frac{\alpha}{4}}}$ , which only depends on  $\alpha$ . Define an intermediate estimator of the minimum  $M(f)$  by

$$\hat{f}_1 = \frac{1}{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}} \min_{i_L < i \leq i_R} \bar{X}_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}, i}.$$

Let  $F_n$  be the cumulative distribution function of  $\tilde{v}_n = \max\{v_1, \dots, v_n\}$ , where  $v_1, \dots, v_n \stackrel{i.i.d}{\sim} N(0, 1)$ , and define

$$S_{n, \beta} = F_n^{-1}(1 - \beta). \quad (2.3.9)$$

In other words,  $S_{n, \beta}$  is the  $(1 - \beta)$  quantile of the distribution of the maximum of  $n$  *i.i.d.* standard normal variables. Let

$$f_{lo} = \hat{f}_1 - z_{\alpha/4} \frac{\sqrt{3}\varepsilon}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}} - \frac{\sqrt{3}\varepsilon}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}}, \quad f_{hi} = \hat{f}_1 + S_{i_R - i_L, \frac{\alpha}{4}} \cdot \frac{\sqrt{3}\varepsilon}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}}.$$

Then the  $(1 - \alpha)$  level confidence interval for  $M(f)$  is defined as

$$CI_{m, \alpha} = [f_{lo}, f_{hi}]. \quad (2.3.10)$$

### 2.3.2. Statistical Optimality

Now we establish the optimality of the adaptive procedures constructed in Section 2.3.1. The results show that the data-driven estimators and the confidence intervals achieves within a constant factor of their corresponding benchmarks simultaneously for all  $f \in \mathcal{F}$ . We begin with the estimator of the minimizer.

**Theorem 2.3.1** (Estimation of Minimizer). *The estimator  $\hat{Z}$  defined in (2.3.5) satisfies*

$$\mathbb{E}_f |\hat{Z} - Z(f)| < 35\rho_z(\varepsilon; f) \leq C_z R_z(\varepsilon; f), \quad \text{for all } f \in \mathcal{F},$$

where  $C_z > 0$  is an absolute constant.

The following result holds for the confidence interval  $CI_{z,\alpha}$ .

**Theorem 2.3.2** (Confidence Interval for the Minimizer). *Let  $0 < \alpha < 0.3$ . The confidence interval  $CI_{z,\alpha}$  given in (2.3.6) is a  $(1 - \alpha)$  level confidence interval for the minimizer  $Z(f)$  and its expected length satisfies*

$$\mathbb{E}_f L(CI_{z,\alpha}) \leq (24 \times 2^{K_\alpha} - 3) \times 17.5 \times \rho_z(\varepsilon; f) \leq C_{z,\alpha} L_{z,\alpha}(\varepsilon; f), \quad \text{for all } f \in \mathcal{F},$$

where  $K_\alpha = \lceil \frac{\log \alpha}{\log \Phi(-2)} \rceil$  and  $C_{z,\alpha}$  is a constant depending on  $\alpha$  only.

Similarly, the estimator and confidence interval for the minimum  $M(f)$  are within a constant factor of the benchmarks simultaneously for all  $f \in \mathcal{F}$ .

**Theorem 2.3.3** (Estimation of Minimum). *The estimator  $\hat{M}$  defined in (2.3.7) satisfies*

$$\mathbb{E}_f |\hat{M} - M(f)| < 449\rho_m(\varepsilon; f) \leq C_m R_m(\varepsilon; f), \quad \text{for all } f \in \mathcal{F},$$

where  $C_m > 0$  is an absolute constant.

**Theorem 2.3.4** (Confidence Interval for the Minimum). *The confidence interval  $CI_{m,\alpha}$  given in (2.3.10) is a  $(1 - \alpha)$  confidence interval for the minimum  $M(f)$  and when  $0 < \alpha <$*

0.3, its expected length satisfies

$$\mathbb{E}_f L(CI_{m,\alpha}) \leq c_{m,\alpha} \rho_m(\varepsilon; f) \leq C_{m,\alpha} L_{m,\alpha}(\varepsilon; f), \quad \text{for all } f \in \mathcal{F},$$

where  $c_{m,\alpha}$  and  $C_{m,\alpha}$  are constants depending on  $\alpha$  only.

## 2.4. Nonparametric Regression

We have so far focused on the white noise model. The procedures and results presented in the previous sections can be extended to nonparametric regression, where we observe

$$y_i = f(x_i) + \sigma z_i, \quad i = 0, 1, 2, \dots, n, \quad (2.4.1)$$

with  $x_i = \frac{i}{n}$ , and  $z_i \stackrel{i.i.d}{\sim} N(0, 1)$ . The noise level  $\sigma$  is assumed to be known. The tasks are the same as before: construct optimal estimators and confidence intervals for the minimizer and minimum of  $f \in \mathcal{F}$ .

### 2.4.1. Benchmarks and Discretization Errors

Analogous to the benchmarks for the white noise model defined in Equations (2.1.1), (2.1.2), (2.1.4), (2.1.5), we define similar benchmarks for the nonparametric regression model (2.4.1) with  $n + 1$  equally spaced observations. Denote by  $\mathcal{I}_{z,\alpha,n}(\mathfrak{F})$  and  $\mathcal{I}_{m,\alpha,n}(\mathfrak{F})$  respectively the collections of  $(1 - \alpha)$  level confidence intervals for  $Z(f)$  and  $M(f)$  on a function class  $\mathfrak{F}$  under the regression model (2.4.1) and let

$$\begin{aligned} \tilde{R}_{z,n}(\sigma; f) &= \sup_{g \in \mathcal{F}} \inf_{\hat{Z}} \max_{h \in \{f, g\}} \mathbb{E}_h |\hat{Z} - Z(h)|, \\ \tilde{R}_{m,n}(\sigma; f) &= \sup_{g \in \mathcal{F}} \inf_{\hat{M}} \max_{h \in \{f, g\}} \mathbb{E}_h |\hat{M} - M(h)|, \\ \tilde{L}_{z,\alpha,n}(\sigma; f) &= \sup_{g \in \mathcal{F}} \inf_{CI \in \mathcal{I}_{z,\alpha,n}(\{f, g\})} \mathbb{E}_f L(CI), \\ \tilde{L}_{m,\alpha,n}(\sigma; f) &= \sup_{g \in \mathcal{F}} \inf_{CI \in \mathcal{I}_{m,\alpha,n}(\{f, g\})} \mathbb{E}_f L(CI). \end{aligned} \quad (2.4.2)$$

Compared with the white noise model, estimation and inference for both  $Z(f)$  and  $M(f)$  incur additional discretization errors, even in the noiseless case. See the appendix Section A.1.9 for further discussion.

#### 2.4.2. Data-driven Procedures

Similar to the white noise model, we first split the data into three independent copies and then construct the estimators and confidence intervals for  $Z(f)$  and  $M(f)$  in three major steps: localization, stopping, and estimation/inference.

##### Data Splitting

Let  $z_{1,0}, z_{1,1}, \dots, z_{1,n}, z_{2,0}, z_{2,1}, \dots, z_{2,n}$  be i.i.d. standard normal random variables, and all be independent of the observed data  $\{y_1, \dots, y_n\}$ . We construct the following three sequences:

$$\begin{aligned} y_{l,i} &= y_i + \frac{\sqrt{2}}{2}\sigma z_{1,i} + \frac{\sqrt{6}}{2}\sigma z_{2,i}, \\ y_{s,i} &= y_i + \frac{\sqrt{2}}{2}\sigma z_{1,i} - \frac{\sqrt{6}}{2}\sigma z_{2,i}, \\ y_{e,i} &= y_i - \sqrt{2}\sigma z_{1,i}, \end{aligned} \tag{2.4.3}$$

for  $i = 0, \dots, n$ . For convenience, let  $y_{l,i} = y_{s,i} = y_{e,i} = \infty$  for  $i \notin \{0, 1, \dots, n\}$ . It is easy to see that these random variables are all independent with the same variance  $3\sigma^2$  for  $i \in \{0, 1, \dots, n\}$ . We will use  $\{y_{l,i}\}$  for localization,  $\{y_{s,i}\}$  for devising the stopping rule, and  $\{y_{e,i}\}$  for constructing the final estimation and inference procedures.

Let  $J = \lfloor \log_2(n+1) \rfloor$ . For  $j = 0, 1, \dots, J$ ,  $i = 1, 2, \dots, \lfloor \frac{n+1}{2^{J-j-1}} \rfloor$ , the  $i$ -th block at level  $j$  consists of  $\{x_{(i-1)2^{J-j}}, x_{(i-1)2^{J-j}+1}, \dots, x_{i \cdot 2^{J-j-1}}\}$ . Denote the sum of the observations in the  $i$ -th block at level  $j$  for the sequence  $u$  ( $u = l, s, e$ ) as

$$Y_{j,i,u} = \sum_{k=(i-1)2^{J-j}}^{i \cdot 2^{J-j-1}} y_{u,k}, \text{ for } u = l, s, e.$$

Again, let  $Y_{j,i,u} = +\infty$  when  $i \in \mathbb{Z} \setminus \{1, 2, \dots, \lfloor \frac{n+1}{2^{J-j-1}} \rfloor\}$ , for  $u = l, s, e$ .

### Localization

We now use  $\{y_{l,i}, i = 0, \dots, n\}$  to construct a localization procedure. Let  $\hat{\mathbf{i}}_0 = 1$ , and for  $j = 1, 2, \dots, J$ , let

$$\hat{\mathbf{i}}_j = \arg \min_{\max\{2\hat{\mathbf{i}}_{j-1}-2, 1\} \leq i \leq \min\{2\hat{\mathbf{i}}_{j-1}+1, \lfloor \frac{n+1}{2^{J-j}} \rfloor\}} Y_{j,i,l}.$$

This is similar to the localization step in the white noise model. In each iteration, the blocks at the previous level are split into two sub-blocks. The  $i$ -th block at level  $j-1$  is split into two blocks, the  $(2i-1)$ -st block and  $2i$ -th block, at level  $j$ . For a given  $\hat{\mathbf{i}}_{j-1}$ ,  $\hat{\mathbf{i}}_j$  is the subblock with the smallest sum among the two subblocks of  $(j-1)$ -level-block  $\hat{\mathbf{i}}_{j-1}$  and their immediate neighboring subblocks.

### Stopping Rule

Similar to the stopping rule for the white noise model, define the statistic  $\mathbf{T}_j$  as

$$\mathbf{T}_j = \min\{Y_{j,\hat{\mathbf{i}}_j+6,s} - Y_{j,\hat{\mathbf{i}}_j+5,s}, Y_{j,\hat{\mathbf{i}}_j-6,s} - Y_{j,\hat{\mathbf{i}}_j-5,s}\}.$$

Let  $\tilde{\sigma}_j^2 = 6 \times 2^{J-j} \sigma^2$ . It is easy to see that when  $Y_{j,\hat{\mathbf{i}}_j+6,s} - Y_{j,\hat{\mathbf{i}}_j+5,s} < \infty$ ,

$$Y_{j,\hat{\mathbf{i}}_j+6,s} - Y_{j,\hat{\mathbf{i}}_j+5,s} | \hat{\mathbf{i}}_j \sim N\left(\sum_{k=(\hat{\mathbf{i}}_j+4)2^{J-j}}^{(\hat{\mathbf{i}}_j+5)2^{J-j}-1} f(x_{k+2^{J-j}}) - f(x_k), \tilde{\sigma}_j^2\right). \quad (2.4.4)$$

Define

$$\check{j} = \begin{cases} \min\{j : \mathbf{T}_j \leq 2\tilde{\sigma}_j\} & \text{if } \{j : \mathbf{T}_j \leq 2\tilde{\sigma}_j\} \cap \{0, 1, 2, \dots, J\} \neq \emptyset \\ \infty & \text{otherwise} \end{cases}$$

and terminate the algorithm at level  $\hat{\mathbf{j}} = \min\{J, \check{j}\}$ . So, either  $\mathbf{T}_j$  triggers the stopping rule

for some  $0 \leq j \leq J$  or the algorithm reaches the highest possible level  $J$ .

With the localization strategy and the stopping rule, the final block, the  $\hat{\mathbf{i}}_{\hat{\mathbf{j}}}$ -th block at level  $\hat{\mathbf{j}}$ , is given by  $\{x_k : (\hat{\mathbf{i}}_{\hat{\mathbf{j}}} - 1)2^{J-\hat{\mathbf{j}}} \leq k \leq \hat{\mathbf{i}}_{\hat{\mathbf{j}}}2^{J-\hat{\mathbf{j}}} - 1\}$ .

### Estimation and Inference

After we have our final block,  $\hat{\mathbf{i}}_{\hat{\mathbf{j}}}$ -th block at level  $\hat{\mathbf{j}}$ , we use it to construct estimators and confidence intervals for the minimizer  $Z(f)$  and the minimum  $M(f)$ . We start with the estimation of  $Z(f)$ . The estimator of  $Z(f)$  is given as follows:

$$\hat{Z} = \begin{cases} -\frac{1}{2n} + \frac{1}{n}(2^{J-\hat{\mathbf{j}}}\hat{\mathbf{i}}_{\hat{\mathbf{j}}} - 2^{J-\hat{\mathbf{j}}-1}), & \check{j} < \infty \\ \frac{1}{n} \arg \min_{\hat{\mathbf{i}}_{\hat{\mathbf{j}}-2} \leq i \leq \hat{\mathbf{i}}_{\hat{\mathbf{j}}+2} y_{e,i-1} - \frac{1}{n}, & \check{j} = \infty \end{cases} \quad (2.4.5)$$

To construct the confidence interval for  $Z(f)$ , we take a few adjacent blocks to the left and right of  $\hat{\mathbf{i}}_{\hat{\mathbf{j}}}$ -th block at level  $\hat{\mathbf{j}}$ . Let

$$\mathbf{L} = \max\{0, \hat{\mathbf{i}}_{\hat{\mathbf{j}}} - 12 \times 2^{K_{\alpha/2}} + 1\} \quad \text{and} \quad \mathbf{U} = \min\{[(n+1)2^{\hat{\mathbf{j}}-J}], \hat{\mathbf{i}}_{\hat{\mathbf{j}}} + 12 \times 2^{K_{\alpha/2}} - 2\}.$$

When  $\check{j} < \infty$ , let

$$t_{lo} = \frac{2^{J-\hat{\mathbf{j}}}}{n}\mathbf{L} - \frac{1}{2n} \quad \text{and} \quad t_{hi} = \frac{2^{J-\hat{\mathbf{j}}}}{n}\mathbf{U} - \frac{1}{2n}.$$

When  $\check{j} = \infty$ ,  $t_{lo}$  and  $t_{hi}$  are calculated by the following Algorithm 1. Note that  $\check{j} = \infty$  means that the procedure is forced to end and the discretization error can be dominant.

Algorithm 1 first iteratively shrinks the original interval  $[t_{lo} - \frac{1}{n}, t_{hi} + \frac{1}{n}]$  to find the minimizer  $\frac{im}{n}$  of the function  $f$  among the  $n+1$  sample points with high probability. In each iteration, the algorithm tests whether the slopes of the segments on both ends are positive or negative. It shrinks the left end with negative slope (on the left), or shrinks the right end with positive slope (on the right), or stops if no further shrinking is needed on either side.



Note that the minimizer of any convex function with given values at these  $n + 1$  points is smaller than the intersection of the following two lines:

$$y = f\left(\frac{i_m}{n}\right) \quad \text{and} \quad y = \frac{f\left(\frac{i_m+2}{n}\right) - f\left(\frac{i_m+1}{n}\right)}{1/n} \left(t - \frac{i_m+1}{n}\right) + f\left(\frac{i_m+1}{n}\right). \quad (2.4.6)$$

Note that these two lines are determined by  $f\left(\frac{i_m}{n}\right)$ ,  $f\left(\frac{i_m+1}{n}\right)$  and  $f\left(\frac{i_m+2}{n}\right)$  only. Given the noisy observations at these three points,  $\frac{i_m}{n}$ ,  $\frac{i_m+1}{n}$ , and  $\frac{i_m+2}{n}$ , the range of these two lines and the intersection can be inferred, and the right side of the interval can then be shrunk accordingly.

Same is done for the left side of the confidence interval. In addition, boundary cases and other complications need to be considered, which are handled in Algorithm 1.

Note that our construction and the theoretical results only rely on convexity. In particular, the existence of second order derivative is not needed as it is commonly assumed in the literature. This is an important contributing factor to optimality under the non-asymptotic local minimax framework.

---

**Algorithm 1** Computing  $t_{lo}$  and  $t_{hi}$  when  $\check{j} = \infty$

---

$L \leftarrow \max\{1, \hat{\mathbf{i}}_{\check{j}} - 12 \times 2^{K_{\alpha/2}}\} - 1$ ,  $U \leftarrow \min\{n+1, \hat{\mathbf{i}}_{\check{j}} + 12 \times 2^{K_{\alpha/2}}\} - 1$ ,  $\alpha_1 \leftarrow \frac{\alpha}{8}$ ,  $\alpha_2 = \alpha/24$

Generate  $z_{3,0}, z_{3,1} \dots, z_{3,n} \stackrel{i.i.d.}{\sim} N(0, 1)$

$i_l \leftarrow \min\{\{U\} \cup \{i \in [L, U-1] : y_{e,i} + \sqrt{3}\sigma z_{3,i} - (y_{e,i+1} + \sqrt{3}\sigma z_{3,i+1}) \leq 2\sqrt{3}\sigma z_{\alpha_1}\}\}$

$i_r \leftarrow \max\{\{L-1\} \cup \{i \in [L, U-1] : y_{e,i} + \sqrt{3}\sigma z_{3,i} - (y_{e,i+1} + \sqrt{3}\sigma z_{3,i+1}) \geq -2\sqrt{3}\sigma z_{\alpha_1}\}\}$

**if**  $i_l = U$  **then**

**if**  $i_l = n$  and  $y_{e,n-2} - y_{e,n-1} - \sqrt{3}\sigma(z_{3,n-2} - z_{3,n-1}) + 2\sqrt{6}\sigma z_{\alpha_2} > 0$  **then**

$t_{lo} \leftarrow \left( \left( -\frac{y_{e,n} - y_{e,n-1} - \sqrt{3}\sigma(z_{3,n} - z_{3,n-1}) + 2\sqrt{6}\sigma z_{\alpha_2}}{n(y_{e,n-2} - y_{e,n-1} - \sqrt{3}\sigma(z_{3,n-2} - z_{3,n-1}) + 2\sqrt{6}\sigma z_{\alpha_2})} + \frac{n-1}{n} \right) \vee \frac{n-1}{n} \right) \wedge \frac{n}{n}$ ,  $t_{hi} \leftarrow 1$

**else**

$t_{lo} = t_{hi} = U/n$

**end if**

**end if**

**if**  $i_r = L-1$  **then**

**if**  $i_r = -1$  and  $y_{e,2} - y_{e,1} - \sqrt{3}\sigma(z_{3,2} - z_{3,1}) + 2\sqrt{6}\sigma z_{\alpha_2} > 0$  **then**

$t_{hi} \leftarrow \left( \left( \frac{y_{e,0} - y_{e,1} - \sqrt{3}\sigma(z_{3,0} - z_{3,1}) + 2\sqrt{6}\sigma z_{\alpha_2}}{n(y_{e,2} - y_{e,1} - \sqrt{3}\sigma(z_{3,2} - z_{3,1}) + 2\sqrt{6}\sigma z_{\alpha_2})} + \frac{1}{n} \right) \vee \frac{0}{n} \right) \wedge \frac{1}{n}$ ,  $t_{lo} = 0$

**else**

$t_{lo} = t_{hi} = 0$

**end if**

**end if**

**if**  $(i_l - U)(i_r - L + 1) \neq 0$  **then**

$i_{lo} \leftarrow (i_l - 1) \vee L$ ,  $i_{hi} \leftarrow (i_r + 2) \wedge U$

**if**  $i_{hi} - i_{lo} \geq 3$  or  $(i_{hi} - n)i_{lo} = 0$  **then**

$t_{lo} = i_{lo}/n$ ,  $t_{hi} = i_{hi}/n$

**else if**  $y_{e,i_{hi}+1} - y_{e,i_{hi}} - \sqrt{3}\sigma(z_{3,i_{hi}+1} - z_{3,i_{hi}}) \leq -2\sqrt{6}\sigma z_{\alpha_2}$  or  $y_{e,i_{lo}-1} - y_{e,i_{lo}} - \sqrt{3}\sigma(z_{3,i_{lo}-1} - z_{3,i_{lo}}) \leq -2\sqrt{6}\sigma z_{\alpha_2}$  **then**

$t_{lo} = t_{hi} = (i_{hi} + i_{lo})/2n$

**else**

$t_{hi} \leftarrow \left( \left( \frac{y_{e,i_{hi}-1} - y_{e,i_{hi}} - \sqrt{3}\sigma(z_{3,i_{hi}-1} - z_{3,i_{hi}}) + 2\sqrt{6}\sigma z_{\alpha_2}}{n(y_{e,i_{hi}+1} - y_{e,i_{hi}} - \sqrt{3}\sigma(z_{3,i_{hi}+1} - z_{3,i_{hi}}) + 2\sqrt{6}\sigma z_{\alpha_2})} + \frac{i_{hi}}{n} \right) \vee \frac{i_{hi}-1}{n} \right) \wedge \frac{i_{hi}}{n}$

$t_{lo} \leftarrow \left( \left( -\frac{y_{e,i_{lo}+1} - y_{e,i_{lo}} - \sqrt{3}\sigma(z_{3,i_{lo}+1} - z_{3,i_{lo}}) + 2\sqrt{6}\sigma z_{\alpha_2}}{n(y_{e,i_{lo}-1} - y_{e,i_{lo}} - \sqrt{3}\sigma(z_{3,i_{lo}-1} - z_{3,i_{lo}}) + 2\sqrt{6}\sigma z_{\alpha_2})} + \frac{i_{lo}}{n} \right) \vee \frac{i_{lo}}{n} \right) \wedge \frac{i_{lo}+1}{n}$

**end if**

**end if**

---

The  $(1 - \alpha)$ -level confidence interval for the minimizer  $Z(f)$  is given by

$$\text{CI}_{z,\alpha} = [t_{lo} \wedge t_{hi}, t_{hi}] \quad (2.4.7)$$

We now construct the estimator and confidence interval for the minimum  $M(f)$ . Let  $\Delta =$

$\mathbb{1}\{Y_{\hat{\mathbf{j}}, \hat{\mathbf{i}}_{\hat{\mathbf{j}}+6, s} - Y_{\hat{\mathbf{j}}, \hat{\mathbf{i}}_{\hat{\mathbf{j}}+5, s} \leq 2\sqrt{6}\sigma\sqrt{2^{J-\hat{\mathbf{j}}}}\} - \mathbb{1}\{Y_{\hat{\mathbf{j}}, \hat{\mathbf{i}}_{\hat{\mathbf{j}}-6, s} - Y_{\hat{\mathbf{j}}, \hat{\mathbf{i}}_{\hat{\mathbf{j}}-5, s} \leq 2\sqrt{6}\sigma\sqrt{2^{J-\hat{\mathbf{j}}}}\}$  and define

$$\tilde{\mathbf{i}}_{\hat{\mathbf{j}}} = \begin{cases} \hat{\mathbf{i}}_{\hat{\mathbf{j}}} + 2\Delta & \text{if } \check{j} < \infty \\ \arg \min_{\hat{\mathbf{i}}_{\hat{\mathbf{j}}-2} \leq i \leq \hat{\mathbf{i}}_{\hat{\mathbf{j}}+2} y_{e, i-1} & \text{if } \check{j} = \infty \end{cases}. \quad (2.4.8)$$

The estimator of  $M(f)$  is then given by the average of the observations of the copy for estimation and inference in the  $\tilde{\mathbf{i}}_{\hat{\mathbf{j}}}$ -th block at level  $\hat{\mathbf{j}}$ ,

$$\hat{M} = \frac{1}{2^{J-\hat{\mathbf{j}}}} Y_{\hat{\mathbf{j}}, \tilde{\mathbf{i}}_{\hat{\mathbf{j}}}, e}. \quad (2.4.9)$$

To construct the confidence interval for  $M(f)$ , we specify two levels  $j_s$  and  $j_l$ , with

$$j_s = \max\{0, \hat{\mathbf{j}} - K_{\frac{\alpha}{4}} - 1\} \quad \text{and} \quad j_l = \min\{J, \hat{\mathbf{j}} + \tilde{K}_{\frac{\alpha}{4}}\},$$

where  $\tilde{K}_{\frac{\alpha}{4}}$  is defined as in Equation (2.3.8). It will be shown that at level  $j_s$ ,  $Z(f)$  is within four blocks of the chosen block with probability at least  $1 - \frac{\alpha}{4}$ , and at level  $j_l$ , with probability at least  $1 - \frac{\alpha}{4}$ , the length of the block is no larger than  $\rho_z(\frac{\sigma}{\sqrt{n}}; f)$ . Define

$$I_{lo} = \max\{1, 2^{j_l-j_s}(\hat{\mathbf{i}}_{j_s} - 5)\}, \quad I_{hi} = \min\{2^{j_l-j_s}(\hat{\mathbf{i}}_{j_s} + 4) + 1, \lceil \frac{n+1}{2^{J-j_l}} \rceil\}.$$

It can be shown that the minimizer  $Z(f)$  lies with high probability in the interval

$$\left[ \frac{2^{J-j_l}(I_{lo} - 1)}{n}, \frac{2^{J-j_l}I_{hi} - 1}{n} \right] \cap [0, 1].$$

Define an intermediate estimator for  $M(f)$  by

$$\hat{\mathbf{f}}_1 = \min_{I_{lo} \leq i \leq I_{hi}} \frac{1}{2^{J-j_l}} Y_{j_l, i, e}.$$

Let

$$\mathbf{f}_{hi} = \hat{\mathbf{f}}_1 + S_{I_{hi}-I_{lo}+1, \frac{\alpha}{4}} \frac{\sqrt{3}\sigma}{\sqrt{2^{J-j_l}}}$$

where  $S_{n,\beta}$  is defined in Equation (2.3.9) in Section 2.3. This is the upper limit of the confidence interval, now we define the lower limit  $\mathbf{f}_{lo}$ .

When  $\hat{\mathbf{j}} + \tilde{K}_{\frac{\alpha}{4}} \leq J$ , let

$$\mathbf{f}_{lo} = \hat{\mathbf{f}}_1 - (z_{\alpha/4} + 1) \frac{\sqrt{3}\sigma}{\sqrt{2^{J-j_l}}}.$$

When  $\hat{\mathbf{j}} + \tilde{K}_{\frac{\alpha}{4}} > J$ , we compute  $\mathbf{f}_{lo}$  by Algorithm 2, which is based on the geometric property of the convex function  $f$  that for any  $1 \leq k \leq n-2$ ,

$$\min\{f(x_k), f(x_{k+1})\} \geq \inf_{t \in [\frac{k}{n}, \frac{k+1}{n}]} \max \left\{ \frac{f(x_{k+2}) - f(x_{k+1})}{1/n} (t - x_{k+1}) + f(x_{k+1}), \right. \\ \left. \frac{f(x_k) - f(x_{k-1})}{1/n} (t - x_k) + f(x_k) \right\}.$$

---

**Algorithm 2** Computing  $\mathbf{f}_{lo}$  when  $\hat{\mathbf{j}} + \tilde{K}_{\frac{\alpha}{4}} > J$

---

```

 $H \leftarrow S_{I_{hi}-I_{lo}+3, \frac{1}{8}} \sqrt{3}\sigma, k_l \leftarrow I_{lo} - 1, k_r \leftarrow I_{hi} - 2$ 
if  $I_{lo} = 1$  then
   $v_{r,0}(t) \leftarrow \frac{y_{e,2}-y_{e,1}+2H}{1/n} (t - 1/n) + y_{e,1} - H, h(0) \leftarrow \min_{t \in [0, 1/n]} v_{r,0}(t), k_l \leftarrow I_{lo}$ 
end if
if  $I_{hi} - 1 = n$  then
   $v_{l,n-1}(t) \leftarrow \frac{y_{e,n-1}-y_{e,n-2}+2H}{1/n} (t - \frac{n-1}{n}) + y_{e,n-1} - H, h(n-1) = \min_{t \in [\frac{n-1}{n}, 1]} v_{l,n-1}(t),$ 
   $k_r \leftarrow I_{hi} - 3$ 
end if
for  $i = k_l, \dots, k_r$  do
  Define two linear functions:
   $v_{l,i}(t) = \frac{y_{e,i}-y_{e,i-1}+2H}{1/n} (t - x_i) + y_{e,i} - H, v_{r,i} = \frac{y_{e,i+2}-y_{e,i+1}+2H}{1/n} (t - x_{i+1}) + y_{e,i+1} - H$ 
   $h(i) = \min_{t \in [x_i, x_{i+1}]} \max\{v_{l,i}(t), v_{r,i}(t)\}$ 
end for
 $\mathbf{f}_{lo} \leftarrow \min\{h(i) : I_{lo} - 1 \leq i \leq I_{hi} - 2\} \wedge \mathbf{f}_{hi}$ 

```

---

Note that  $h(i)$  in Algorithm 2 is derived from one or two linear functions, so given the relationship of the function values at two end points of the corresponding interval, it has an explicit form. Hence the procedure is still computationally efficient.

The  $(1 - \alpha)$ -level confidence interval for the minimum  $M(f)$  is given by

$$\mathbf{CI}_{m,\alpha} = [\mathbf{f}_{lo}, \mathbf{f}_{hi}]. \quad (2.4.10)$$

*Remark 2.4.1.* As mentioned in the introduction, Agarwal et al. (2011) proposes an algorithm for stochastic convex optimization with bandit feedback. While both our procedures and the method in Agarwal et al. (2011) include an ingredient trying to localize the minimizer through shrinking intervals by exploiting the convexity of the underlying function, the two methods are essentially different due to the significant differences in both the designs and loss functions. The goal of exploiting convexity in Agarwal et al. (2011) is mainly for deciding the direction of shrinking their intervals, while ours is mainly for deciding when to stop and what to do after stopping.

### 2.4.3. Statistical Optimality

Now we establish the optimality of the adaptive procedures constructed in Section 2.4.2. The regression model is similar to the white noise model, but with additional discretization errors. The results show that our data-driven procedures are simultaneously optimal (up to a constant factor) for all  $f \in \mathcal{F}$ . We begin with the estimator of the minimizer.

**Theorem 2.4.1** (Estimation of the Minimizer). *The estimator  $\hat{Z}$  of the minimizer  $Z(f)$  defined in (2.4.5) satisfies*

$$\mathbb{E}_f |\hat{Z} - Z(f)| \leq C_1 \tilde{R}_{z,n}(\sigma; f), \quad \text{for all } f \in \mathcal{F}, \quad (2.4.11)$$

where  $C_1 > 0$  is an absolute constant.

The following result holds for the confidence interval  $\text{CI}_{z,\alpha}$  of  $Z(f)$ .

**Theorem 2.4.2.** *Let  $0 < \alpha < 0.3$ . The confidence interval  $\text{CI}_{z,\alpha}$  given in (2.4.7) is a  $(1 - \alpha)$ -level confidence interval for the minimizer  $Z(f)$  and its expected length satisfies*

$$\mathbb{E}_f L(\text{CI}_{z,\alpha}) \leq C_{2,\alpha} \tilde{L}_{z,\alpha,n}(\sigma; f), \quad \text{for all } f \in \mathcal{F},$$

where  $C_{2,\alpha}$  is a constant depending on  $\alpha$  only.

Similarly, the estimator and confidence interval for the minimum  $M(f)$  are within a constant

factor of the benchmarks simultaneously for all  $f \in \mathcal{F}$ .

**Theorem 2.4.3** (estimation for the minimum). *The estimator  $\hat{M}$  defined in (2.4.9) satisfies*

$$\mathbb{E}_f |\hat{M} - M(f)| \leq C_3 \tilde{R}_{m,n}(\sigma; f), \quad \text{for all } f \in \mathcal{F},$$

where  $C_3$  is an absolute constant.

**Theorem 2.4.4.** *Let  $0 < \alpha < 0.3$ . The confidence interval  $\text{CI}_{m,\alpha}$  given in (2.4.10) is a  $(1 - \alpha)$ -level confidence interval and its expected length satisfies*

$$\mathbb{E}_f L(\text{CI}_{m,\alpha}) \leq C_{4,\alpha} \tilde{L}_{m,\alpha,n}(\sigma; f), \quad \text{for all } f \in \mathcal{F},$$

where  $C_{4,\alpha}$  is a constant depending only on  $\alpha$ .

#### 2.4.4. Comparison with constrained least squares methods

The convexity-constrained least squares (CLS) estimator is perhaps the most commonly used method for estimating a convex regression function globally. Estimation and inference methods for the minimizer based on the CLS estimator have been proposed and investigated in the literature (e.g., Shoung et al. (2001); Ghosal and Sen (2017); Deng et al. (2020)). Theoretical analyses typically assume that the second or higher order derivatives exist with an even order derivative being positive and all lower order derivatives being zero at the minimizer. It is unclear how the CLS estimator behaves under our nonasymptotic framework or even asymptotically in general when the underlying convex function is nonsmooth at the minimizer. As for estimation and inference for the minimum, to the best of our knowledge, there is no CLS based method with theoretical guarantees.

It is interesting to compare with the CLS confidence interval for the minimizer proposed in Deng et al. (2020). Let  $\hat{f}_n = \min_{f \in \mathcal{F}} \sum_{i=1}^n (y_i - f(x_i))^2$  be the CLS estimator. Let  $\hat{m}_n$  be the anti-mode of  $\hat{f}_n$ ,  $\hat{v}_m$  (resp.  $\hat{u}_m$ ) be the first kink of  $\hat{f}_n$  to the right (resp. left) of  $\hat{m}_n$ . Under the assumption that the second order derivative exists and is positive around

the minimizer, Deng et al. (2020) introduces the following  $(1 - \alpha)$ -level confidence interval,

$$CLSCI_\alpha = [\hat{m}_n \pm c_\alpha^m(\hat{v}_m - \hat{u}_m)] \cap [0, 1], \quad (2.4.12)$$

where  $c_\alpha^m$  is a constant depending on  $\alpha$  only.

Denote by  $\mathcal{F}_2$  the collection of convex functions with continuous positive second order derivative around the minimizer. Deng et al. (2020) shows that the confidence interval  $CLSCI_\alpha$  has desired coverage probability asymptotically over  $\mathcal{F}_2$ . The following result shows that  $CLSCI_\alpha$  defined in (2.4.12) is sub-optimal under the local minimax framework.

**Proposition 2.4.1.** *For any sample size  $n \geq 5$ ,*

$$\sup_{f \in \mathcal{F}_2} \frac{\mathbb{E}_f L(CLSCI_\alpha)}{\mathbb{E}_f L(\mathbf{CI}_{z,\alpha})} = \infty. \quad (2.4.13)$$

This result shows that for any given  $n \geq 5$ , there exists  $f \in \mathcal{F}_2$  such that the length of the confidence interval  $CLSCI_\alpha$  at  $f$  is much larger than the length of our proposed confidence interval  $\mathbf{CI}_{z,\alpha}$ . The non-asymptotic nature of our framework and the asymptotic nature of  $CLSCI_\alpha$  are a key contributing factor to this phenomenon. In the appendix Section A.3.1, through an example, we intuitively demonstrate the sub-optimality in the construction of the CLS confidence interval. In short, only looking at the kinks does not fully utilize the convexity property.

For estimation of the minimizer, all the existing analyses of the CLS estimator are based on the limiting distribution under strong regularity assumptions. So they are asymptotic in nature. For example, the rate of convergence of the CLS estimator is  $n^{-1/5}$  for the minimizer of over the function class  $\mathcal{F}_2$ . It can be shown that our estimator  $\hat{Z}$  of the minimizer given in (2.4.5) also achieves the same rate over  $\mathcal{F}_2$ . The properties of the CLS estimator under the non-asymptotic local minimax framework are unclear and difficult to analyze. We investigate the empirical performance of the CLS estimator through simulations. Simulation results are summarized in Section 2.4.5, with details given in the appendix Section A.4.

### 2.4.5. Numerical Results

The proposed algorithms are easy to implement and computationally fast. We implement the algorithms in R and the code is available at <https://github.com/chenrancece/MMCF>. The data splitting procedure in our proposed algorithm was introduced to create independence, which is purely for technical reasons, we also include a variant of our method without the data splitting step. That is, the original data set is used in the localization, stopping, and estimation/inference steps. Simulation studies are carried out to investigate the numerical performance of the proposed algorithms and this non-split variant as well as make comparisons with the CLS confidence interval  $CLSCI_\alpha$  in (2.4.12) proposed by Deng et al. (2020) and the CLS estimator for the minimizer. For reasons of space, we provide a brief summary of the numerical results here and give the detailed simulation results and discussions in the appendix Section A.4.

The simulation studies use 10 test functions with different levels of smoothness around the minimizer, 6 sample sizes ranging from 100 to 50,000, 5 confidence levels for the confidence intervals, and 100 replications. We compared the proposed methods, their non-split variant, and the CLS methods in terms of computational time, average absolute error (for the estimators), and coverage probability and length (for the confidence intervals). We also investigated the relationship with the benchmarks when the benchmarks can be calculated explicitly. The results can be summarized as follows.

- **Computational cost:** Our methods are significantly faster than CLS methods. For small sample sizes, all methods run relatively fast. For  $n \geq 5000$ , our procedures are at least 10 times faster than the CLS methods for all functions. In many cases, they are more than 100 times faster. This gap is further increased as the sample size grows.
- **Confidence interval for the minimizer:** Our methods achieve the nominal coverage consistently and the empirical lengths are proportional to the benchmark. In comparison, the coverage probability of  $CLSCI_\alpha$  can be far below the nominal level



for a variety of functions, including functions that are not differentiable at the minimizer or have vanishing second order derivative around the minimizer. For piecewise linear function such as  $100 \cdot |2x - 1|$ ,  $CLSCI_\alpha$  is long and its length remains roughly a constant as the sample size increases, while the benchmark goes to zero.

- **Estimation of the minimizer:** The numerical performances of our methods and the CLS estimator are comparable. Interestingly, in the cases where the benchmarks can be calculated explicitly, the performance of the CLS estimator relative to the benchmarks (and our methods) deteriorates with increasing smoothness of the function around the minimizer, while the performance of our estimator remains steady relative to the benchmarks.
- **Estimation and CI for the minimum:** We are unaware of theoretically guaranteed CLS estimator or confidence interval for the minimum, so we only examined the performance of our methods. The empirical absolute error for estimator and the lengths of the confidence intervals for the minimum exhibit linear relationship with the corresponding benchmarks (when calculable). The nominal coverages of the confidence intervals are achieved in all the settings.

## 2.5. Discussion

In this chapter of dissertation, we studied optimal estimation and inference for the minimizer and minimum of a convex function in the white noise and nonparametric regression models under a non-asymptotic local minimax framework. It is shown in the appendix Section A.3.2 the results obtained in this chapter can be readily used to establish the optimal rates of convergence over the convex smoothness classes under the classical minimax framework: the lower bounds under this framework can be easily transferred into the ones under the conventional minimax framework and the optimal procedures under this framework is automatically adaptively optimal under the conventional framework. The converse is not true: procedures that are minimax optimal in the classical sense can be sub-optimal under

the local minimax framework.

A key advantage of our non-asymptotic local minimax framework is that it enables the characterization of the difficulty for estimating individual functions, and makes establishing the non-superefficiency type of results conceptually possible. Another significant advantage is that our framework manifests novel phenomena that cannot be seen in the classical minimax theory. The Uncertainty Principle established in this chapter shows the fundamental tension between the estimation accuracy for the minimizer and that for the minimum of a convex function. Analogous results also hold for the inference accuracy. It would be interesting to establish uncertainty principles in other statistical problems such as stochastic optimization with bandit feedback under the shape constraints.

The present work can be extended in different directions. For estimation, the absolute error was used as the loss function in this chapter of the dissertation. The results can be easily generalized to the  $\ell_q$  loss for  $q > 1$ . In this chapter, we focused on the minimizer and minimum of a univariate convex function. In the next chapter, consider the multivariate setting with the convexity constraint on individual nonzero components. It would also be interesting to further extend to the high-dimensional sparse additive model with the convexity constraint on individual nonzero components. It is also interesting to consider the extremum under more general shape constraints such as  $s$ -convexity. In addition, estimation and inference for other nonlinear functionals such as the quadratic functional, entropies, and divergences under a similar non-asymptotic local minimax framework can be studied. We expect the penalty-of-superefficiency property to hold in these problems and our approach to be particularly helpful for the construction of the confidence intervals.

We believe the non-asymptotic local minimax framework is most advantageous when the difficulty of estimation/inference varies significantly from function to function. Another important direction is to apply our non-asymptotic local minimax framework to other statistical models such as estimation and inference the mode and the maximum of a log concave density function based on i.i.d. observations. We expect similar Uncertainty Principles to

hold in this problem.

## 2.6. Proofs

We prove Theorems 2.2.1 and 2.2.2 here. To avoid interrupting the logic flow, other results are proved in the appendix.

### 2.6.1. Proof of Theorem 2.2.1

We begin with the lower bounds by first proving that  $R_z(\varepsilon; f) \geq \Phi(-0.5)\omega_z(\varepsilon; f)$ . The proof for  $R_m(\varepsilon; f) \geq \Phi(-0.5)\omega_m(\varepsilon; f)$  is analogous and will hence be omitted.

Let  $f \in \mathcal{F}$ . Let  $g \in \mathcal{F}$ , which we will specify later. Take  $\theta \in \{1, -1\}$  as a parameter to be estimated and let  $f_1 = f$  and  $f_{-1} = g$ .

Any estimator  $\hat{Z}$  of the minimizer  $Z(f_\theta)$  gives an estimator of  $\theta$  by

$$\hat{\theta} = \frac{\hat{Z} - \frac{Z(f_1) + Z(f_{-1})}{2}}{\frac{Z(f_1) - Z(f_{-1})}{2}},$$

and therefore  $\mathbb{E}_\theta |\hat{Z} - Z(f_\theta)| = |Z(f_1) - Z(f_{-1})| \mathbb{E}_\theta \frac{|\hat{\theta} - \theta|}{2}$ . On the other hand, a sufficient statistic for  $\theta$  is given by

$$W = \frac{\int_0^1 (f_1(t) - f_{-1}(t)) dY(t) - \frac{1}{2} \int_0^1 (f_1(t)^2 - f_{-1}(t)^2) dt}{\varepsilon \|f_1 - f_{-1}\|}. \quad (2.6.1)$$

Let  $\mathbb{P}_\theta$  be the probability measure associated with the white noise model corresponding to  $f_\theta$ . Then

$$W \sim N\left(\frac{\theta}{2} \cdot \frac{\|f_1 - f_{-1}\|}{\varepsilon}, 1\right) \quad \text{under } \mathbb{P}_\theta.$$

Note that for any  $\omega_z(\varepsilon; f) > \delta > 0$  there exists  $h_\delta \in \mathcal{F}$  such that  $\|f - h_\delta\|_2 = \varepsilon$  and that  $|Z(f) - Z(h_\delta)| \geq \omega_z(\varepsilon; f) - \delta$ , we let  $g = h_\delta$ . Then we have  $R_z(\varepsilon; f) \geq (\omega_z(\varepsilon; f) - \delta) \cdot r_1$ ,

where  $r_1$  is the minimax risk of the two-point problem based on an observation  $X \sim N(\frac{\theta}{2}, 1)$ ,

$$r_1 = \inf_{\hat{\theta}} \max_{\theta=\pm 1} \mathbb{E}_{\theta} \frac{|\hat{\theta} - \theta|}{2}.$$

It is easy to see that  $r_1 = \Phi(-0.5)$ . Taking  $\delta \rightarrow 0^+$ , we have  $R_z(\varepsilon; f) \geq \Phi(-0.5)\omega_z(\varepsilon; f)$ . So we have  $a_1 \geq \Phi(-0.5) \approx 0.309$ .

Next, we show for  $0 < \alpha < 0.3$  that  $L_{z,\alpha}(\varepsilon; f) \geq b_{\alpha}\omega_z(\varepsilon/3; f)$  where  $b_{\alpha} = 0.6 - 2\alpha$ . A lower bound for  $L_{m,\alpha}(\varepsilon; f)$  can be derived following a similar argument. We begin by recalling a lemma from Cai and Guo (2017).

**Lemma 2.6.1** (Cai and Guo, 2017). *For any  $CI \in \mathcal{I}_{z,\alpha}(\{f, g\})$ ,*

$$\mathbb{E}_f L(CI) \geq |Z(f) - Z(g)|(1 - 2\alpha - \text{TV}(P_f, P_g)),$$

where  $\text{TV}$  denotes the total variation distance between the two distributions of the white noise models corresponding to  $f$  and  $g$ . Similarly, for any  $CI \in \mathcal{I}_{m,\alpha}(\{f, g\})$ ,

$$\mathbb{E}_f L(CI) \geq |M(f) - M(g)|(1 - 2\alpha - \text{TV}(P_f, P_g)).$$

Again let  $g \in \mathcal{F}$ . Then for  $CI \in \mathcal{I}_{z,\alpha}(\{f, g\})$ , by Lemma 2.6.1,

$$\mathbb{E}_f L(CI) \geq |Z(f) - Z(g)|(1 - 2\alpha - \text{TV}(P_f, P_g)).$$

It is well known that  $\text{TV}(P_f, P_g) \leq \sqrt{\chi^2(P_f, P_g)}$ , where

$$\chi^2(P_f, P_g) = \int \left( \frac{dP_f}{dP_g} \right)^2 dP_g - 1$$

is the  $\chi^2$  distance between  $P_f$  and  $P_g$ . By Girsanov's theorem we can obtain the likelihood ratio

$$\frac{dP_f}{dP_g} = \exp \left( \int \frac{f(t) - g(t)}{\varepsilon^2} dY(t) - \frac{1}{2} \int \frac{f(t)^2 - g(t)^2}{\varepsilon^2} dt \right),$$

and hence

$$\begin{aligned}
\chi^2(P_f, P_g) &= \int \exp \left( 2 \int \frac{f(t) - g(t)}{\varepsilon^2} dY(t) - \int \frac{f(t)^2 - g(t)^2}{\varepsilon^2} dt \right) dP_g - 1 \\
&= \exp \left( -\frac{\|f - g\|^2}{\varepsilon^2} \right) \mathbb{E} \exp \left( 2 \int \frac{f(t) - g(t)}{\varepsilon} dW(t) \right) - 1 \\
&= \exp \left( \frac{\|f - g\|^2}{\varepsilon^2} \right) - 1.
\end{aligned}$$

Using it to bound the total variation distance, we get

$$\mathbb{E}_f L(CI) \geq |Z(f) - Z(g)| \left( 1 - 2\alpha - \sqrt{\exp \left( \frac{\|f - g\|^2}{\varepsilon^2} \right) - 1} \right).$$

We continue by specifying  $g$ . For any  $\omega_z(\varepsilon/3; f) > \delta > 0$ , picking  $g = g_\delta \in \mathcal{F}$  such that  $\|f - g_\delta\| = \varepsilon/3$  and  $|Z(f) - Z(g_\delta)| \geq \omega_z(\varepsilon/3; f) - \delta$ , we have  $\mathbb{E}_f L(CI) \geq (0.6 - 2\alpha)(\omega_z(\varepsilon/3; f) - \delta)$ .

By taking  $\delta \rightarrow 0^+$ , we have

$$L_{z,\alpha}(\varepsilon; f) \geq (0.6 - 2\alpha) \omega_z(\varepsilon/3; f).$$

Now we turn to the upper bounds. We introduce the following two lemmas, one for the minimum and another for the minimizer, that will be proved later.

**Lemma 2.6.2.** *For  $0 < \alpha \leq 0.3$  and any  $f \in \mathcal{F}$ ,*

$$R_m(\varepsilon; f) \leq A_m \rho_m(\varepsilon; f) \leq A_m \omega_m(\varepsilon; f), \tag{2.6.2}$$

$$L_{m,\alpha}(\varepsilon; f) \leq B_{m,\alpha} \rho_m(\varepsilon; f) \leq B_{m,\alpha} \omega_m(\varepsilon; f), \tag{2.6.3}$$

where  $A_m = 1.03$  and  $0 < B_{m,\alpha} \leq 3(1 - 2\alpha)z_\alpha$ .

**Lemma 2.6.3.** *For  $0 < \alpha \leq 0.3$  and any  $f \in \mathcal{F}$ ,*

$$R_z(\varepsilon; f) \leq A_z \rho_z(\varepsilon; f) \leq A_z \omega_z(\varepsilon; f), \tag{2.6.4}$$

$$L_{z,\alpha}(\varepsilon; f) \leq B_{z,\alpha} \rho_z(\varepsilon; f) \leq B_{z,\alpha} \omega_z(\varepsilon; f), \tag{2.6.5}$$

where  $A_z = 1.5$  and  $0 < B_{z,\alpha} \leq 3(1 - 2\alpha) \min\{z_\alpha, (2z_\alpha)^{2/3}\}$ .

The theorem follows as  $B_\alpha \geq \max\{B_{z,\alpha}, B_{m,\alpha}\}$  and  $A_1 \geq \max\{A_m, A_z\}$ .  $\square$

*Proof of Lemma 2.6.2.* For any function  $g \in \mathcal{F}$ , define  $f_\theta$  with  $\theta \in \{-1, 1\}$  and  $f_{-1} = f$  and  $f_1 = g$ . Recall that for  $W$  defined in (2.6.1),  $W \sim N(\theta \cdot \frac{\|f_1 - f_{-1}\|}{2\varepsilon}, 1)$ . Let

$$\hat{M} = \text{sign}(W) \cdot \frac{M(g) - M(f)}{2} + \frac{M(g) + M(f)}{2}.$$

Then  $\mathbb{E}_f(|\hat{M} - M(f)|) = |M(f) - M(g)|\Phi(-\frac{\|g-f\|}{2\varepsilon}) = \mathbb{E}_g(|\hat{M} - M(g)|)$ . Therefore,

$$\begin{aligned} R_m(\varepsilon; f) &\leq \sup_{g \in \mathcal{F}} |M(f) - M(g)|\Phi(-\frac{\|g-f\|}{2\varepsilon}) \stackrel{(i)}{\leq} \sup_{c>0} \omega_m(c\varepsilon; f)\Phi(-\frac{c}{2}) \\ &\stackrel{(ii)}{\leq} \max\{3\rho_m(\varepsilon; f) \sup_{0<c\leq 1} c^{\frac{2}{3}}\Phi(-\frac{c}{2}), \sup_{c\geq 1} \omega_m(c\varepsilon; f)\Phi(-\frac{c}{2})\} \\ &\stackrel{(iii)}{\leq} \max\{3\rho_m(\varepsilon; f)\Phi(-\frac{1}{2}), \sup_{c\geq 1} \omega_m(c\varepsilon; f)\Phi(-\frac{c}{2})\}, \end{aligned}$$

where (i) is due to the definition of  $\omega_m(c\varepsilon; f)$  in Equation (2.2.2), (ii) follows from Proposition 2.2.1, (iii) is due to the fact that  $c^{\frac{2}{3}}\Phi(-\frac{c}{2})$  increases in  $c \in [0, 1]$ . Furthermore we have,

$$\begin{aligned} \sup_{c\geq 1} \omega_m(c\varepsilon; f)\Phi(-\frac{c}{2}) &\stackrel{(iv)}{\leq} \sup_{c\geq 1} 3\rho_m(c\varepsilon; f)\Phi(-\frac{c}{2}) \stackrel{(v)}{\leq} 3\rho_m(\varepsilon; f) \cdot \sup_{c\geq 1} c\Phi(-\frac{c}{2}) \\ &\stackrel{(vi)}{\leq} 3\rho_m(\varepsilon; f) \times 0.3423 \stackrel{(vii)}{\leq} 1.03\omega_m(\varepsilon; f), \end{aligned}$$

where (iv) is due to Proposition 2.2.2, (v) and (vii) are due to Proposition 2.2.1, and (vi) is due to a bound for  $\sup_{c\geq 1} c\Phi(-\frac{c}{2})$ , which follows from the elementary inequalities:  $\Phi(-c/2) \leq \frac{1}{c}\sqrt{\frac{2}{\pi}}\exp(-\frac{c^2}{8})$  for  $c > 0$ ;  $\frac{\partial(c\Phi(-c/2))}{\partial c} = \Phi(-c/2) - \frac{c}{2}\sqrt{\frac{1}{2\pi}}\exp(-\frac{c^2}{8}) < 0$  for  $c > 2$ ; and  $\sup_{c \in [k/100, (k+1)/100]} c\Phi(-c/2) \leq 0.01(k+1)\Phi(-0.01 \times k/2)$  for  $k = \{100, 101, \dots, 200\}$ . Therefore, we can take  $A_m = \max\{3\Phi(-1/2), 1.03\} = 1.03$ .

For inference of the minimum, consider the following confidence interval:

$$CI_{m,\alpha} = \begin{cases} \{M(f)\} & W < -z_\alpha + \frac{\|f-g\|}{2\varepsilon} \\ \{M(g)\} & W \geq (z_\alpha - \frac{\|f-g\|}{2\varepsilon}) \vee (-z_\alpha + \frac{\|f-g\|}{2\varepsilon}) \\ [M(f) \wedge M(g), M(f) \vee M(g)] & \text{otherwise} \end{cases}.$$

Clearly, we have  $P_f(M(f) \notin CI_{m,\alpha}) \leq \alpha$  and  $P_g(M(g) \notin CI_{m,\alpha}) \leq \alpha$ . Note that for  $\theta \in \{0, 1\}$ ,

$$\begin{aligned} \mathbb{E}_{f_\theta} L(CI_{m,\alpha}) &\leq |M(f) - M(g)| P_{f_\theta}(-z_\alpha + 0.5 \frac{\|f-g\|}{\varepsilon} \leq W < z_\alpha - 0.5 \frac{\|f-g\|}{\varepsilon}) \\ &\leq |M(f) - M(g)| (\Phi(z_\alpha - \frac{\|f-g\|}{\varepsilon}) - \alpha)_+. \end{aligned}$$

Therefore, it follows from Proposition 2.2.1 that

$$\begin{aligned} L_{m,\alpha}(\varepsilon; f) &\leq \sup_{g \in \mathcal{F}} |M(f) - M(g)| (\Phi(z_\alpha - \frac{\|f-g\|}{\varepsilon}) - \alpha)_+ \\ &\leq \sup_{c>0} \omega_m(c\varepsilon; f) (\Phi(z_\alpha - c) - \alpha)_+ \\ &\leq \max\{\omega_m(\varepsilon; f) (\Phi(z_\alpha) - \alpha)_+, \sup_{c>1} \omega_m(c\varepsilon; f) (\Phi(z_\alpha - c) - \alpha)_+\} \\ &= \max\{\omega_m(\varepsilon; f) (1 - 2\alpha), \sup_{c>1} \omega_m(c\varepsilon; f) (\Phi(z_\alpha - c) - \alpha)_+\}. \end{aligned}$$

Further, recalling  $\alpha < 0.3$ , we have  $2z_\alpha > 1$ , thus

$$\begin{aligned} \sup_{c>1} \omega_m(c\varepsilon; f) (\Phi(z_\alpha - c) - \alpha)_+ &\leq \sup_{c>1} 3\rho_m(c\varepsilon; f) (\Phi(z_\alpha - c) - \alpha)_+ \\ &\leq 3\rho_m(\varepsilon; f) \sup_{c>1} c (\Phi(z_\alpha - c) - \alpha)_+ = 3\rho_m(\varepsilon; f) \sup_{2z_\alpha > c > 1} c (\Phi(z_\alpha - c) - \alpha) \\ &\stackrel{\text{(viii)}}{\leq} 3\rho_m(\varepsilon; f) [(1 - 2\alpha)z_\alpha \mathbb{1}\{z_\alpha \geq 1\} + (0.5 - \alpha) \cdot 2z_\alpha \mathbb{1}\{z_\alpha < 1\}] \\ &\leq 3\omega_m(\varepsilon; f) (1 - 2\alpha)z_\alpha, \end{aligned}$$

where (viii) follows from  $\sup_{c \in [A, B]} c(\Phi(z_\alpha - c) - \alpha) \leq B(\Phi(z_\alpha - A) - \alpha)$  for any  $1 \leq A \leq$

$B \leq 2z_\alpha$ . In conclusion,

$$L_{m,\alpha}(\varepsilon; f) \leq 3(1 - 2\alpha)z_\alpha\rho_m(\varepsilon; f) \leq 3(1 - 2\alpha)z_\alpha\omega_m(\varepsilon; f).$$

□

*Proof of Lemma 2.6.3.* For any  $g \in \mathcal{F}$ , consider  $f_\theta$  with  $\theta \in \{-1, 1\}$ ,  $f_{-1} = f$  and  $f_1 = g$ . Recall that for  $W$  defined in (2.6.1),  $W \sim N(\theta \cdot \frac{\|f_1 - f_{-1}\|}{2\varepsilon}, 1)$ . Let

$$\hat{Z} = \text{sign}(W) \cdot \frac{Z(g) - Z(f)}{2} + \frac{Z(g) + Z(f)}{2}.$$

Then  $\mathbb{E}_f(|\hat{Z} - Z(f)|) = |Z(f) - Z(g)|\Phi(-\frac{\|g - f\|}{2\varepsilon}) = \mathbb{E}_g(|\hat{Z} - Z(g)|)$ . Therefore,

$$\begin{aligned} R_z(\varepsilon; f) &\leq \sup_{g \in \mathcal{F}} |Z(f) - Z(g)|\Phi(-\frac{\|g - f\|}{2\varepsilon}) \leq \sup_{c > 0} \omega_z(c\varepsilon; f)\Phi(-\frac{c}{2}) \\ &\leq \max\{0.5\omega_z(\varepsilon; f), \sup_{c \geq 1} \omega_z(c\varepsilon; f)\Phi(-\frac{c}{2})\}. \end{aligned} \quad (2.6.6)$$

In addition,

$$\begin{aligned} \sup_{c \geq 1} \omega_z(c\varepsilon; f)\Phi(-\frac{c}{2}) &\leq \sup_{c \geq 1} 3\rho_z(c\varepsilon; f)\Phi(-\frac{c}{2}) \\ &\leq 3 \sup_{c \geq 1} \min\{c, (2c)^{\frac{2}{3}}\} \rho_z(\varepsilon; f)\Phi(-\frac{c}{2}) \leq 1.03\rho_z(\varepsilon; f). \end{aligned} \quad (2.6.7)$$

Inequalities (2.6.7) and (2.6.6) together with Proposition 2.2.1 show that we can take  $A_z = 1.5$ .

For inference of the minimizer, let

$$CI_{z,\alpha} = \begin{cases} \{Z(f)\} & W < -z_\alpha + 0.5\frac{\|f-g\|}{\varepsilon} \\ \{Z(g)\} & W \geq (z_\alpha - \frac{\|f-g\|}{2\varepsilon}) \vee (-z_\alpha + \frac{\|f-g\|}{2\varepsilon}) \\ [Z(f) \wedge Z(g), Z(f) \vee Z(g)] & \text{otherwise} \end{cases}$$



Clearly, we have  $P_f(Z(f) \notin CI_{z,\alpha}) \leq \alpha, P_g(Z(g) \notin CI_{z,\alpha}) \leq \alpha$ . For the expected length, similar to the proof for Lemma 2.6.2, we have for  $\theta \in \{-1, 1\}$ ,

$$\mathbb{E}_{f_\theta} L(CI_{z,\alpha}) \leq |Z(f) - Z(g)|(\Phi(z_\alpha - \frac{\|f - g\|}{\varepsilon}) - \alpha)_+. \quad (2.6.8)$$

Therefore

$$\begin{aligned} L_{z,\alpha}(\varepsilon; f) &\leq \sup_{g \in \mathcal{F}} |Z(f) - Z(g)|(\Phi(z_\alpha - \frac{\|f - g\|}{\varepsilon}) - \alpha)_+ \leq \sup_{c > 0} \omega_z(c\varepsilon; f)(\Phi(z_\alpha - c) - \alpha)_+ \\ &\leq \max\{\omega_z(\varepsilon; f)(\Phi(z_\alpha) - \alpha)_+, \sup_{c > 1} \omega_z(c\varepsilon; f)(\Phi(z_\alpha - c) - \alpha)_+\} \\ &\leq \max\{\omega_z(\varepsilon; f)(1 - 2\alpha), \sup_{c > 1} \omega_z(c\varepsilon; f)(\Phi(z_\alpha - c) - \alpha)_+\}. \end{aligned}$$

Note that  $0 < \alpha < 0.3$  implies  $2z_\alpha > 1$ . Hence

$$\begin{aligned} \sup_{c > 1} \omega_z(c\varepsilon; f)(\Phi(z_\alpha - c) - \alpha)_+ &\leq \sup_{c > 1} 3\rho_z(c\varepsilon; f)(\Phi(z_\alpha - c) - \alpha)_+ \\ &\leq 3\rho_z(\varepsilon; f) \sup_{c > 1} \min\{c, (2c)^{2/3}\}(\Phi(z_\alpha - c) - \alpha)_+ \\ &\leq 3\rho_z(\varepsilon; f) \max\{(1 - 2\alpha) \min\{z_\alpha, (2z_\alpha)^{2/3}\} \mathbb{1}\{z_\alpha \geq 1\}, (0.5 - \alpha) \min\{2z_\alpha, (4z_\alpha)^{2/3}\}\} \\ &\leq 3\rho_z(\varepsilon; f)(1 - 2\alpha) \min\{z_\alpha, (2z_\alpha)^{2/3}\} \\ &\leq 3\omega_z(\varepsilon; f)(1 - 2\alpha) \min\{z_\alpha, (2z_\alpha)^{2/3}\}. \end{aligned}$$

In conclusion,  $L_{z,\alpha}(\varepsilon; f) \leq 3(1 - 2\alpha) \min\{z_\alpha, (2z_\alpha)^{2/3}\} \omega_z(\varepsilon; f)$ . □

### 2.6.2. Proof of Theorem 2.2.2

It follows from Theorem 2.2.1 and Proposition 2.2.2 that

$$A_1^3 \omega_z(\varepsilon; f) \cdot \omega_m(\varepsilon; f)^2 \geq R_z(\varepsilon; f) \cdot R_m(\varepsilon; f)^2 \geq a_1^3 \omega_z(\varepsilon; f) \cdot \omega_m(\varepsilon; f)^2$$

and

$$\rho_z(\varepsilon; f) \cdot \rho_m(\varepsilon; f)^2 \leq \omega_z(\varepsilon; f) \cdot \omega_m(\varepsilon; f)^2 \leq 27\rho_z(\varepsilon; f) \cdot \rho_m(\varepsilon; f)^2.$$

Furthermore,

$$\frac{\varepsilon^2}{2} \leq \rho_z(\varepsilon; f) \cdot \rho_m(\varepsilon; f)^2 \leq 3\varepsilon^2. \quad (2.6.9)$$

This can be shown as follows. Let  $u = \rho_m(\varepsilon; f) + M(f)$  and define  $f_u(t) = \max\{f(t), u\}$  as in Section 2.2.1. Note that  $\|f - f_u\|_\infty \leq \rho_m(\varepsilon; f)$  and it follows from the definition of  $\rho_m(\varepsilon; f)$  that  $\|f - f_u\|_2 = \varepsilon$ . As illustrated in Figure 2.1 in Section 2.2.1 (with special attention to the rectangle ABCD and the triangle EDF),

$$\begin{aligned} 2\rho_z(\varepsilon; f) \cdot \rho_m(\varepsilon; f)^2 &\geq \int_0^1 (f(t) - f_u(t))^2 dt = \varepsilon^2 \\ &\geq \max \left\{ \int_0^{Z(f)} (f(t) - f_u(t))^2 dt, \int_{Z(f)}^1 (f(t) - f_u(t))^2 dt \right\} \geq \frac{1}{3} \rho_z(\varepsilon; f) \cdot \rho_m(\varepsilon; f)^2. \end{aligned}$$

To conclude, we have for any  $f \in \mathcal{F}$

$$274\varepsilon^2 > 81A_1^3\varepsilon^2 \geq R_z(\varepsilon; f) \cdot R_m(\varepsilon; f)^2 \geq \frac{a_1^3}{2}\varepsilon^2 \geq \frac{\Phi(-0.5)^3}{2}\varepsilon^2.$$

Similarly, we have

$$L_{z,\alpha}(\varepsilon; f) \cdot L_{m,\alpha}(\varepsilon; f)^2 \geq (0.6 - 2\alpha)^3 \cdot \omega_z\left(\frac{\varepsilon}{3}; f\right) \cdot \omega_m\left(\frac{\varepsilon}{3}; f\right)^2 \geq \frac{(0.6 - 2\alpha)^3}{18}\varepsilon^2,$$

and

$$L_{z,\alpha}(\varepsilon; f) \cdot L_{m,\alpha}(\varepsilon; f)^2 \leq B_\alpha^3 \omega_z(\varepsilon; f) \omega_m(\varepsilon; f)^2 \leq 3^7 \cdot (1 - 2\alpha)^3 \varepsilon^2. \quad \square$$

## CHAPTER 3

### Optimal Estimation and Inference for Minimizer and Minimum of Multivariate Additive Convex Functions

#### 3.1. Introduction

Chapter 2 establishes minimax rates for both estimation and inference for both minimizer and minimum under a non-asymptotic local minimax framework for univariate convex function.

In the present chapter, we consider optimal estimation and inference for the minimizer and minimum of *multivariate additive convex functions* under suitable non-asymptotic framework that can characterize the difficulty of the problem at individual functions.

We consider both white noise model and nonparametric regression. We first focus on the white noise model, which is given by

$$dY(\mathbf{t}) = \mathbf{f}(\mathbf{t})d\mathbf{t} + \varepsilon d\mathbf{W}(\mathbf{t}), \mathbf{t} \in [0, 1]^s, \quad (3.1.1)$$

where  $\mathbf{W}(\mathbf{t})$  is a standard  $(s, 1)$ -Brownian sheet on  $[0, 1]^s$ ,  $\varepsilon > 0$  is the noise level. The drift function  $\mathbf{f}$  is assume to be in  $\mathcal{F}_s$ , the collection of  $s$ -dimensional additive convex functions defined as follows. Function  $\mathbf{f}$  is said to be an additive convex function if it can be written in the following form:

$$\mathbf{f}(\mathbf{t}) = f_0 + \sum_{i=1}^s f_i(t_i), \mathbf{t} = (t_1, t_2, \dots, t_s) \in [0, 1]^s, \quad (3.1.2)$$

where  $f_0$  is a real number and for  $1 \leq i \leq s$ ,  $f_i$  is in  $\mathcal{F}$ , the collection of univariate convex functions with unique minimizer, and  $f_i$  also satisfies  $\int_0^1 f_i(t)dt = 0$ . Note that for any function  $\mathbf{f}$  that can be written in the aforementioned decomposition (3.1.2), the decomposition is unique. And for  $s = 1$ ,  $\mathcal{F}_s = \mathcal{F}$ . For clarity, we also write  $Y_{\mathbf{f}}$  for  $Y$  under

$\mathbf{f}$  to specify the true function. The goal is to optimally estimate the minimizer  $Z(\mathbf{f}) = \arg \min_{\mathbf{t} \in [0,1]^s} \mathbf{f}(\mathbf{t})$  and minimum  $M(\mathbf{f}) = \min_{\mathbf{t} \in [0,1]^s} \mathbf{f}(\mathbf{t})$  and also construct confidence hypercube for  $Z(\mathbf{f})$  and confidence interval for  $M(\mathbf{f})$ . Estimation and inference for the minimizer  $Z(\mathbf{f})$  and minimum  $M(\mathbf{f})$  under nonparametric setting will be discussed later in section 3.4.

### 3.1.1. Non-asymptotic Function-specific Benchmarks

The first step toward evaluating the performance of a procedure at individual convex functions in  $\mathcal{F}_s$  is to define function-specific benchmarks for estimation and inference for minimizer. For estimation and inference of minimum and estimation of minimizer, we investigate it under local minimax framework (Cai and Low, 2015), which is also used in estimation and inference for univariate convex functions in Chapter 2. For inference of minimizer, the same two-point local minimax framework is not as appropriate and we take a non-asymptotic function-specific benchmark that measures exactly the best behavior that any method can achieve.

For estimation of the minimizer, the hardness of the problem at an individual function is naturally captured by the expected squared distance. Further, under the local minimax framework, the benchmark is given by

$$R_z(\varepsilon; \mathbf{f}) = \sup_{\mathbf{g} \in \mathcal{F}_s} \inf_{\hat{Z}} \max_{\mathbf{h} \in \{\mathbf{f}, \mathbf{g}\}} \mathbb{E} \left( \|\hat{Z} - Z(\mathbf{h})\|^2 \right). \quad (3.1.3)$$

For any given  $\mathbf{f} \in \mathcal{F}_s$ , the benchmark  $R_z(\varepsilon; \mathbf{f})$  quantifies the estimation accuracy at  $\mathbf{f}$  of the minimizer  $Z(\mathbf{f})$  against the hardest alternative of  $\mathbf{f}$  within the function class  $\mathcal{F}_s$ .

For estimation of the minimum, the hardness of the problem at an individual function  $\mathbf{f}$  is naturally captured by the expected squared error. Further, under the local minimax framework, it is given by

$$R_m(\varepsilon; \mathbf{f}) = \sup_{\mathbf{g} \in \mathcal{F}_s} \inf_{\hat{M}} \max_{\mathbf{h} \in \{\mathbf{f}, \mathbf{g}\}} \mathbb{E}_{\mathbf{h}} \left( \|\hat{M} - M(\mathbf{h})\|^2 \right). \quad (3.1.4)$$

For any given function  $\mathbf{f} \in \mathcal{F}_s$ , benchmark  $R_m(\varepsilon; \mathbf{f})$  quantifies the estimation accuracy of the minimum  $M(\mathbf{f})$  at  $\mathbf{f}$  against the hardest alternative of  $\mathbf{f}$  within function class  $\mathcal{F}_s$ .

For estimation problems, we show that the benchmarks are valid good benchmarks in the sense that if it is significantly out performed at function  $\mathbf{f} \in \mathcal{F}_s$ , then a penalty need to be paid at another function  $\mathbf{f}_1 \in \mathcal{F}_s$ . We establish sharp minimax rates for these benchmarks and construct procedures attain the minimax rates, up to a constant factor depending on dimension  $s$ , simultaneously for all  $\mathbf{f} \in \mathcal{F}_s$ .

For confidence hyper cube of the minimizer with a pre-specified coverage, the hardness of the problem is naturally captured by the expected volume. Let  $\mathcal{I}_{z,\alpha}(\mathcal{S})$  be the collection of confidence hyper cubes for the minimizer  $Z(\mathbf{f})$  with guaranteed coverage probability  $1 - \alpha$  for all  $\mathbf{f} \in \mathcal{S}$ . The benchmark under a non-asymptotic function-specific framework, at  $\mathbf{f}$ , is given by the minimum expected volume at  $\mathbf{f}$  for all confidence hyper cube in  $\mathcal{I}_{z,\alpha}(\mathcal{F}_s)$ :

$$L_{\alpha,z}(\varepsilon; \mathbf{f}) = \inf_{CI_{z,\alpha} \in \mathcal{I}_{z,\alpha}(\mathcal{F}_s)} \mathbb{E}_{\mathbf{f}}(V(CI_{z,\alpha})), \quad (3.1.5)$$

where  $V(CI_{z,\alpha})$  is the volume of the confidence hyper cubes. Unlike local minimax framework, which measures the best a confidence hyper cube with the pre-specified probability coverage at  $\mathbf{f}$  and a hardest  $\mathbf{g} \in \mathcal{F}_s$  can achieve, this benchmark takes hyper cubes in  $\mathcal{I}_{z,\alpha}(\mathcal{F}_s)$  (i.e. it has pre-specified probability coverage for all  $\mathbf{g} \in \mathcal{F}_s$ ). It is easy to see that this benchmark depends on  $\mathbf{f}$  and is the best that any method can achieve at  $\mathbf{f}$ .

For confidence interval of the minimum with a pre-specified coverage, the hardness of the problem is naturally captured by the expected length. Let  $\mathcal{I}_{m,\alpha}(\mathcal{S})$  be the collection of confidence intervals for the minimum  $M(\mathbf{f})$  with guaranteed coverage probability  $1 - \alpha$  for all  $\mathbf{f} \in \mathcal{S}$ . Under the local minimax framework, the benchmark is given by

$$L_{\alpha,m}(\varepsilon; \mathbf{f}) = \sup_{\mathbf{g} \in \mathcal{F}_s} \inf_{CI_{m,\alpha} \in \mathcal{I}_{m,\alpha}(\{\mathbf{f}, \mathbf{g}\})} \mathbb{E}_{\mathbf{f}}(|CI_{m,\alpha}|), \quad (3.1.6)$$

### 3.1.2. Projection Representation and Optimal Procedures

Another major step in our analysis is developing data-driven and computationally efficient algorithms for the construction of estimators and confidence interval (hyper cube) as well as establishing the optimality of these procedures at each  $f \in \mathcal{F}$ .

An interesting observation is that  $Y_{\mathbf{f}}$  admits a *projection representation*,

$$\mathfrak{P}(Y_{\mathbf{f}}) = (\boldsymbol{\pi}_1(Y_{\mathbf{f}}), \dots, \boldsymbol{\pi}_s(Y_{\mathbf{f}}), \mathbf{er}(Y_{\mathbf{f}})),$$

such that  $\boldsymbol{\pi}_i(Y_{\mathbf{f}})$  is a sufficient statistic for  $f_i$  and all elements in  $\mathfrak{P}(Y_{\mathbf{f}})$  are independent. Also  $Y_{\mathbf{f}}$  can be fully recovered from  $\mathfrak{P}(Y_{\mathbf{f}})$ . The estimators and confidence interval (hyper cube) are constructed based on this observation by doing estimation and inference on each component and carefully join them together.

The key idea behind the construction for each component of the optimal procedures is to first iteratively localize the minimizer by comparing the integrals over relevant subintervals together with a very carefully constructed stopping rule controlled by a user-specified parameter, and then add an additional estimation/inference procedure. The final estimation/inference is to carefully choose the control parameter of the component-wise stopping rule and put together the output for each axis.

The resulting estimators,  $\hat{Z}$  for  $Z(\mathbf{f})$  and  $\hat{M}$  for  $M(\mathbf{f})$ , are shown to attain within a dimension-dependent constant of the benchmarks  $R_z(\varepsilon; \mathbf{f})$   $R_m(\varepsilon; \mathbf{f})$  simultaneously for all  $\mathbf{f} \in \mathcal{F}_s$ ,

$$\mathbb{E}_{\mathbf{f}} \left( \|\hat{Z} - Z(\mathbf{f})\|^2 \right) \leq C_{z,s} R_z(\varepsilon; \mathbf{f}), \quad (3.1.7)$$

$$\mathbb{E}_{\mathbf{f}} \left( \|\hat{M} - M(\mathbf{f})\|^2 \right) \leq C_{m,s} R_m(\varepsilon; \mathbf{f}), \quad (3.1.8)$$

for constants  $C_{z,s}$  and  $C_{m,s}$  depending on dimension  $s$  only.

The resulting confidence interval (hyper cube),  $CI_{z,\alpha}$  for  $Z(\mathbf{f})$  and  $CI_{m,\alpha}$  for  $M(\mathbf{f})$ , are shown to have the pre-specified coverage  $(1 - \alpha)$  while having expected length (volume) being adaptive to  $\mathbf{f}$  and attaining within a coverage-dimension-dependent constant of the benchmarks  $L_{\alpha,z}(\varepsilon; \mathbf{f}), L_{\alpha,m}(\varepsilon; \mathbf{f})$  for all  $\mathbf{f} \in \mathcal{F}_s$ . That is,

$$\mathbb{E}_{\mathbf{f}}(V(CI_{z,\alpha})) \leq C_{z,s,\alpha} L_{\alpha,z}(\varepsilon; \mathbf{f}), \quad (3.1.9)$$

$$\mathbb{E}_{\mathbf{f}}(|CI_{m,\alpha}|) \leq C_{m,s,\alpha} L_{\alpha,m}(\varepsilon; \mathbf{f}), \quad (3.1.10)$$

where  $C_{z,s,\alpha}$  and  $C_{m,s,\alpha}$  are constants depending on dimension  $s$  and  $\alpha$  only.

### 3.1.3. Organization of this Chapter

In Section 3.2, we analyze local minimax risks, relating them to appropriate local modulus of continuity, in turn providing rate-sharp upper and lower bounds. We also provide lower bound for the benchmark for inference of the minimizer in Section 3.2. In Section 3.3, we introduce projection representation of the observation, provide computationally efficient adaptive procedures and show their optimality. In Section 3.4, we consider the nonparametric regression model. We introduce the corresponding benchmarks, propose adaptive procedures and establish the optimality. Proofs are given in appendix Section A.5.

### 3.1.4. Notation

We conclude this section with some notation that will be used in the section. The cdf of the standard normal distribution is denoted by  $\Phi$ . For  $0 < \alpha < 1$ ,  $z_\alpha = \Phi^{-1}(1 - \alpha)$ . For  $\alpha = 0$ ,  $z_\alpha = \infty$ . We use  $\|\cdot\|$  to denote the  $L_2$  norm for vectors, univariate functions and multivariate functions, depending on the setting. We use  $\mathbb{1}\{A\}$  to denote indicator function that takes 1 when event  $A$  happens and 0 otherwise. We use bold symbols to denote multivariate functions, e.g.  $\mathbf{f}, \mathbf{g}, \mathbf{h}$ . We use  $f_1, \dots, f_s$  to denote the component functions for  $\mathbf{f}$  and  $f_0$  for constant part for  $\mathbf{f}$ , similar convention for  $\mathbf{g}, \mathbf{h}$ . Let  $a \wedge b = \min\{a, b\}$ ,  $a \vee b = \max\{a, b\}$  for real numbers  $a$  and  $b$ . We use  $Z(\cdot)$  to denote the minimizer operator, and  $M(\cdot)$  to denote

the minimum operator, for both  $\mathbf{f} \in \mathcal{F}_s$  and  $f \in \mathcal{F}$ . Note that we use  $\mathcal{I}_{z,\alpha}(\mathcal{S})$  to denote the collection of confidence hyper cubes for the minimizer with guaranteed coverage probability  $1 - \alpha$  for all functions in  $\mathcal{S}$ . This can be generalized into univariate case when  $\mathcal{S} \subset \mathcal{F}$  and the hyper cube becomes interval.

We use  $\mathcal{I}_{m,\alpha}(\mathcal{S})$  to denote the collection of confidence intervals for the minimum with guaranteed coverage probability  $1 - \alpha$  for all functions in  $\mathcal{S}$ . This can be generalized into univariate case when  $\mathcal{S} \subset \mathcal{F}$ .

### 3.2. Local Minimax Rates and Lower Bounds

In this section, we discuss the local minimax rates and the lower bound for inference of the minimizer. We introduce the local moduli of continuity and use it to characterize the benchmarks for estimation of minimizer and estimation and inference of minimum introduced in Section 3.1.1. We provide rate-sharp bounds for the continuity moduli based on geometry properties of the functions. As we use a different benchmark for inference of minimizer, we provide lower bound of it in this section.

#### 3.2.1. Local Modulus of Continuity.

For any given function  $\mathbf{f} \in \mathcal{F}_s$ , we define the following local moduli of continuity for the minimizer and minimum.

$$\omega_z(\varepsilon; \mathbf{f}) = \sup\{\|Z(\mathbf{f}) - Z(\mathbf{g})\|^2 : \|\mathbf{f} - \mathbf{g}\|_2 \leq \varepsilon, \mathbf{g} \in \mathcal{F}_s\} \quad (3.2.1)$$

$$\omega_m(\varepsilon; \mathbf{f}) = \sup\{\|M(\mathbf{f}) - M(\mathbf{g})\|^2 : \|\mathbf{f} - \mathbf{g}\|_2 \leq \varepsilon, \mathbf{f} \in \mathcal{F}_s\}, \quad (3.2.2)$$

$$\tilde{\omega}_m(\varepsilon; \mathbf{f}) = \sup\{\|M(\mathbf{f}) - M(\mathbf{g})\| : \|\mathbf{f} - \mathbf{g}\|_2 \leq \varepsilon, \mathbf{f} \in \mathcal{F}_s\}. \quad (3.2.3)$$

As in the case of linear functionals or in the case of minimizer and minimum operators for univariate convex functions, the local moduli  $\omega_z(\varepsilon; \mathbf{f})$ ,  $\omega_m(\varepsilon; \mathbf{f})$ ,  $\tilde{\omega}_m(\varepsilon; \mathbf{f})$  clearly depends on  $\mathbf{f}$  and can be regarded as an analogue of inverse Fisher Information in regular parametric



model.

The following theorem characterizes the benchmarks for estimation and inference in terms of the corresponding local moduli of continuity.

**Theorem 3.2.1** (Sharp Lower Bounds). *Let  $R_z(\varepsilon; \mathbf{f})$  be defined in (3.1.3),  $R_m(\varepsilon; \mathbf{f})$  be defined in (3.1.4), and  $L_{\alpha,m}(\varepsilon; \mathbf{f})$  be defined in (3.1.6). Let  $0 < \alpha \leq 0.1$ . Then*

$$a\omega_z(\varepsilon; \mathbf{f}) \leq R_z(\varepsilon; \mathbf{f}) \leq A\omega_z(\varepsilon; \mathbf{f}), \quad (3.2.4)$$

$$a\omega_m(\varepsilon; \mathbf{f}) \leq R_m(\varepsilon; \mathbf{f}) \leq A\omega_m(\varepsilon; \mathbf{f}) \quad (3.2.5)$$

$$b_\alpha \tilde{\omega}_m(\varepsilon; \mathbf{f}) \leq L_{\alpha,m}(\varepsilon; \mathbf{f}) \leq B_\alpha \tilde{\omega}_m(\varepsilon; \mathbf{f}) \quad (3.2.6)$$

where the constants  $a, A, b_\alpha, B_\alpha$  can be taken as  $a = 0.1$ ,  $A = 3.1$ ,  $b_\alpha = 0.6 - \alpha$ , and  $B_\alpha = 2(1 - 2\alpha)z_\alpha$ .

Theorem 3.2.1 shows that the benchmarks can be characterized in terms of continuity moduli of continuity. However, this continuity moduli is hard to compute. We now recollect two related geometry quantities to facilitate bounding the continuity moduli used in univariate case in Chapter 2. For  $f \in \mathcal{F}$ ,  $u \in \mathbb{R}$  and  $\varepsilon > 0$ , let  $f_u(t) = \max\{f(t), u\}$ ,  $M(f) = \min_{x \in [0,1]} f(x)$ , and define

$$\rho_m(\varepsilon; f) = \sup\{u - \min\{f(x) : x \in [0, 1]\} : \|f - f_u\| \leq \varepsilon\}, \quad (3.2.7)$$

$$\rho_z(\varepsilon; f) = \sup\{|t - Z(f)| : f(t) \leq \rho_m(\varepsilon; f) + M(f), t \in [0, 1]\}. \quad (3.2.8)$$

With the geometric quantity  $\rho_z(\varepsilon; f)$ , we can establish a rate-sharp bound of modulus of continuity for the minimizer.

**Theorem 3.2.2** (Geometry Representation for Modulus of Continuity for Minimzer). *Let  $\rho_z(\varepsilon; f)$  be defined in (3.2.8) for  $f \in \mathcal{F}$ , and let  $\mathbf{f} \in \mathcal{F}_s$ . Let  $\omega_z(\varepsilon; \mathbf{f})$  be defined in (3.2.1). Then*

$$\frac{1}{3}s^{-\frac{2}{3}} \sum_{i=1}^s \rho_z(\varepsilon; f_i)^2 \leq \omega_z(\varepsilon; \mathbf{f}) \leq \sum_{i=1}^s 9\rho_z(\varepsilon; f_i)^2. \quad (3.2.9)$$

And for any  $\beta \leq s$ , there exists  $\mathbf{f} \in \mathcal{F}_s$  such that  $\sum_{i=1}^s \rho_z(\varepsilon; f_i)^2 = \beta$  and

$$\omega_z(\varepsilon; \mathbf{f}) \leq 9s^{-\frac{2}{3}} \sum_{i=1}^s \rho_z(\varepsilon; f_i)^2. \quad (3.2.10)$$

And for any  $\beta \leq s$ , and  $\delta_0 > 0$ , there exists  $\mathbf{f} \in \mathcal{F}_s$  such that  $\sum_{i=1}^s \rho_z(\varepsilon; f_i)^2 = \beta$  and

$$\omega_z(\varepsilon; \mathbf{f}) \geq \rho_z(\varepsilon; f_i)^2 - \delta_0. \quad (3.2.11)$$

Theorem 3.2.2 shows that the modulus of continuity for minimizer varies within an absolute constant multiple times of

$$s^{-\frac{2}{3}} \sum_{i=1}^s \rho_z(\varepsilon; f_i)^2 \text{ and } \sum_{i=1}^s \rho_z(\varepsilon; f_i)^2,$$

with the order of both upper and lower bound attainable for some  $\mathbf{f} \in \mathcal{F}_s$ .

With the geometric quantity  $\rho_z(\varepsilon; f)$  and  $\rho_m(\varepsilon; f)$ , we can establish a rate-sharp bound of moduli of continuity for the minimum.

**Theorem 3.2.3** (Geometry Representation for Modulus of Continuity for Minimum). *Let  $\rho_z(\varepsilon; \mathbf{f})$  be defined in (3.2.8) and  $\rho_m(\varepsilon; \mathbf{f})$  be defined in (3.2.7) for  $f \in \mathcal{F}$ . Let  $\omega_m(\varepsilon; \mathbf{f})$  be defined in (3.2.2) and  $\tilde{\omega}_m(\varepsilon; \mathbf{f})$  be defined in (3.2.3) for  $\mathbf{f} \in \mathcal{F}_s$ . Then*

$$\frac{1}{1 + \sum_{i=1}^s (1 \wedge 2\rho_z(\varepsilon; f_i))} \sum_{i=1}^s \rho_m(\varepsilon; f_i)^2 \leq \omega_m(\varepsilon; \mathbf{f}) \leq 9(1 + \frac{1}{s}) \sum_{i=1}^s \rho_m(\varepsilon; f_i)^2, \quad (3.2.12)$$

$$\sqrt{\frac{1}{1 + \sum_{i=1}^s (1 \wedge 2\rho_z(\varepsilon; f_i))} \sum_{i=1}^s \rho_m(\varepsilon; f_i)^2} \leq \tilde{\omega}_m(\varepsilon; \mathbf{f}) \leq \sqrt{9(1 + \frac{1}{s}) \sum_{i=1}^s \rho_m(\varepsilon; f_i)^2}. \quad (3.2.13)$$

Theorem 3.2.3 shows that the modulus of continuity for minimum  $\omega_m(\varepsilon; \mathbf{f})$  is of the order  $\sum_{k=1}^s \rho_m(\varepsilon; f_k)^2$  and  $\tilde{\omega}_m(\varepsilon; \mathbf{f})$  is of the order  $\sqrt{\sum_{k=1}^s \rho_m(\varepsilon; f_k)^2}$ .

Now we have done establishing the local minimax rates for three tasks, we turn to estab-

lishing the lower bound for the benchmark of inference of the minimizer.

**Theorem 3.2.4** (Lower Bound for Expected Volume of Confidence Hyper Cube for Minimizer). *Let  $L_{\alpha,z}(\varepsilon; \mathbf{f})$  be defined in (3.1.5) for  $\mathbf{f} \in \mathcal{F}_s$  and  $\rho_z(\varepsilon; f)$  be defined in (3.2.8) for  $f \in \mathcal{F}$ . Then we have*

$$L_{\alpha,z}(\varepsilon; \mathbf{f}) \geq C_{\alpha,s} \Pi_{i=1}^s \rho_z(\varepsilon; f_i), \quad (3.2.14)$$

where  $C_{\alpha,s}$  is a positive constant depending on  $\alpha$  and  $s$ .

### 3.2.2. Penalty for Super-efficiency

We have shown that the estimation benchmarks  $R_z(\varepsilon; \mathbf{f})$  and  $R_m(\varepsilon; \mathbf{f})$  can be characterized by intrinsic geometric quantities of  $\mathbf{f}$ . Now we show that these benchmarks can not be essentially uniformly out performed. That is, if the benchmark is significantly out performed at function  $\mathbf{f} \in \mathcal{F}_s$ , then it needs to pay a penalty at another function  $\mathbf{f}_1 \in \mathcal{F}_s$ . These benchmarks, similar to that in the univariate case, play a role analogous to the information lower bound in the classic statistic.

**Theorem 3.2.5** (Penalty for Supper-Efficiency). *For any estimator of the minimizer  $\hat{Z}$ , if  $\mathbb{E}_{\mathbf{f}} \left( \|\hat{Z} - Z(\mathbf{f})\|^2 \right) \leq \gamma R_z(\varepsilon; \mathbf{f})$  for  $\mathbf{f} \in \mathcal{F}_s$  and  $\gamma < \gamma_0$ , where  $\gamma_0$  is a positive constant, then there exists  $\mathbf{f}_1 \in \mathcal{F}_s$  such that*

$$\mathbb{E}_{\mathbf{f}_1} \left( \|\hat{Z} - Z(\mathbf{f}_1)\|^2 \right) \geq c_{z,s} \left( \log \frac{1}{\gamma} \right)^{\frac{2}{3}} R_z(\varepsilon; \mathbf{f}_1), \quad (3.2.15)$$

where  $c_{z,s}$  is a constant depending on  $s$  only.

Similarly, for any estimator of the minimum  $\hat{M}$ , if  $\mathbb{E}_{\mathbf{f}} (|\hat{M} - M(\mathbf{f})|^2) \leq \gamma R_m(\varepsilon; \mathbf{f})$  for  $\mathbf{f} \in \mathcal{F}_s$  and  $\gamma < \gamma_0/s$ , where  $\gamma_0$  is a positive constant, then there exists  $\mathbf{f}_1 \in \mathcal{F}_s$  such that

$$\mathbb{E}_{\mathbf{f}_1} \left( |\hat{M} - M(\mathbf{f}_1)|^2 \right) \geq c_{m,s} \left( \log \frac{1}{\gamma} \right)^{\frac{2}{3}} R_m(\varepsilon; \mathbf{f}_1), \quad (3.2.16)$$

where  $c_{m,s}$  is a constant depending on  $s$  only.

### 3.3. Projection Representation and Adaptive Optimal Procedures.

We now turn to the construction of data-driven and computationally efficient algorithms for estimation and inference of minimizer and minimum for white noise model. Our construction is based on an information-preserving representation of the observation  $Y_{\mathbf{f}}$ , which we call *Projection Representation*. We show that our procedures achieve, up to a universal constant depending on dimension  $s$  and confidence level  $1-\alpha$ , the corresponding benchmarks  $R_z(\varepsilon; \mathbf{f})$ ,  $R_m(\varepsilon; \mathbf{f})$ ,  $L_{\alpha,z}(\varepsilon; \mathbf{f})$ ,  $L_{\alpha,m}(\varepsilon; \mathbf{f})$ , simultaneously for all  $\mathbf{f} \in \mathcal{F}_s$ .

#### 3.3.1. Projection Representation.

The construction of the procedures is based on an interesting property of the observation  $Y_{\mathbf{f}}$  (or  $Y$ ) that  $Y$  admits a nice information-preserving *projection representation*, which maps  $Y$  to an  $s+1$ -tuple, where first  $s$  elements can roughly be considered as a projection of the original stochastic process on each coordinate, and the last element is an  $s$ -dimensional stochastic process that can be considered as a remaining error.

**Definition 3.3.1** (Projection Representation). *For each  $1 \leq i \leq s$ , the  $i$ -th projection of  $Y$ ,  $\pi_i(Y)$ , is a univariate stochastic process that satisfies for  $0 \leq a_i < A_i \leq 1$ ,*

$$\int_{[a_i, A_i]} d\pi_i(Y) = \int_{t_i \in [a_i, A_i], \mathbf{t}_{-i} \in [0, 1]^{s-1}} dY - (A_i - a_i) \int_{[0, 1]^s} dY, \quad (3.3.1)$$

where  $\mathbf{t}_{-i} = \{t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_s\}$ .

$\mathbf{er}(Y)$  is a stochastic process on  $[0, 1]^s$ , such that for  $\mathcal{A} = [a_1, A_1] \times [a_2, A_2] \times \dots \times [a_s, A_s] \subset [0, 1]^s$ , we have

$$\int_{\mathcal{A}} d\mathbf{er}(Y) = \int_{\mathcal{A}} dY - \sum_{i=1}^s \Pi_{j \neq i}(A_j - a_j) \int_{a_i}^{A_i} d\pi_i(Y). \quad (3.3.2)$$

The projection representation mapping  $\mathfrak{P}(\cdot)$  of  $Y$  is

$$\mathfrak{P}(Y) = (\pi_1(Y), \pi_2(Y), \dots, \pi_s(Y), \mathbf{er}(Y)). \quad (3.3.3)$$

The reasons we call it a *projection representation* mapping are that  $\mathfrak{P}(Y)$  preserves all information of  $Y$ , that  $\mathfrak{P}(Y)$  has all of its elements, the projections and error, being mutually independent, and that its first  $s$  elements are sufficient statistics for corresponding component function  $f_i$ . More specifically, we have Proposition 3.3.1 summarizing the properties of projection representation.

**Proposition 3.3.1** (Property of Projection Representation). *Let  $\mathfrak{P}(\cdot)$  be defined as in equation (3.3.3). Denote the class of stochastic process defined in (3.1.1) as  $\mathfrak{Y}$ . Then we have the followings.*

- $\mathfrak{P}(\cdot)$  is a bijection from  $\mathfrak{Y}$  to  $\mathfrak{P}(\mathfrak{Y})$ .
- $\mathfrak{P}(Y)$  has all elements being independent.
- $\pi_i(Y)$  is a sufficient statistic for  $f_i$ , for  $i \in \{1, 2, \dots, s\}$ .

Also, it's easy to check that  $\mathbf{er}(Y)$  only depends on  $f_0$ , thus not carrying information for  $Z(\mathbf{f})$  by itself. Instead, it carries part of the information of  $M(\mathbf{f})$ . Note that the minimizer  $Z(\mathbf{f})$  can be written as  $Z(\mathbf{f}) = (Z(f_1), Z(f_2), \dots, Z(f_s))$ , so its  $i$ -th element only depends on  $f_i$ . Similarly  $M(\mathbf{f})$  can be written as  $M(\mathbf{f}) = f_0 + \sum_{k=1}^s M(f_k)$ , so each component in  $\mathfrak{P}(Y_{\mathbf{f}})$  serves as a sufficient statistics for each of the adding components of  $M(\mathbf{f})$ . The information preserving representation  $\mathfrak{P}(\cdot)$  plays the role of separating the relevant information of  $s$  coordinates into independent random variables.

### 3.3.2. Adaptive Procedures.

Now we are ready to introduce the construction of data-driven and computationally efficient algorithms for estimation and confidence interval (hyper cube) for the minimum  $M(\mathbf{f})$  and

the minimizer  $Z(\mathbf{f})$  under the white noise model in this section. The procedures constructed in this section are shown in Section 3.3.3 to be adaptive to each individual function  $\mathbf{f} \in \mathcal{F}_s$  in the sense that they simultaneously achieve, up to a universal constant depending on dimension  $s$  and confidence level  $1 - \alpha$ , the corresponding benchmarks, simultaneously for all  $\mathbf{f} \in \mathcal{F}_s$ .

Similar to the construction in Chapter 2, we have three blocks: localization, stopping, and estimation/inference. But since  $\pi_i(Y)$  has different distribution with that in the univariate case, and we also need to account for the dimension, our procedures are carefully tailored to accommodate for the new challenges.

### Sample Splitting

For technical reasons, we split the first  $s$  coordinates of the projection representation (i.e.  $\mathfrak{P}(Y)$ ),  $V = (\pi_1(Y), \pi_2(Y), \dots, \pi_s(Y))$ , into three independent pieces to ensure independence of the data used in the three steps.

Let  $B_1^1(t), B_1^2(t), B_2^1(t), B_2^2(t), \dots, B_s^1(t), B_s^2(t)$  be  $2s$  independent standard Brownian motions that are also independent from  $Y$ . Let data vectors  $V_l = (\mathbf{v}_1^l, \mathbf{v}_2^l, \dots, \mathbf{v}_s^l)$ ,  $V_r = (\mathbf{v}_1^r, \mathbf{v}_2^r, \dots, \mathbf{v}_s^r)$  and  $V_e = (\mathbf{v}_1^e, \mathbf{v}_2^e, \dots, \mathbf{v}_s^e)$  be defined as follows.

$$\begin{aligned} \mathbf{v}_i^l(t) &= \pi_i(Y)(t) + \frac{\sqrt{2}}{2}\varepsilon \left( B_i^1(t) - t \int_0^1 B_i^1(x) dx \right) + \frac{\sqrt{6}}{2}\varepsilon \left( B_i^2(t) - t \int_0^1 B_i^2(x) dx \right), \\ \mathbf{v}_i^r(t) &= \pi_i(Y)(t) + \frac{\sqrt{2}}{2}\varepsilon \left( B_i^1(t) - t \int_0^1 B_i^1(x) dx \right) - \frac{\sqrt{6}}{2}\varepsilon \left( B_i^2(t) - t \int_0^1 B_i^2(x) dx \right), \\ \mathbf{v}_i^e(t) &= \pi_i(Y)(t) - \sqrt{2}\varepsilon \left( B_i^1(t) - t \int_0^1 B_i^1(x) dx \right). \end{aligned} \tag{3.3.4}$$

Then the concatenate vector of vectors  $V_l, V_r, V_e$  has all of its  $3s$  elements being independent,

and for each axis  $i \in \{1, 2, \dots, s\}$ ,  $\mathbf{v}_i^l(t), \mathbf{v}_i^r(t), \mathbf{v}_i^e(t)$  can be written as

$$\begin{aligned} d\mathbf{v}_i^l(t) &= f_i(t)dt + \sqrt{3}\varepsilon d\tilde{W}_i^l, \\ d\mathbf{v}_i^r(t) &= f_i(t)dt + \sqrt{3}\varepsilon d\tilde{W}_i^r, \\ d\mathbf{v}_i^e(t) &= f_i(t)dt + \sqrt{3}\varepsilon d\tilde{W}_i^e, \end{aligned} \tag{3.3.5}$$

where  $\tilde{W}_i^l, \tilde{W}_i^r, \tilde{W}_i^e$  are independent standard Brownian Bridges.

### Localization

We use  $V_l$  for localization step, and for each axis  $k \in \{1, 2, \dots, s\}$ , localization is based on  $\mathbf{v}_k^l$ .

We take an iterative localization procedure similar to that in Chapter 2 on  $\mathbf{v}_k^l$ . For iterations (levels)  $j = 0, 1, \dots$ , and possible location index at  $j$ th level  $i = 0, 1, \dots, 2^j$ , we denote the sub-interval length, sub-interval end points, and the index of the sub-interval containing the minimizer at level  $j$  to be

$$m_j = 2^{-j}, \quad t_{j,i} = i \cdot m_j, \quad \text{and} \quad i_{j,k}^* = \max\{i : Z(f_k) \in [t_{j,i-1}, t_{j,i}]\}. \tag{3.3.6}$$

For  $j = 0, 1, \dots$ , and  $i = 1, 2, \dots, 2^j$ , define

$$X_{j,i,k} = \int_{t_{j,i-1}}^{t_{j,i}} d\mathbf{v}_k^l(t),$$

where  $\mathbf{v}_k^l$  is one of the three independent copies constructed above through sample splitting.

For convenience, we define  $X_{j,i,k} = +\infty$  for  $j = 0, 1, \dots$ , and  $i \in \mathbb{Z} \setminus \{1, 2, \dots, 2^j\}$ .

Let  $\hat{i}_{0,k} = 1$  and for  $j = 1, 2, \dots$ , let

$$\hat{i}_{j,k} = \arg \min_{2\hat{i}_{j-1}-2 \leq i \leq 2\hat{i}_{j-1}+1} X_{j,i,k}.$$

Note that given the value of  $\hat{i}_{j-1,k}$  at level  $j-1$ , in the next iteration the procedure halves the interval  $[t_{\hat{i}_{j-1,k}-1}, t_{\hat{i}_{j-1,k}}]$  into two subintervals and selects the interval  $[t_{\hat{i}_{j,k}-1}, t_{\hat{i}_{j,k}}]$  at level  $j$  from these and their immediate neighboring subintervals. So  $i$  only ranges over 4 possible values at level  $j$ .

### Stopping Rule

For each axis, it is necessary to have a stopping rule to select a final subinterval constructed in the localization iterations and carry out the estimation/inference based on that. But unlike a unified stopping rule in univariate case, we construct a series of stopping rules based on a user select parameter  $\zeta > 0$ , which we will specify later in the specific estimation/inference procedures. Again, for any  $1 \leq k \leq s$ , we focus on the stopping rules for  $k$ -th axis.

We use another independent copy  $\mathbf{v}_k^r$  constructed in the sample splitting step to devise the stopping rules. For  $j = 0, 1, \dots$ , and  $i = 1, 2, \dots, 2^j$ , let

$$\tilde{X}_{j,i,k} = \int_{t_{j,i-1}}^{t_{j,i}} d\mathbf{v}_k^r(t).$$

Again, for convenience, we define  $\tilde{X}_{j,i,k} = +\infty$  for  $j = 0, 1, \dots$ , and  $i \in \mathbb{Z} \setminus \{1, 2, \dots, 2^j\}$ .

Let the statistic  $T_{j,k}$  be defined as

$$T_{j,k} = \min\{\tilde{X}_{j,\hat{i}_{j,k}+6,k} - \tilde{X}_{j,\hat{i}_{j,k}+5,k}, \tilde{X}_{j,\hat{i}_{j,k}-6,k} - \tilde{X}_{j,\hat{i}_{j,k}-5,k}\},$$

where we use the convention  $+\infty - x = +\infty$  and  $\min\{+\infty, x\} = x$ , for any  $-\infty \leq x \leq \infty$ .

The stopping rule indexed by the parameter  $\zeta > 0$  is based on the value of  $T_{j,k}$ . Before we formally go into the stopping rule, it's helpful to look at the distribution of the elements defining  $T_{j,k}$ . Let  $\sigma_j^2 = 6m_j\varepsilon^2$ , some calculations show that when  $\tilde{X}_{j,\hat{i}_{j,k}+6,k} - \tilde{X}_{j,\hat{i}_{j,k}+5,k} <$



$\infty$ , we have

$$\left. \frac{\tilde{X}_{j, \hat{i}_j, k+6, k} - \tilde{X}_{j, \hat{i}_j, k+5, k}}{\sigma_j} \right|_{\hat{i}_j, k} \sim N \left( \frac{m_j \sqrt{m_j}}{\sqrt{6}\varepsilon} \times \frac{1}{m_j} \int_{t_{j, \hat{i}_j+5, k}}^{t_{j, \hat{i}_j+6, k}} \frac{f_k(t+m_j) - f_k(t)}{m_j} dt, 1 \right). \quad (3.3.7)$$

Note that the term

$$S_p(j, k) = \frac{1}{m_j} \int_{t_{j, \hat{i}_j+5, k}}^{t_{j, \hat{i}_j+6, k}} \frac{f_k(t+m_j) - f_k(t)}{m_j} dt$$

can be interpreted as an average slope across the interval  $[t_{j, \hat{i}_j+5, k}, t_{j, \hat{i}_j+6, k}]$  of the line determined by two points  $(t, f(t))$  and  $(t+m_j, f(t+m_j))$ . Basic property of convex function shows that  $S_p(j, k)$  is non-increasing with the increasing of  $j$ , and that  $S_p(j, k) < 0$  implies  $i_{j, k}^* \geq \hat{i}_j + 5$ . These mean that a small number of  $\frac{\tilde{X}_{j, \hat{i}_j, k+6, k} - \tilde{X}_{j, \hat{i}_j, k+5, k}}{\sigma_j}$  indicates either localization procedure's choice of a far away sub-interval from the one minimizer lies in or a negligible signal which implies little or no gain in continuing the localization procedure.

Analogous results hold for  $\frac{\tilde{X}_{j, \hat{i}_j, k-6, k} - \tilde{X}_{j, \hat{i}_j, k-5, k}}{\sigma_j}$ .

Finally, the iteration stops at level  $\hat{j}(\zeta, k)$ , where

$$\hat{j}(\zeta, k) = \min\{j : \frac{T_{j, k}}{\sigma_j} \leq z_\zeta\}. \quad (3.3.8)$$

The subinterval containing the minimizer  $Z(f_k)$  is localized to be

$$[t_{\hat{j}(\zeta, k), \hat{i}_{\hat{j}(\zeta, k), k}-1}, t_{\hat{j}(\zeta, k), \hat{i}_{\hat{j}(\zeta, k), k}}].$$

## Estimation and Inference

After obtaining, for each axis  $k \in \{1, 2, \dots, s\}$ , a stopping step  $\hat{j}(\zeta_k, k)$ , an associated index at the stopping step  $\hat{i}_{\hat{j}(\zeta_k, k), k}$ , and a final interval  $[t_{\hat{j}(\zeta_k, k), \hat{i}_{\hat{j}(\zeta_k, k), k}-1}, t_{\hat{j}(\zeta_k, k), \hat{i}_{\hat{j}(\zeta_k, k), k}}]$ ,

all controlled by a parameter  $\zeta_k > 0$ , we use them to construct estimator and confidence interval (hyper cube) for the minimum  $M(\mathbf{f})$  and the minimizer  $Z(\mathbf{f})$ .

For estimation of the minimizer, we set  $\zeta_k = \zeta = \Phi(-2)$ , for  $k \in \{1, 2, \dots, s\}$ . The  $k$ -th axis of the estimator  $\hat{Z}$  is given by the mid point of final interval:

$$\hat{Z}_k = \frac{t_{\hat{j}(\zeta, k), \hat{i}_{\hat{j}(\zeta, k), k-1}} + t_{\hat{j}(\zeta, k), \hat{i}_{\hat{j}(\zeta, k), k}}}{2}. \quad (3.3.9)$$

The final estimator  $\hat{Z}$  is given by

$$\hat{Z} = (\hat{Z}_1, \hat{Z}_2, \dots, \hat{Z}_s), \quad (3.3.10)$$

with  $\hat{Z}_k$  defined in (3.3.9).

For inference of the minimizer, we set  $\zeta_k = \zeta = \alpha/s$ , for  $k \in \{1, 2, \dots, s\}$ . The  $k$ -th axis  $CI_k$  of the hyper cube  $CI_{z, \alpha}$  is given by

$$CI_k = \left[ 2^{-\hat{j}(\zeta, k)+1} \left( \hat{i}_{\hat{j}(\zeta, k)-1, k} - 7 \right), 2^{-\hat{j}(\zeta, k)+1} \left( \hat{i}_{\hat{j}(\zeta, k)-1, k} + 6 \right) \right] \cap [0, 1]. \quad (3.3.11)$$

The confidence cube  $CI$  for the minimizer is give by

$$CI_{z, \alpha} = CI_1 \times CI_2 \times \dots \times CI_s, \quad (3.3.12)$$

where  $CI_k$  is defined in (3.3.11).

For estimation and inference of the minimum, let

$$\bar{X}_{j, i, k} = \int_{t_{j, i-1}}^{t_{j, i}} d\mathbf{v}_k^e(t),$$

for  $1 \leq i \leq 2^j$ , and  $+\infty$  for  $i \notin \{1, 2, \dots, 2^j\}$ .

For estimation of the minimum  $M(\mathbf{f})$ , let  $\zeta_k = \zeta = \Phi(-2)$  for  $k = 1, 2, \dots, s$ . Let the *final*

index for estimator construction for  $k$ -th coordinate be

$$i_{F,k} = \hat{i}_{\hat{j}(\zeta,k)-1,k} + 2 \left( \mathbb{1}\{\tilde{X}_{\hat{j}(\zeta,k),\hat{i}_{\hat{j}(\zeta,k)}+6,k} - \tilde{X}_{\hat{j}(\zeta,k),\hat{i}_{\hat{j}(\zeta,k)}+5,k} \leq 2\sigma_{\hat{j}(\zeta,k)}\} \right. \\ \left. - \mathbb{1}\{\tilde{X}_{\hat{j}(\zeta,k),\hat{i}_{\hat{j}(\zeta,k)}-6,k} - \tilde{X}_{\hat{j}(\zeta,k),\hat{i}_{\hat{j}(\zeta,k)}-5,k} \leq 2\sigma_{\hat{j}(\zeta,k)}\} \right). \quad (3.3.13)$$

The estimator of the minimum is given by

$$\hat{M} = Y(1, 1, \dots, 1) - Y(0, 0, \dots, 0) + \sum_{k=1}^s 2^{\hat{j}(\zeta,k)} \bar{X}_{\hat{j}(\zeta,k), i_{F,k}, k}. \quad (3.3.14)$$

For inference of the minimum, let  $\zeta_k = \zeta = \alpha/4s$  for  $k = 1, 2, \dots, s$ . Define an intermediate estimator of the minimum by

$$\hat{\mathbf{f}}_1 = Y(1, 1, \dots, 1) - Y(0, 0, \dots, 0) + \sum_{k=1}^s 2^{\hat{j}(\zeta,k)+3} \min_{16(\hat{i}_{\hat{j}(\zeta,k),k-1}-7) < i \leq 16(\hat{i}_{\hat{j}(\zeta,k),k-1}+6)} \bar{X}_{\hat{j}(\zeta,k)+2,i,k}. \quad (3.3.15)$$

Let  $U_n$  be the cumulative distribution function of  $\tilde{u} = \max\{u_1, \dots, u_n\}$ , where

$$u_1, \dots, u_n \stackrel{i.i.d}{\sim} N(0, 1),$$

and define

$$S_{n,\beta} = U_n^{-1}(1 - \beta). \quad (3.3.16)$$

In other words,  $S_{n,\beta}$  is the  $(1 - \beta)$  quantile of the distribution of the maximum of  $n$  *i.i.d.* standard normal variables.

Let

$$\mathbf{f}_{hi} = \hat{\mathbf{f}}_1 + S_{208,\alpha/8s} \times \sqrt{3}\varepsilon \sum_{k=1}^s 2^{\frac{\hat{j}(\zeta,k)+3}{2}} + z_{\alpha/8} \sqrt{3}\varepsilon s \\ \mathbf{f}_{lo} = \hat{\mathbf{f}}_1 - z_{\alpha/4} \sqrt{3}\varepsilon \sqrt{1 + \sum_{k=1}^s 2^{\hat{j}(\zeta,k)+3}} - \sum_{k=1}^s z_{\alpha/4s} \sqrt{3} \cdot 2\varepsilon \cdot 2^{\frac{\hat{j}(\zeta,k)+3}{2}}. \quad (3.3.17)$$

Then the  $(1 - \alpha)$  level confidence interval for  $M(\mathbf{f})$  is

$$CI_{m,\alpha} = [\mathbf{f}_{lo}, \mathbf{f}_{hi}]. \quad (3.3.18)$$

### 3.3.3. Statistical Optimality.

In this section, we establish the optimality of the adaptive procedures constructed in Section 3.3.2. The results show that the data driven estimators and the confidence interval (hyper cube) achieve within a universal constant factor depending on  $s$  and  $\alpha$  only of their corresponding benchmarks simultaneously for all  $\mathbf{f} \in \mathcal{F}_s$ . These results are non-asymptotic and function-specific, which are much stronger than the conventional minimax framework. We start with estimation of the minimizer.

**Theorem 3.3.1** (Estimation for Minimizer). *The estimator  $\hat{Z}$  defined by (3.3.10) satisfies*

$$\mathbb{E}_{\mathbf{f}} \left( \|\hat{Z} - Z(\mathbf{f})\|^2 \right) \leq C_{z,s} R_z(\varepsilon; \mathbf{f}), \text{ for all } \mathbf{f} \in \mathcal{F}_s, \quad (3.3.19)$$

where  $C_{z,s} > 0$  is a constant depending on dimension  $s$ .

The following holds for the confidence hyper cube  $CI_{z,\alpha}$ .

**Theorem 3.3.2** (Confidence Hyper-cube for Minimizer). *For  $0 < \alpha \leq 0.3$ , the confidence hyper cube  $CI_{z,\alpha}$  defined by (3.3.12) is a  $1 - \alpha$  level confidence hyper cube for the minimizer  $Z(\mathbf{f})$ . Its expected volume satisfies*

$$\mathbb{E}_{\mathbf{f}} (V(CI)) \leq C_{z,s,\alpha} L_{\alpha,z}(\varepsilon; \mathbf{f}),$$

where  $C_{z,s,\alpha}$  is a positive constant depending on  $s$  and  $\alpha$ .

**Theorem 3.3.3** (Estimation for Minimum). *The estimation  $\hat{M}$  defined in (3.3.14) satisfies*

$$E \left( (\hat{M} - M(\mathbf{f}))^2 \right) \leq C_{m,s} R_m(\varepsilon; \mathbf{f}), \quad (3.3.20)$$

where  $C_{m,s}$  is a positive constant depending on dimension  $s$ .

**Theorem 3.3.4** (Confidence Interval for Minimum). *For  $0 < \alpha \leq 0.3$ , the confidence interval defined by (3.3.18) is a  $1 - \alpha$  level confidence interval for the minimum  $M(\mathbf{f})$  satisfying*

$$\mathbb{E}(|CI_{m,\alpha}|) \leq C_{m,s,\alpha} L_{\alpha,m}(\varepsilon; \mathbf{f}), \quad (3.3.21)$$

where  $C_{m,s,\alpha}$  is a positive constant depending on  $\alpha$  and  $s$ .

### 3.4. Nonparametric Regression

We have so far focused on the white noise model. The procedures and results presented in the previous sections can be extended to nonparametric regression, where we observe

$$y_{i_1, i_2, \dots, i_s} = \mathbf{f}(i_1/n, i_2/n, \dots, i_s/n) + \sigma z_{i_1, i_2, \dots, i_s}, 0 \leq i_k \leq n, \text{ for } 1 \leq k \leq s, \quad (3.4.1)$$

with  $z_{i_1, i_2, \dots, i_s} \stackrel{i.i.d}{\sim} N(0, 1)$ ,  $\mathbf{f} \in \mathcal{F}_s$ . The noise level  $\sigma$  is assumed to be known. The tasks are the same as before: constructing optimal estimators and confidence interval (hyper cube) for the minimizer  $Z(\mathbf{f})$  and the minimum  $M(\mathbf{f})$ , for  $\mathbf{f} \in \mathcal{F}_s$ . For simplicity of notation, we take  $\mathbf{i} = (i_1, i_2, \dots, i_s)$ . To avoid trivial case, we suppose  $n \geq 2$ .

#### 3.4.1. Local Minimax Rates, Discretization Error and Separable Representation

Analogous to the benchmarks for the white noise model defined in Equations (3.1.3), (3.1.4), (3.1.6), we define similar benchmarks for the nonparametric regression model (3.4.1) with  $n + 1$  equally spaced observations. Denote by  $\mathcal{I}_{m,\alpha,n}(\mathfrak{F})$  the collection of  $(1 - \alpha)$  level confidence intervals for  $M(f)$  on a function class  $\mathfrak{F}$  under the regression model (3.4.1) and

let

$$\begin{aligned}
\tilde{\mathbf{R}}_{z,n}(\sigma; \mathbf{f}) &= \sup_{\mathbf{g} \in \mathcal{F}_s} \inf_{\hat{Z}} \max_{h \in \{\mathbf{f}, \mathbf{g}\}} \mathbb{E}_h \|\hat{Z} - Z(h)\|^2, \\
\tilde{\mathbf{R}}_{m,n}(\sigma; \mathbf{f}) &= \sup_{\mathbf{g} \in \mathcal{F}_s} \inf_{\hat{M}} \max_{h \in \{\mathbf{f}, \mathbf{g}\}} \mathbb{E}_h (\hat{M} - M(h))^2, \\
\tilde{\mathbf{L}}_{m,\alpha,n}(\sigma; \mathbf{f}) &= \sup_{\mathbf{g} \in \mathcal{F}_s} \inf_{CI_{m,\alpha} \in \mathcal{I}_{m,\alpha,n}(\{\mathbf{f}, \mathbf{g}\})} \mathbb{E}_{\mathbf{f}} |CI_{m,\alpha}|.
\end{aligned} \tag{3.4.2}$$

For confidence hyper cube for minimizer, denote  $\mathcal{I}_{z,\alpha,n}(\mathfrak{F})$  the collection of  $(1 - \alpha)$  level confidence hyper cube on a function class  $\mathfrak{F}$  under the regression model (3.4.1) and let

$$\tilde{\mathbf{L}}_{z,\alpha,n}(\sigma; \mathbf{f}) = \inf_{CI \in \mathcal{I}_{z,\alpha,n}(\mathcal{F}_s)} \mathbb{E}_{\mathbf{f}} V(CI). \tag{3.4.3}$$

It is clear that the expected volume for confidence hyper cube of the minimizer can not be smaller than  $\tilde{\mathbf{L}}_{z,\alpha,n}(\sigma; \mathbf{f})$ , which is also function-specific, i.e. depending on  $\mathbf{f}$ .

Compared with white noise model, in addition to the difference in the probability structure caused by discrete observations, estimation and inference for both  $Z(\mathbf{f})$  and  $M(\mathbf{f})$  incur additional discretization errors, even in the noiseless case. See the appendix Section A.5.12 for further discussion.

### Separable Representation

Analogous to the white noise model, the observation under nonparametric setting also admits a separable representation, as defined in Definition 3.4.1.

**Definition 3.4.1** (Projection Representation for Nonparametric Regression Model). *For  $k \in \{1, 2, \dots, s\}$ , the  $k$ -th projection of  $\{y_{\mathbf{i}}\}$ ,  $\boldsymbol{\pi}_k(\{y_{\mathbf{i}}\})$ , is an  $n + 1$ -long random vector,*

$$\begin{aligned}
\boldsymbol{\pi}_k(\{y_{\mathbf{i}}\}) &= \\
&\left( \frac{\sum_{\mathbf{i}: i_k=1} y_{\mathbf{i}}}{(n+1)^{s-1}} - \frac{\sum_{\mathbf{i}} y_{\mathbf{i}}}{(n+1)^s}, \frac{\sum_{\mathbf{i}: i_k=2} y_{\mathbf{i}}}{(n+1)^{s-1}} - \frac{\sum_{\mathbf{i}} y_{\mathbf{i}}}{(n+1)^s}, \dots, \frac{\sum_{\mathbf{i}: i_k=s} y_{\mathbf{i}}}{(n+1)^{s-1}} - \frac{\sum_{\mathbf{i}} y_{\mathbf{i}}}{(n+1)^s} \right).
\end{aligned} \tag{3.4.4}$$

$\mathbf{er}(\{y_{\mathbf{i}}\})$  is an  $s$ -dimension tensor with

$$\mathbf{er}(\{y_{\mathbf{i}}\})_{i_1, i_2, \dots, i_s} = y_{i_1, i_2, \dots, i_s} - \sum_{k=1}^s \pi_k(\{y_{\mathbf{i}}\})_{i_k}, \quad (3.4.5)$$

for  $0 \leq i_k \leq n$ ,  $1 \leq k \leq s$ .

The projection representation mapping  $\mathfrak{P}(\cdot)$  of observation  $\{y_{\mathbf{i}}\}$  is given by

$$\mathfrak{P}(\{y_{\mathbf{i}}\}) = (\pi_1(\{y_{\mathbf{i}}\}), \pi_2(\{y_{\mathbf{i}}\}), \pi_s(\{y_{\mathbf{i}}\}), \mathbf{er}(\{y_{\mathbf{i}}\})). \quad (3.4.6)$$

Similar to white noise model,  $\mathfrak{P}(\cdot)$  preserves the information of  $\{y_{\mathbf{i}}\}$ ; has its  $s + 1$  elements being mutually independent; and separates the information for the  $s$  univariate component functions of  $\mathbf{f}$  into its first  $s$  random variables, as shown in Proposition 3.4.1.

**Proposition 3.4.1** (Property of Projection Representation). *Let  $\mathfrak{P}(\cdot)$  be define in equation (3.4.6). Then we have*

- $\mathfrak{P}(\cdot)$  is invertible,
- $\mathfrak{P}(\{y_{\mathbf{i}}\})$  has its  $s + 1$  elements being independent,
- $\pi_k(\{y_{\mathbf{i}}\})$  is sufficient statistic for  $f_k$ .

### 3.4.2. Optimal Procedures

Similar to the white noise model, we split the data into three independent copies and then construct the estimators and confidence interval (hyper cube) for  $Z(\mathbf{f})$  and  $M(\mathbf{f})$  for  $\mathbf{f} \in \mathcal{F}_s$  in three major steps: localization, stopping, and estimation/inference.

### Data Splitting

Let  $z_{k,i}^j \stackrel{i.i.d}{\sim} N(0, 1)$ , with  $1 \leq k \leq s$ ,  $1 \leq i \leq n$ ,  $1 \leq j \leq 2$ .

For each  $1 \leq k \leq s$ , we construct the following three sequences based on  $\boldsymbol{\pi}_k(\{y_i\})$ :

$$\begin{aligned}\nu_{k,i}^l &= \boldsymbol{\pi}_k(\{y_i\})_i + \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \left\{ \frac{\sqrt{2}}{2} \left( z_{k,i}^1 - \frac{\sum_{l=0}^n z_{k,l}^1}{n+1} \right) + \frac{\sqrt{6}}{2} \left( z_{k,i}^2 - \frac{\sum_{l=0}^n z_{k,l}^2}{n+1} \right) \right\}, \\ \nu_{k,i}^r &= \boldsymbol{\pi}_k(\{y_i\})_i + \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \left\{ \frac{\sqrt{2}}{2} \left( z_{k,i}^1 - \frac{\sum_{l=0}^n z_{k,l}^1}{n+1} \right) - \frac{\sqrt{6}}{2} \left( z_{k,i}^2 - \frac{\sum_{l=0}^n z_{k,l}^2}{n+1} \right) \right\}, \\ \nu_{k,i}^e &= \boldsymbol{\pi}_k(\{y_i\})_i - \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \sqrt{2} \left( z_{k,i}^1 - \frac{\sum_{l=0}^n z_{k,l}^1}{n+1} \right),\end{aligned}\tag{3.4.7}$$

for  $i = 0, \dots, n$ . For convenience, let  $\nu_{k,i}^l = \nu_{k,i}^r = \nu_{k,i}^e = \infty$  for  $i \notin \{0, 1, \dots, n\}$ . It is easy to see that the three sequences for each axis  $k$  are independent, and the  $s$  collections of the three sequences are also independent. For each  $k$ , we will use  $\{\nu_{k,\cdot}^l\}$  for localization,  $\{\nu_{k,\cdot}^r\}$  for stopping rule, and  $\{\nu_{k,\cdot}^e\}$  for construction of the final estimation and inference procedures.

Let  $J = \lfloor \log_2(n+1) \rfloor$ . For  $j = 0, 1, \dots, J$ ,  $i = 1, 2, \dots, \lfloor \frac{n+1}{2^{J-j}} \rfloor$ , the  $i$ -th block at level  $j$  consists of  $\{\frac{(i-1)2^{J-j}}{n}, \frac{(i-1)2^{J-j}+1}{n}, \frac{i \cdot 2^{J-j}-1}{n}\}$ . Denote the sum of observations in the  $i$ -th block at level  $j$  for the axis  $k$ , sequence  $u$  ( $u=l,r,e$ ) as

$$Y_{k,j,i}^u = \sum_{h=(i-1)2^{J-j}}^{i \cdot 2^{J-j}-1} \nu_{k,h}^u.\tag{3.4.8}$$

Again, let  $Y_{k,j,i}^u = +\infty$  when  $i \notin \{1, 2, \dots, \lfloor \frac{n+1}{2^{J-j}} \rfloor\}$  for  $k \in \{1, 2, \dots, s\}$ ,  $u \in \{l, r, e\}$ ,  $j \in \{0, 1, \dots, J\}$ .

### Localization

For  $k$ -th axis, we use  $\{\nu_{k,h}^l, h \in \{0, 1, \dots, n\}\}$  to construct a localization procedure. Let  $\hat{\mathbf{i}}_{k,0} = 1$ , and for  $j = 1, 2, \dots, J$ , let

$$\hat{\mathbf{i}}_{k,j} = \arg \min_{\max\{2\hat{\mathbf{i}}_{k,j-1}-2, 1\} \leq i \leq \min\{2\hat{\mathbf{i}}_{k,j-1}+1, \lfloor \frac{n+1}{2^{J-j}} \rfloor\}} Y_{k,j,i}^l.\tag{3.4.9}$$



This is similar to the localization step in the white noise model. In each iteration, the blocks at the previous level are split into two sub-blocks. The  $i$ -th block at level  $j - 1$  is split into two blocks, the  $(2i - 1)$ -th block and the  $2i$ -th block, at level  $j$ . For a given  $\hat{\mathbf{i}}_{k,j-1}$ ,  $\hat{\mathbf{i}}_{k,j}$  is the sub-block with the smallest sum (i.e.  $\mathbf{Y}_{k,j,i}^l$ ) among the two sub-blocks of  $\hat{\mathbf{i}}_{k,j-1}$  and their immediate neighboring sub-blocks.

### Stopping Rule

Similar to the stopping rule for the white noise model, for axis  $k$ , define the statistic  $\mathbf{T}_{k,j}$  based on the sequence  $\mathbf{Y}_{k,\cdot,\cdot}^r$  as

$$\mathbf{T}_{k,j} = \min\{\mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}+6}^r - \mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}+5}^r, \mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}-6}^r - \mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}-5}^r\}.$$

Let  $\tilde{\sigma}_{k,j}^2 = 6 \times 2^{J-j} \times \frac{\sigma^2}{(n+1)^{s-1}}$ . It is easy to see that when  $\mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}+6}^r - \mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}+5}^r < \infty$ ,

$$\mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}+6}^r - \mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}+5}^r \Big| \hat{\mathbf{i}}_{k,j} \sim N \left( \sum_{h=(\hat{\mathbf{i}}_{k,j}+4)2^{J-j}}^{(\hat{\mathbf{i}}_{k,j}+5)2^{J-j}-1} \left( f_k\left(\frac{h+2^{J-j}}{n}\right) - f_k\left(\frac{h}{n}\right) \right), \tilde{\sigma}_{k,j}^2 \right). \quad (3.4.10)$$

Similar to white noise model, we define a series of stopping rules controlled by a parameter  $\zeta > 0$ .

Define a *stopping step precursor*  $\check{\mathbf{j}}_k(\zeta)$  as

$$\check{\mathbf{j}}_k(\zeta) = \begin{cases} \min\{j : \mathbf{T}_{k,j} \leq z_\zeta \tilde{\sigma}_{k,j}\} & \text{if } \{j : \mathbf{T}_{k,j} \leq z_\zeta \tilde{\sigma}_{k,j}\} \cap \{0, 1, 2, \dots, J\} \neq \emptyset \\ \infty & \text{otherwise} \end{cases}$$

and terminate the algorithm at level  $\hat{\mathbf{j}}_k(\zeta) = \min\{J, \check{\mathbf{j}}_k(\zeta)\}$ . So either  $\mathbf{T}_{k,j}$  triggers the stopping for some  $0 \leq j \leq J$  or the algorithm reaches the highest possible level  $J$ .

With the localization strategy and the stopping rule, the final block, the  $\hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\zeta)}$ -th block

at level  $\hat{\mathbf{j}}_k(\zeta)$  is given by

$$\left\{ \frac{h}{n} : (\hat{\mathbf{i}}_{k, \hat{\mathbf{j}}_k(\zeta)} - 1)2^{J - \hat{\mathbf{j}}_k(\zeta)} \leq h \leq \hat{\mathbf{i}}_{k, \hat{\mathbf{j}}_k(\zeta)}2^{J - \hat{\mathbf{j}}_k(\zeta)} - 1 \right\}.$$

### Estimation and Inference

After we have, for each axis  $k \in \{1, 2, \dots, s\}$ , our stopping step precursor  $\check{\mathbf{j}}_k(\zeta)$ , stopping step  $\hat{\mathbf{j}}_k(\zeta)$ , index associated with the stopping step  $\hat{\mathbf{i}}_{k, \hat{\mathbf{j}}_k(\zeta)}$ , and the final block, we use them to construct estimator and confidence hyper cube for the minimizer of  $\mathbf{f} \in \mathcal{F}_s$ , as well as estimator and confidence interval for the minimum of  $\mathbf{f} \in \mathcal{F}_s$ .

For estimation of the minimizer, let  $\zeta = \Phi(-2)$ . The  $k$ -th coordinate of  $\hat{Z}$ ,  $\hat{Z}_k$ , is defined as

$$\hat{Z}_k = \begin{cases} -\frac{1}{2n} + \frac{1}{n} \left( 2^{J - \hat{\mathbf{j}}_k(\zeta)} - 2^{J - \hat{\mathbf{j}}_k(\zeta) - 1} \right), & \check{\mathbf{j}}_k(\zeta) < \infty \\ \frac{1}{n} \arg \min_{\hat{\mathbf{i}}_{k, J-2 \leq i \leq \hat{\mathbf{i}}_{k, J+2}} \nu_{k, i-1}^e - \frac{1}{n}, & \check{\mathbf{j}}_k(\zeta) = \infty \end{cases}. \quad (3.4.11)$$

The final estimator  $\hat{Z}$  is defined as

$$\hat{Z} = (\hat{Z}_1, \hat{Z}_2, \dots, \hat{Z}_s), \quad (3.4.12)$$

where  $\hat{Z}_k$  is defined in (3.4.11) for  $k \in \{1, 2, \dots, s\}$ .

To construct the confidence hyper cube for  $Z(\mathbf{f})$ , for each axis  $k \in \{1, \dots, s\}$ , we set the parameter for stopping rule to be  $\zeta_k = \alpha/2s$  and take a few adjacent blocks at level  $\hat{\mathbf{j}}_k(\zeta_k) - 1$  to the left and right of  $\hat{\mathbf{i}}_{k, \hat{\mathbf{j}}_k(\zeta_k) - 1}$ -th block.

Let

$$L_k = \max\{0, 2 \cdot (\hat{\mathbf{i}}_{k, \hat{\mathbf{j}}_k(\alpha/2s) - 1} - 7)\}, U_k = \min\{2 \cdot (\hat{\mathbf{i}}_{k, \hat{\mathbf{j}}_k(\alpha/2s)} + 6), \lceil (n+1)2^{\hat{\mathbf{j}}_k(\alpha/2s) - J} \rceil\}.$$

When  $\check{\mathbf{j}}_k(\alpha/2s) < \infty$ , let

$$t_{k,lo} = \frac{2^{J-\hat{\mathbf{j}}_k(\alpha/2s)}}{n} L_k - \frac{1}{2n}, t_{k,hi} = \min\left\{\frac{2^{J-\hat{\mathbf{j}}_k(\alpha/2s)}}{n} U_k - \frac{1}{2n}, 1\right\}. \quad (3.4.13)$$

When  $\check{\mathbf{j}}_k(\alpha/2s) = \infty$ ,  $t_{k,lo}$  and  $t_{k,hi}$  are calculated by the following Algorithm 3.

The key ideas of Algorithm 3 are as follows.

$\check{\mathbf{j}}_k(\alpha/2s) = \infty$  means that  $\mathbf{T}_{k,j}$  never triggers the stopping, which is a strong indicator that the signal is strong and discretization error could dominate. Algorithm 3 first specifies a range that the minimizer lies in with high probability (e.g.  $1 - \alpha/2s$ ), and then shrinks the interval to locate the minimizer among the grid points within the original interval. After this step, the minimizer(s) among the grids are in the shrunk interval with high probability (e.g.  $1 - 3\alpha/4s$ ). Then in the case that shrunk interval detects only one grid-wise minimizer ( $i_m/n$ ) and this minimizer does not indicate a discretization error larger or equal than  $1/n$  (i.e.  $i_m = 1$  or  $i_m = n - 1$ ), we use a geometry property of convex functions to determine the final interval. Basically, the right most possible minimizer is or is infinitely near to the intersection of two lines :  $y = f(i_m/n)$ , and the line joining  $(\frac{i_m+1}{n}, f(\frac{i_m+1}{n}))$  with  $(\frac{i_m+2}{n}, f(\frac{i_m+2}{n}))$ . With observation  $\nu_{k,i_m}^e, \nu_{k,i_m+1}^e, \nu_{k,i_m+2}^e$ , we can infer the intersection of the aforementioned two lines and specify the right end point of the interval accordingly.

The  $k$ -th axis of confidence hyper cube  $CI_{z,\alpha}$  is given by

$$CI_{k,\alpha} = [t_{k,lo}, t_{k,hi}]. \quad (3.4.14)$$

The  $(1 - \alpha)$ -level confidence hyper cube  $CI_{z,\alpha}$  is given by

$$CI_{z,\alpha} = CI_{1,\alpha} \times CI_{2,\alpha} \times \cdots \times CI_{s,\alpha}, \quad (3.4.15)$$

where  $CI_{k,\alpha}$  is defined in (3.4.14).

---

**Algorithm 3** Computing  $t_{k,lo}$  and  $t_{k,hi}$  when  $\check{\mathbf{j}}_k(\zeta) = \infty$

---

$L_k \leftarrow \max\{0, 2\hat{\mathbf{i}}_{k,\check{\mathbf{j}}_k(\alpha/2s)-1} - 15\}, U_k = \min\{n, 2\hat{\mathbf{i}}_{k,\check{\mathbf{j}}_k(\alpha/2s)-1} + 12\}, \alpha_1 = \alpha/8s, \alpha_2 = \alpha/24s$

Generate  $z_{k,0}^3, z_{k,2}^3, \dots, z_{k,n}^3 \stackrel{i.i.d.}{\sim} N(0, 1)$

$$i_l \leftarrow \min\{\{U\} \cup \{i \in [L, U-1] : \nu_{k,i}^e - \nu_{k,i+1}^e + \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (z_{k,i}^3 - z_{k,i+1}^3 - 2z_{\alpha_1}) \leq 0\}$$

$$i_r \leftarrow \max\{\{L-1\} \cup \{i \in [L, U-1] : \nu_{k,i}^e - \nu_{k,i+1}^e + \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (z_{k,i}^3 - z_{k,i+1}^3 + 2z_{\alpha_1}) \geq 0\}$$

**if**  $i_l \leq i_r$  **then**

$$t_{k,lo} = \max\{0, \frac{i_l-1}{n}\}, t_{k,hi} = \max\{1, \frac{i_r+2}{n}\}$$

**end if**

**if**  $i_l = i_r + 1$  and  $i_l \leq n - 2$  **then**

**if**  $\nu_{k,i_l+2}^e - \nu_{k,i_l+1}^e - \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (z_{k,i_l+2}^3 - z_{k,i_l+1}^3 - 2\sqrt{2}z_{\alpha_2}) > 0$  **then**

$$t_{hi} \leftarrow \left( \left( \frac{\nu_{k,i_l}^e - \nu_{k,i_l+1}^e - \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (z_{k,i_l}^3 - z_{k,i_l+1}^3 - 2\sqrt{2}z_{\alpha_2})}{n \left( \nu_{k,i_l+2}^e - \nu_{k,i_l+1}^e - \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (z_{k,i_l+2}^3 - z_{k,i_l+1}^3 - 2\sqrt{2}z_{\alpha_2}) \right)} + \frac{1}{n} \right) + \frac{i_l}{n} \right) \wedge \frac{i_l+1}{n}$$

**else**

$$t_{hi} \leftarrow \frac{i_l}{n}$$

**end if**

**end if**

**if**  $i_l = i_r + 1$  and  $i_l \geq n - 1$  **then**

$$t_{k,hi} = 1$$

**end if**

**if**  $i_l = i_r + 1$  and  $i_l \geq 2$  **then**

**if**  $\nu_{k,i_l-2}^e - \nu_{k,i_l-1}^e - \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (z_{k,i_l-2}^3 - z_{k,i_l-1}^3 - 2\sqrt{2}z_{\alpha_2}) > 0$  **then**

$$t_{k,lo} \leftarrow \left( \left( - \frac{\nu_{k,i_l}^e - \nu_{k,i_l-1}^e - \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (z_{k,i_l}^3 - z_{k,i_l-1}^3 - 2\sqrt{2}z_{\alpha_2})}{n \left( \nu_{k,i_l-2}^e - \nu_{k,i_l-1}^e - \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (z_{k,i_l-2}^3 - z_{k,i_l-1}^3 - 2\sqrt{2}z_{\alpha_2}) \right)} + \frac{i_l-1}{n} \right) \wedge \frac{i_l}{n} \right)$$

**else**

$$t_{k,lo} \leftarrow \frac{i_l}{n}$$

**end if**

**end if**

**if**  $i_l = i_r + 1$  and  $i_l \leq 1$  **then**

$$t_{k,lo} = 0$$

**end if**

---

Now we turn to the construction of the estimator and confidence interval for the minimum.

We start with estimation for the minimum  $M(\mathbf{f})$ . Let  $\zeta = \Phi(-2)$ . For axis  $k$ , let

$$\Delta_k = \mathbb{1}\{\mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}+6}^r - \mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}+5}^r \leq z_\zeta \tilde{\sigma}_{k,\hat{\mathbf{j}}_k(\zeta)}^2\} - \mathbb{1}\{\mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}-6}^r - \mathbf{Y}_{k,j,\hat{\mathbf{i}}_{k,j}-5}^r \leq z_\zeta \tilde{\sigma}_{k,\hat{\mathbf{j}}_k(\zeta)}^2\}.$$

The estimator for  $M(\mathbf{f})$  is given as follows.

We define  $s$  intermediate estimators  $\hat{M}_k$  as

$$\hat{M}_k = \begin{cases} 2^{\hat{\mathbf{j}}_k(\zeta)-J} \mathbf{Y}_{k,\hat{\mathbf{j}}_k(\zeta),\hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\zeta)}+2\Delta_k}^e, & \check{\mathbf{j}}_k(\zeta) < \infty \\ \min_{\hat{\mathbf{i}}_{k,J-2} \leq i \leq \hat{\mathbf{i}}_{k,J+2}} \nu_{k,i-1}^e, & \check{\mathbf{j}}_k(\zeta) = \infty \end{cases}. \quad (3.4.16)$$

The final estimator  $\hat{M}$  is defined as

$$\hat{M} = \frac{1}{(n+1)^s} \sum_{\mathbf{i} \in \{0,1,2,\dots,n\}^s} \mathbf{er}(\{y_{\mathbf{i}}\}) + \sum_{k=1}^s \hat{M}_k. \quad (3.4.17)$$

Now we continue with the confidence interval for the minimum  $M(\mathbf{f})$ . Let  $\zeta_k = \zeta = \alpha/4s$ .

Define the step number that will be used for constructing the interval as

$$j_{F,k} = \begin{cases} \check{\mathbf{j}}_k(\zeta) + 3, & \text{for } \check{\mathbf{j}}_k(\zeta) \leq J \\ \infty, & \text{for } \check{\mathbf{j}}_k(\zeta) = \infty \end{cases} \quad (3.4.18)$$

Basically, we go three steps forward from the step that the test statistic  $\mathbf{T}_{k,j}$  triggers the stopping rule.

Define

$$\begin{aligned} I_{k,lo} &= 2^{(j_{F,k} \wedge J) - \hat{\mathbf{j}}_k(\zeta) + 1} \times \left( \hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\zeta)-1} - 7 \right), \\ I_{k,hi} &= 2^{(j_{F,k} \wedge J) - \hat{\mathbf{j}}_k(\zeta) + 1} \times \left( \hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\zeta)-1} + 6 \right) + 1 \end{aligned} \quad (3.4.19)$$

We first define 3 sets of  $s$  intermediate estimators  $\{\tilde{M}_{k,md} : 1 \leq k \leq s\}, \{\tilde{M}_{k,hi} : 1 \leq k \leq$

$s\}, \{\tilde{M}_{k,lo} : 1 \leq k \leq s\}$  as

$$\tilde{M}_{k,md} = \min_{I_{k,lo} \leq i \leq I_{k,hi}} \mathbf{Y}_{k,(j_{F,k} \wedge J),i}^e \times 2^{(j_{F,k} \wedge J) - J}, \quad (3.4.20)$$

$$\tilde{M}_{k,hi} = \tilde{M}_{k,md} + S_{210,\alpha/8s} \times \sqrt{3} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \times 2^{\frac{(j_{F,k} \wedge J) - J}{2}} \quad (3.4.21)$$

and

$$\tilde{M}_{k,lo} = \tilde{M}_{k,md} - \frac{3\sigma(z_{\alpha/4s} + 1)}{(n+1)^{\frac{s-1}{2}}} \times 2^{\frac{j_{F,k} - J}{2}} - S_{210,\alpha/8s} \times \sqrt{3} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \times 2^{\frac{(j_{F,k} \wedge J) - J}{2}} \text{ for } j_{F,k} \leq J. \quad (3.4.22)$$

Let  $\tilde{M}_{k,lo}$  be computed by Algorithm 4 when  $j_{F,k} > J$ . Algorithm 4 is based on the geometric property of the convex function  $f$  that for any  $1 \leq i \leq n-2$ ,

$$\inf_{t \in [\frac{i}{n}, \frac{i+1}{n}]} f(t) \geq \inf_{t \in [\frac{i}{n}, \frac{i+1}{n}]} \max \left\{ \frac{f_k(\frac{i+2}{n}) - f_k(\frac{i+1}{n})}{1/n} (t - \frac{i+1}{n}) + f_k(\frac{i+1}{n}), \right. \\ \left. \frac{f_k(\frac{i}{n}) - f_k(\frac{i-1}{n})}{1/n} (t - \frac{i}{n}) + f_k(\frac{i}{n}) \right\}.$$

---

**Algorithm 4** Computing  $\tilde{M}_{k,lo}$  when  $j_{F,k} > J$ 


---

$k_l \leftarrow \max\{0, I_{k,lo} - 1\}, k_r \leftarrow \min\{n - 1, I_{k,hi} - 2\}, H \leftarrow S_{k_r - k_l + 4, \frac{\alpha}{24s}} \sqrt{3} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} +$   
 $z_{\frac{\alpha}{48s}} \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s}{2}}}$   
**if**  $k_l = 0$  **then**  
 $v_{r,0}(t) \leftarrow \frac{\nu_{k,2}^e - \nu_{k,1}^e + 2H}{1/n} (t - 1/n) + \nu_{k,1}^e - H, h(0) \leftarrow \min_{t \in [0, 1/n]} v_{r,0}(t)$   
**end if**  
**if**  $k_r = n - 1$  **then**  
 $v_{l,n-1}(t) \leftarrow \frac{\nu_{k,n-1}^e - \nu_{k,n-2}^e - 2H}{1/n} (t - \frac{n-1}{n}) + \nu_{k,n-1}^e - H, h(n-1) = \min_{t \in [\frac{n-1}{n}, 1]} v_{l,n-1}(t)$   
**end if**  
**for**  $i = (k_l \vee 1), \dots, (k_r \wedge n - 2)$  **do**  
Define two linear functions:  

$$v_{l,i}(t) = \frac{\nu_{k,i}^e - \nu_{k,i-1}^e - 2H}{1/n} (t - \frac{i}{n}) + \nu_{k,i}^e - H,$$

$$v_{r,i} = \frac{\nu_{k,i+2}^e - \nu_{k,i+1}^e + 2H}{1/n} (t - \frac{i+1}{n}) + \nu_{k,i+1}^e - H$$

$$h(i) = \min_{t \in [\frac{i}{n}, \frac{i+1}{n}]} \max\{v_{l,i}(t), v_{r,i}(t)\}$$
  
**end for**  
 $\tilde{M}_{k,lo} \leftarrow \min\{h(i) : k_l \leq i \leq k_r\} \wedge \tilde{M}_{k,hi}$

---

Let

$$\tilde{M}_{hi} = \frac{1}{(n+1)^s} \sum_{\mathbf{i} \in \{0,1,2,\dots,n\}^s} \mathbf{er}(\{y_{\mathbf{i}}\}) + \sum_{k=1}^s \tilde{M}_{k,hi} + z_{\alpha/8} \cdot 2\sqrt{3} \frac{\sigma}{(n+1)^{\frac{s}{2}}} s, \quad (3.4.23)$$

$$\tilde{M}_{lo} = \frac{1}{(n+1)^s} \sum_{\mathbf{i} \in \{0,1,2,\dots,n\}^s} \mathbf{er}(\{y_{\mathbf{i}}\}) + \sum_{k=1}^s \tilde{M}_{k,lo} - z_{\alpha/8} \cdot 2\sqrt{3} \frac{\sigma}{(n+1)^{\frac{s}{2}}} s. \quad (3.4.24)$$

The confidence interval for the minimum  $M(\mathbf{f})$  is given by

$$CI_{m,\alpha} = [\tilde{M}_{lo}, \tilde{M}_{hi}]. \quad (3.4.25)$$

### 3.4.3. Statistical Optimality

Now we establish the optimality of the adaptive procedures constructed in Section 3.4.2. The results show that our procedures are simultaneously optimal (up to a constant depending on dimension and confidence level) for  $\mathbf{f} \in \mathcal{F}_s$  in terms our benchmarks introduced in (3.4.2) and (3.4.3).

We begin with the estimator of the minimizer.

**Theorem 3.4.1** (Estimation for Minimizer). *The estimator  $\hat{Z}$  defined in (3.4.12) satisfies*

$$\mathbb{E}_{\mathbf{f}} \left( \|\hat{Z} - Z(\mathbf{f})\|^2 \right) \leq Q_{z,s} \tilde{\mathbf{R}}_{z,n}(\sigma; \mathbf{f}), \text{ for all } \mathbf{f} \in \mathcal{F}_s \quad (3.4.26)$$

where  $Q_{z,s}$  is a positive constant depending on  $s$ .

For the confidence hyper cube  $CI_{z,\alpha}$  of  $Z(\mathbf{f})$ , we have the following result.

**Theorem 3.4.2** (Inference for Minimizer). *For  $0 < \alpha \leq 0.3$ , confidence cube  $CI_{z,\alpha}$  defined in (3.4.15) is a  $(1 - \alpha)$ -level confidence cube for the minimizer  $Z(\mathbf{f})$  and its expected volume satisfies*

$$\mathbb{E}_{\mathbf{f}} (V(CI_{z,\alpha})) \leq Q_{z,s,\alpha} \tilde{\mathbf{L}}_{z,\alpha,n}(\sigma; \mathbf{f}), \text{ for all } \mathbf{f} \in \mathcal{F}_s \quad (3.4.27)$$

where  $Q_{z,s,\alpha}$  is a positive constant depending on  $s$  and  $\alpha$  only.

Similarly, the estimator and confidence interval for the minimizer  $M(\mathbf{f})$  also achieve within a constant depending on  $s$  and  $\alpha$  of the corresponding benchmark simultaneously for all  $\mathbf{f} \in \mathcal{F}_s$ .

**Theorem 3.4.3** (Estimation for Minimum). *The estimator  $\hat{M}$  defined in (3.4.17) satisfies*

$$\mathbb{E} \left( \hat{M} - M(\mathbf{f}) \right)^2 \leq Q_{m,s} \tilde{\mathbf{R}}_{m,n}(\sigma; \mathbf{f}) \quad (3.4.28)$$

where  $Q_{m,s}$  is a positive constant depending on  $s$ .

**Theorem 3.4.4** (Inference for Minimum). *For  $0 < \alpha \leq 0.3$ , the confidence interval  $CI_{m,\alpha}$*



defined in (3.4.25) is a  $(1 - \alpha)$  level confidence interval for minimum  $M(\mathbf{f})$  and its expected length satisfies

$$\mathbb{E}(|CI_{m,\alpha}|) \leq Q_{m,s,\alpha} \tilde{\mathbf{L}}_{m,\alpha,n}(\sigma; \mathbf{f}), \quad (3.4.29)$$

where  $Q_{m,s,\alpha}$  is a positive constant depending on dimension  $s$  and  $\alpha$ .

## CHAPTER 4

### Interplay Between Statistical Accuracy and Running Time Cost: A Framework and Examples

#### 4.1. Introduction

With the advent of iterative methods and the increasing scale of data, computational cost has become a great concern in addition to statistical accuracy. Approaches from different angles have been proposed, including categorizing different methods with the triple of sample size, computation time and statistical error (Chandrasekaran and Jordan, 2013), computational-theoretical approach that differentiates between regions of parameters where the problem is polynomial-time computable or not polynomial-time computable (Wang et al., 2016; Berthet and Rigollet, 2013), reducing the effective sample size (Shender and Lafferty, 2013; Horev et al., 2015; Sussman et al., 2015; Kpotufe and Verma, 2017), and separately investigating both optimization running time and statistical accuracy, when the problem enjoys good properties like a certain form of strong convexity, smoothness or isotropic property (Loh and Wainwright, 2015; Wang et al., 2017; Chen and Wainwright, 2015; Bottou and Bousquet, 2011).

Our approach is to provide theoretically guaranteed iterative optimization algorithm and precise quantification of how iteration number affects the statistical accuracy for a class of problems that admits estimators of a certain general form without imposing artificial or hard-to-verify conditions.

Our approach is different from the computational-theoretical approach in that we quantify the affects of running time on statistical accuracy on a continuous scale rather than a binary answer of polynomial time computability.

Compared with literature that deals with only statistical problem, only statistically rooted optimization problem, or both optimization and statistical aspects of a statistical prob-

lem, our approach provide theoretically guaranteed optimization procedure; our approach provides refined optimization-wise convergence rate that considers the dimension of the statistical problem as a changing quantity rather than a constant; and our approach combines optimization and statistic in a more intrinsic way so that we do not need artificial hard-to-verify conditions to give theoretical guarantee for our optimization procedure in terms of its influence on statistical accuracy.

To further illustrate this, we digress a little into the existing works.

Existing literature usually treats statistical properties and optimization properties separately. Statistical properties (i.e. statistical convergence rate) are usually established for a perfect solution of an optimization problem. And optimization convergence rates are established targeting the perfect solution for a certain method. Literature attempting to consider both aspects jointly also follow this style.

But this separation has three undesired consequences. It requires assumptions that facilitates convergence rate in the sense of conventional optimization. It gives convergence results in the sense of conventional optimization. It deals with problems that's considered interesting in the sense of conventional optimization.

Those assumptions include strongly convex in some form for the objective function and the uniqueness of the solution, among others. However, for the original statistical problems, these assumptions are hard to verify or invalid. For example, strong convexity type condition is hard to verify and always violated in statistical problems, and solutions to the optimization problem can be multiple in over-parametrized settings like neural network and robust Principle Component Analysis (RPCA).

One of our key observations is that these assumptions are not necessary for producing statistically well behaved computed estimators, as we do not need to solve the optimization problem well in the conventional way to guarantee its statistical performance — there is an alternative way of characterizing how well the optimization problem is solved in terms of

solution’s statistical performance. Further, solving it well in the conventional way does not give additional help to statistical analysis.

The convergence results in the sense of conventional optimization are also not enough for statistical consideration. In high dimensional statistics, we are essentially dealing with a class of optimization problems with changing dimensions, and we need to know how optimization-induced statistical error changes in terms of both iteration number and the dimension. Conventional optimization results usually view dimension related quantity as a constant intrinsic to the optimization problem.

Many statistically rooted optimization problems are not considered general enough or interesting enough under conventional optimization sense, but the statistical problems are important from statistical perspective. Therefore, many heuristic optimization methods widely used in statistical literature are not nearly well understood. Many statistically good estimators also lack optimization algorithms. And some optimization results targeting statistically rooted optimization problem generalize the problem in the way making it no longer useful for the root statistical problem.

Our approach is free from all these problems. We propose a framework consists of three parts. We incorporate the consideration of optimization error into the statistical analysis through an approximate optimization problem rather than an approximate optimization solution. We provide a template optimization algorithm. We show its convergence in terms of converging to the optimization problem. Our convergence results takes the possibly growing dimension and other changing geometry quantities into consideration in addition to the iteration number. All three added together, we have a theoretically guaranteed algorithm and a precise quantification of statistical accuracy given iteration number.

In two examples, 1-bit matrix completion (Davenport et al., 2014) and causal inference for panel data (Athey et al., 2021), we apply our framework, which yields novel results for both problems. And our framework can also be applied to network analysis, robust

principle analysis, kernel ridge regression, SVM, simple neural networks, LASSO, etc. We take LASSO for an example. LASSO in (high dimensional sparse) linear regression is a simpler and degenerate case for our framework. Through it, we show that our framework automatically adapt to the setting where stronger assumptions are satisfied (e.g. restricted strong convexity).

In addition to our framework, our statistical analysis of causal inference for panel data using matrix completion is also sharper and yields better statistical convergence rate in the special case that the solution is perfect, which is the case considered in the literature.

#### 4.1.1. Our framework

Our framework deals with statistical problems where the most promising estimator can be written as a solution to an optimization problem of the form

$$\begin{aligned} \min_X \quad & f(X) + g(X) \\ \text{s.t.} \quad & X \in C_1 \cap C_2 \cap \cdots C_J, \end{aligned} \tag{4.1.1}$$

where  $X$  is an  $m \times n$  parameter matrix, with vector being a special case by taking  $n = 1$ ,  $f$  is an  $L(\epsilon)$ -smooth (optimization wise) and  $L_f(\epsilon)$ -Lipschitz convex function on the constraint set and its  $\epsilon$  neighborhood (with  $L(\epsilon), L_f(\epsilon) > 0$ ),  $g$  is a possibly non-smooth but  $L_g(\epsilon)$ -Lipschitz convex function on the same area (with  $L_g(\epsilon) > 0$ ),  $C_1$  to  $C_J$  are convex constraint sets that are easy to project on. Note that  $f$  and  $g$  here are usually data dependent.

In some cases  $f$  is data dependent. Examples include negative log likelihood, sum of least squares in high-dimensional linear regression, or the objective function in principle component analysis (PCA). In these cases  $g$  can be penalty term or 0. In some cases,  $g$  is data-dependent and  $f$  is the regularization term. Examples include soft support vector machine and neural network with Relu activation function.

So this general form includes a wide range of estimators, including constrained maximum log

likelihood estimators, penalized maximum log likelihood estimator, support vector machine, etc.. This wide range of estimators have proved their power by achieving minimax optimality for many statistical problems, especially in high dimensional statistics, or by achieving good empirical performances, especially in machine learning.

Note that we do not require strong convexity, restricted strong convexity or strong convexity of any form for  $f(X)$ , which is almost a conventional assumption in the literature considering both optimization and statistical properties. We will see later that the absence of strong convexity is indeed very common in reality.

A specific example fitting this general form is the 1-bit matrix completion with constrained maximum log likelihood estimators. It's helpful to see how this concrete example fits the general framework.

*Example 4.1.1* (1-bit matrix completion). The statistical setting for 1-bit matrix completion is as follows (Davenport et al., 2014). Given the true parameter matrix  $M \in \mathbb{R}^{d_1 \times d_2}$ , a random subset of indices  $\Omega \subset [d_1] \times [d_2]$  indicating the elements we observe, and a differentiable link function  $l : \mathcal{D} \rightarrow [0, 1]$ , where  $\mathcal{D} \subset \mathbb{R}$ , the observation is a matrix  $Y \in \mathbb{R}^{d_1 \times d_2}$  defined as follows. Entries of  $Y$  are independent.

For  $(i, j) \in \Omega$ ,

$$Y_{i,j} = \begin{cases} +1 & \text{with probability } l(M_{i,j}) \\ -1 & \text{with probability } 1 - l(M_{i,j}) \end{cases}. \quad (4.1.2)$$

For  $(i, j) \notin \Omega$ ,  $Y_{i,j} = 0$ . The assumptions are as follows.  $M$  is nuclear norm bounded ( $\|M\|_* \leq \alpha \sqrt{rd_1 d_2}$ ) and element wise bounded ( $\|M\|_\infty \leq \alpha$ ). The random subset of indices satisfies  $\mathbb{E}|\Omega| = n$  with each entry being chosen with probability  $\frac{n}{d_1 \times d_2}$  independently.

Then the log-likelihood function of this problem is

$$\mathcal{L}_{\Omega, Y}(X) = \sum_{(i,j) \in \Omega} (\mathbb{1}\{Y_{i,j} = 1\} \log(l(X_{i,j})) + \mathbb{1}\{Y_{i,j} = -1\} \log(1 - l(X_{i,j}))). \quad (4.1.3)$$

Davenport et al. (2014) show that the minimax optimal estimator  $\hat{M}$  is a solution of the following optimization problem

$$\begin{aligned} \min_X \quad & -\mathcal{L}_{\Omega,Y}(X) \\ \text{s.t.} \quad & \|X\|_* \leq \alpha\sqrt{rd_1d_2} \text{ and } \|X\|_\infty \leq \alpha. \end{aligned} \tag{4.1.4}$$

If we further assume twice differentiability of the link function, which is true for all link function examples in Davenport et al. (2014), this estimator satisfies our general formulation (4.1.1), with

$$\begin{aligned} f(X) &= -\mathcal{L}_{\Omega,Y}(X), \quad g(X) = 0, \\ C_1 &= [-\alpha, \alpha]^{d_1 \times d_2}, \quad C_2 = \{M \in \mathbb{R}^{d_1 \times d_2} \mid \|M\|_* \leq \alpha\sqrt{rd_1d_2}\}, \\ L_f(\epsilon) &= \sup_{|x| \leq \epsilon + \alpha} \frac{|l'(x)|}{l(x)(1-l(x))}, \quad L_g(\epsilon) = 0, \text{ and} \\ L(\epsilon) &= \sup_{|x| \leq \epsilon + \alpha} \max\left\{ \frac{|l''(x)l(x) - (l'(x))^2|}{l(x)^2}, \frac{|l''(x)(1-l(x)) + (l'(x))^2|}{(1-l(x))^2} \right\}. \end{aligned} \tag{4.1.5}$$

*Remark 4.1.1.* Note that in Example 4.1.1,  $-\mathcal{L}_{\Omega,Y}(X)$  in most cases is not strongly convex, or restricted strongly convex (Loh and Wainwright, 2015; Wang et al., 2017), hence the approach of establishing convergence in parameter space (the space of  $X$ ) for the optimization problem separated from the statistical problem, which is adopted in the literature, is not a good, if possible, approach.

*Remark 4.1.2.* In the original work by Davenport et al. (2014), where Example 4.1.1 arises, they only have a heuristic algorithm computing the solution of optimization problem (4.1.4) with no theoretical guarantee.

*Remark 4.1.3.* Causal inference for panel data (Athey et al., 2021) also satisfies the general formulation (4.1.1). We discuss it in detail in Section 4.4, where we not only develop a theoretically guaranteed optimization algorithm and provide a precise quantification of how iteration number comes in the statistical accuracy based on our framework, but also give a sharper upper bound on statistical accuracy than that in Athey et al. (2021) when the

solution is exact, all of which are interesting results on their own.

*Remark 4.1.4.* Lasso for linear regression is another example satisfying our framework. But it is a severely degenerate case: it is for parameter vector; it does not have constraints; it admits restricted strong convexity in high dimensional sparse setting. We discuss it in detail in Section 4.5.

*Remark 4.1.5.* More examples fit into our framework. For the reason of space, we do not give detailed discussion in this dissertation.

In our framework, to be free from strong convexity of any form or other artificial conditions, we consider  $\tilde{X}$  that has small violations on both constraints and minimum objective function value. We analyze statistical property of  $\tilde{X}$ . The analysis of  $\tilde{X}$  is independent from any optimization procedure and it is an interface between statistical property and optimization error, so we call this step *Statistical-Optimization Interplay*. Then we develop an optimization algorithm and analyze its convergence in terms of those small vanishing violations. Therefore, we can give a precise quantification of how number of iterations in our algorithm translates into statistical accuracy. Given that the number of iteration is the key bottleneck for running time and can not be reduced through parallel computing, this shows how running time could buy statistical accuracy until the statistical limit is reached.

### **Statistical-Optimization Interplay**

The first step of our framework is to integrate the optimization error into statistical analysis before solving the optimization problem.

Given the data, functions  $f$  and  $g$  in optimization problem (4.1.1) are decided. The target estimator  $X^*$  is a solution to the optimization problem (4.1.1). But the exact solution of optimization problem (4.1.1) can be computationally infeasible and only approximate solutions can be computed. We need to consider the statistical property of this approximate solution.

Instead of considering the convergence rate of the computed solution  $\tilde{X}$  to the target es-



timator  $X^*$ , we move the consideration of optimization to the start of statistical analysis. We consider an approximate estimator  $\tilde{X}$  satisfying the approximate conditions in (4.1.6) and investigate its statistical properties. Basically, approximate conditions mean both constraints and optimal objective function can be violated a little ( $\delta, \delta_0, \delta_1, \dots, \delta_J \geq 0$ ).

$$\begin{aligned} f(\tilde{X}) + g(\tilde{X}) &\leq f(X^*) + g(X^*) + \delta, \\ \inf_{Z \in C_i} \|Z - \tilde{X}\|_2 &\leq \delta_i, \text{ for all } 1 \leq i \leq J, \\ \inf_{Z \in C_1 \cap C_2 \cap \dots \cap C_J} \|Z - \tilde{X}\|_2 &\leq \delta_0. \end{aligned} \tag{4.1.6}$$

Note that the target estimator  $X^*$ , the optimizer of Optimization Problem (4.1.1), satisfies these inequalities with  $\delta = \delta_0 = \dots = \delta_J = 0$ . When  $\delta, \delta_0, \dots, \delta_J \rightarrow 0^+$ , the approximate conditions are infinitely close to the original Optimization Problem (4.1.1), and when  $\delta = \delta_0 = \dots = \delta_J = 0$ , the approximate conditions define an equivalent optimization problem as the original one. So this is a way of characterizing how close the computed estimator  $\tilde{X}$  is to the target estimator  $X^*$ . An interesting observation is that the statistical analysis of, or the tools used in the statistical analysis of most constrained  $M$ -estimators, a kind of estimators satisfying the conditions of our framework, can be carried to this approximate version estimator relatively easily. We concrete the idea in three examples, 1-bit matrix completion, causal panel data analysis and LASSO. 1-bit matrix completion problem is analyzed as a representation for constrained log-likelihood estimator. Causal panel data is analyzed as a representation for constrained penalized log-likelihood estimator. Lasso is a representation of a degenerate case for our framework, where we show that the statistical-optimization interplay automatically adapt to simpler settings to give strong results in the simpler setting. For causal panel data, we also sharpened the backbone statistical analysis. And our framework is applied to the sharpened statistical analysis.

Note that in this step, we do not yet have an optimization procedure and the analysis is entirely irrelevant to the optimization procedure. Yet the slightly violated conditions fully characterize the statistical property of computed solution  $\tilde{X}$  in the sense that non-violated

version conditions are the starting point for any statistical analysis for the exact solution. So we do not need the optimization procedure to have a traditional optimization convergence.

Existing work on considering both optimization error and statistical error (e.g. Bottou and Bousquet (2011); Loh and Wainwright (2015)) usually considers the optimization error after the statistical problem is fully analyzed. They consider the optimization wise convergence rate of the computed solution to the true solution. But this approach does not work when the true solution is hard or unable to computed well. One of such setting is when the optimization problem has multiple solutions. Examples include neural network, which is usually over-parametrized, and principal component analysis (PCA). People deals with the problem of multiple solution in PCA through defining a distance that implicitly equalize the solutions, partly leading to a huge volume of literature on non-convex optimization, see Chi et al. (2019) for a review. Another situation that the true solution is hard to be computed well is when the objective function does not enjoy good properties in the sense of optimization, e.g. strong convexity of some form.

### **Optimization Algorithm and Convergence Analysis**

The second step is to develop an optimization procedure with theoretical guarantees in terms of convergence to an estimator satisfying inequalities (4.1.6).

We adopt a double-loop optimization procedure where the outer loop is proximal gradient descent and the inner loop is 3-block ADMM.

We give convergence rate of the optimization procedure that considers both iteration number and statistically important quantity (e.g. dimension). This includes the convergence rate for inexact proximal gradient descent, convergence rate for our inner loop (3-block ADMM), and a bound for a dimension-related geometric quantity involved in the convergence rate.

There can be variants to our optimization procedure ( e.g. using accelerated proximal gradient descent for outer loop, using 2-block ADMM for inner loop when reducible). But

our analysis for outer loop can be easily carried to accelerated version. And our analysis of the geometric quantity can also be easily carried to 2-block ADMM. Another reason for taking 3-block ADMM is that in addition to fitting our two examples, the 3-block ADMM can serve as a building block for a general number of constraints, as is in our general framework.

#### 4.1.2. Related Literature

In addition to the literature mentioned at the beginning of this sections. The problems considered in this paper is also connected to the following problems and literature.

Computational issue for low-rank matrix completion has been studied through a matrix factorization approach which leads to nonconvex optimization problem. See, for example, Wang et al. (2017); Jain et al. (2013); Chen and Wainwright (2015), and the overview paper, Chi et al. (2019). In this line of research, 1-bit matrix completion problem is also correctly studied by Chen and Wainwright (2015). However, this approach requires the exact low rank assumption, the knowledge of the rank, and also at least one other conditions like RIP condition (Jain et al., 2013), restricted convexity (Wang et al., 2017), and incoherence Chen and Wainwright (2015), which are strong and hard-to-verify condition in many settings. Further, the convergence rate for 1-bit matrix problem in Chen and Wainwright (2015) depends on the mostly unknown incoherence, which varies greatly, and the worst case different from the best by order.

Computational issue for  $M$ -estimator is also considered in Loh and Wainwright (2015), where they consider Lasso type estimator. Their work deal with vectors (instead of matrices) with restricted strong convexity (RSC) requirement. Our framework is designed for the more general case: matrix without RSC condition. This includes the simpler setting (vector with RSC condition). And as shown in our third example, our framework automatically adapts to the simpler setting and provides stronger results under stronger conditions.

Schmidt et al. (2011) studied convergence rate for inexact proximal gradient and inexact

accelerated proximal gradient when the non-smooth part is finite. But in our setting, the existence of constraints dictates the infinity of the non-smooth part. Jiang et al. (2012) studied inexact accelerated proximal gradient descent, but it is for linearly constrained convex SDP.

Literature on the convergence of 3-block ADMM includes, for example, Cai et al. (2017); Lin et al. (2018); Hong and Luo (2017); Lin et al. (2016). But they either are not applicable to our setting (Hong and Luo, 2017; Lin et al., 2018), or establishes convergence rate on Lagrange Functions (Cai et al., 2017), or establishes convergence rate on objective function with results weaker than ours in its applicable setting (Lin et al., 2016). Tibshirani (2017) considers projection on intersection of convex sets, but it is for coordinate descent and for vectors instead of matrices, thus not applicable to our setting.

#### 4.1.3. Organization of the Chapter

In Section 4.2 we introduce our general framework and give general results. In Section 4.3 we discuss the results of applying our framework to 1 bit matrix completion example, where we get interesting new results. In section 4.4 we discuss the results for causal panel data example, where in addition to applying our framework we provide tighter back-bone statistical analysis. In Section 4.5, we discuss applying our framework to (high dimensional) linear regression and compare with the results in literature for this degenerated setting. In section 4.6, we discuss some directions for future work. For the reason of space, the proofs are given in the appendix Section A.6.

#### 4.1.4. Notation and Definition

Both  $\|\cdot\|$  and  $\|\cdot\|_F$  stand for Frobenious norm.  $\|\cdot\|_F$  is to give special emphasis for matrices when there might be confusion.  $\|\cdot\|_*$  stands for nuclear norm. We use  $|\mathcal{O}|$  to denote the number of elements in  $\mathcal{O}$  when  $\mathcal{O}$  is a set. We use  $D(A\|B) = \frac{1}{d_1 d_2} \sum_{i,j} D(A_{i,j}\|B_{i,j})$  to denote average KL divergence between  $d_1$  by  $d_2$  probability matrix A and B for 1-bit

matrix completion, where  $D(a\|b) = a \log(\frac{a}{b}) + (1-a) \log(\frac{1-a}{1-b})$ . We use  $\mathfrak{T}\{A\}$  to denote the function where it takes 0 if  $A$  holds and  $\infty$  if  $A$  does not hold. We use  $\mathcal{R}(\varepsilon, C)$  to denote the  $\varepsilon$  neighborhood of convex set  $C$ :  $\mathcal{R}(\varepsilon, C) = \{X : \inf_{Z \in C} \|X - Z\| \leq \varepsilon\}$ . We use  $B_d(x)$  to denote a ball centered at  $x$  with radius  $d$  under Frobenious norm. We use  $\vee$  to denote taking max:  $a \vee b = \max\{a, b\}$ . We use  $\text{Proj}_C(P)$  to denote the projection point of  $P$  on convex set  $C$ , the projection is in terms of Euclidean distance.

Now we introduce the definition of smoothness in optimization sense.

**Definition 4.1.1** (Optimization-wise Smoothness). *A convex function  $f(X)$  is said to be  $L$ -smooth if for any  $X$  in the domain, there is a subgradient  $\partial f(X)$  at  $X$  such that for all  $Y$  in the domain,*

$$f(Y) \leq f(X) + \partial f(X)(Y - X) + \frac{L}{2} \|X - Y\|^2. \quad (4.1.7)$$

## 4.2. General Framework

In this section, we introduce the general framework. The general framework has three parts: statistical-optimization interplay, optimization-template algorithm, and optimization convergence analysis.

### 4.2.1. Statistical-Optimization Interplay

In statistical-optimization interplay, we integrate the optimization consideration into the statistical analysis by considering the statistical accuracy of an estimator coming from an approximate optimization problem instead of just an approximate solution.

Recall that the target estimator  $X^*$  is the solution in (4.1.1). To consider the optimization-induced statistical error, we consider the statistical property of approximate estimator  $\tilde{X}$  satisfying Inequalities (4.1.6). The measurements for how well the optimization problem is eventually solved are  $\delta, \delta_0, \delta_1, \dots, \delta_J$ .

Suppose one of the true parameters of the statistical model is  $X_t$ .

The key ingredient for statistical-optimization interplay is an interesting but natural observation on statistical analysis of estimator of the form (4.1.1). The statistical analysis for  $X^*$  usually starts with the inequality

$$f(X^*) + g(X^*) \leq f(X_t) + g(X_t). \quad (4.2.1)$$

This inequality is usually reduced to simpler form with or without using the constraint conditions. And then the simpler form becomes a solvable inequality for the statistical error or the simpler form is further reduced. Typical tools for further reducing the inequality includes empirical process, which is also where the constraints in (4.1.1) usually comes in.

A reflection on this whole procedure gives that the additive nature of (4.2.1) is untouched, so are the constraints in (4.1.1).

These characteristics of the analysis mean that similar analysis can go through for approximate solution  $\tilde{X}$ , as it adds in the optimization errors (e.g.  $\delta, \delta_0, \dots, \delta_j$ ) in an additive way. Specifically, the analysis for  $\tilde{X}$  starts with

$$f(\tilde{X}) + g(\tilde{X}) \leq f(X_t) + g(X_t) + \delta. \quad (4.2.2)$$

Constraints also enter the analysis with an additional error term.

In this way, the focus is shifted from the final approximate solution  $\tilde{X}$  to the approximate optimization problem (4.1.6). We do not need strong convexity or uniqueness of the solution or other conditions to ensure the fast proximity of the solutions. We only need proximity of the problems, which is the only thing relevant to the statistical accuracy while being much relaxed in terms of optimization.

As statistical analysis varies from problem to problem. We will concrete the idea of analyzing solution satisfying the approximate optimization problem through examples in Section 4.3, Section 4.4 and Section 4.5.

*Remark 4.2.1.* In our framework, we consider problems with constraints, but it is also applicable to the setting where there is no constraints. The problem with no constraints is a degenerated case where we do not need to consider projection in optimization part. We show in Section 4.5 that in a degenerate case, (high dimensional) linear regression, our framework automatically adapts to the simpler setting and stronger conditions to give stronger results.

*Remark 4.2.2.* Statistical-Optimization Interplay, the interface building optimization error into statistical analysis before solving the optimization problem, can work alone. That is, the optimization procedures and analysis can be replaced when needed. Further, the statistical-optimization interplay can also be extended to Z-estimators and other type of estimators coming from equation/inequality system, which is in my future work.

#### 4.2.2. Template Algorithm

The second step of the framework is to have an algorithm finding  $\tilde{X}$  satisfying (4.1.6). Our template algorithm is a double-loop algorithm, where the outer loop is inexact proximal gradient descent and inner loop is a 3-block ADMM to approximately compute quantities in the outer loop. Our inner loop algorithm can be replaced and generalized to fit arbitrary number of constraints, but to avoid unnecessary complexity while being sufficient for our examples, we elaborate on 3-block ADMM and remark on generalized algorithm.

##### Outer Loop

Note that optimization problem (4.1.1) is equivalent to minimizing the following function.

$$F(X) = f(X) + (g(X) + \mathfrak{T}\{X \in C_1\} + \mathfrak{T}\{X \in C_2\} + \cdots + \mathfrak{T}\{X \in C_J\}). \quad (4.2.3)$$

To minimize  $F(X)$ , we do proximal gradient descent but with an “approximate” proximal step, as shown in algorithm 4.2.1.

*Algorithm 4.2.1* (Outer Loop: Inexact Proximal Gradient Descent). Starting point is  $X_0 \in C_1 \cap C_2 \cap \dots \cap C_J$ . Step size is  $\eta > 0$ . For  $k \geq 0$ ,

$$X_{k+0.5} = X_k - \eta \nabla f(X_k), \quad X_{k+1} = \widetilde{\text{Prox}}_{\eta(g(X) + \mathfrak{T}\{C_1 \cap C_2 \cap \dots \cap C_J\})}(X_{k+0.5}), \quad (4.2.4)$$

where  $\widetilde{\text{Prox}}_{\eta(g(X) + \mathfrak{T}\{X \in C_1 \cap C_2 \cap \dots \cap C_J\})}(X_{k+0.5})$  is a close approximation of

$$\begin{aligned} & \text{Prox}_{\eta(g(X) + \mathfrak{T}\{X \in C_1 \cap C_2 \cap \dots \cap C_J\})}(X_{k+0.5}) = \\ & \arg \min_X \left( \frac{1}{2} \|X - X_{k+0.5}\|_F^2 + \eta \left( g(X) + \mathfrak{T}\{X \in C_1 \cap C_2 \cap \dots \cap C_J\} \right) \right). \end{aligned} \quad (4.2.5)$$

$\text{Prox}_{\eta(g(X) + \mathfrak{T}\{X \in C_1 \cap C_2 \cap \dots \cap C_J\})}(\cdot)$  is called a *proximal operator*. However, we do not have an exact solution to (4.2.5) to give  $\text{Prox}_{\eta(g(X) + \mathfrak{T}\{X \in C_1 \cap C_2 \cap \dots \cap C_J\})}(X_{k+0.5})$ . We only have an approximate proximal  $\widetilde{\text{Prox}}_{\eta(g(X) + \mathfrak{T}\{C_1 \cap C_2 \cap \dots \cap C_J\})}(X_{k+0.5})$  in the outer loop by approximately solving the optimization problem corresponding to it, which is our inner loop.

Before we proceed to inner loop, we conclude with a remark that the approximate proximal gradient can be replaced by its accelerated version for outer loop. But given the commonly seen phenomenon that accelerated version of algorithms are usually less robust to errors along the computation, we do not discuss the accelerated version for our setting. Similar discussion can be given for accelerated version.

### Inner Loop

The optimization problem that inner loop aims to solve is

$$\min_X \left( \frac{1}{2} \|X - X_{k+0.5}\|_F^2 + \eta \left( g(X) + \mathfrak{T}\{X \in C_1 \cap C_2 \cap \dots \cap C_J\} \right) \right). \quad (4.2.6)$$

We can write it as

$$\min_P \left( \|P - P_0\|_F^2 + \left( h_1(P) + h_2(P) + \dots + h_m(P) \right) \right), \quad (4.2.7)$$



where  $P_0$  equals to  $X_{k+0.5}$  in (4.2.6), and  $h_i(\cdot)$  are convex functions not necessarily smooth and potentially take infinity value. In the case  $J \geq 1$ , at least one  $h_i(\cdot)$  takes infinity value.

We first consider the case that  $m = 2$ . In this case, Optimization Problem (4.2.7) is equivalent to the following problem:

$$\begin{aligned} \min_{W, Z, P} & \|P - P_0\|_F^2 + h_1(W) + h_2(Z), \\ \text{s.t.} & W = P, Z = P. \end{aligned} \quad (4.2.8)$$

We take 3-block ADMM to solve this problem. The Augmented Lagrange Function for this 3-block ADMM is

$$\mathcal{L}_\beta(W, Z, P, \Lambda_1, \Lambda_2) = \|P - P_0\|_F^2 + h_1(W) + h_2(Z) + \frac{\beta}{2}(\|W - P + \frac{\Lambda_1}{\beta}\|_F^2 + \|Z - P + \frac{\Lambda_2}{\beta}\|_F^2), \quad (4.2.9)$$

where  $\beta > 0$  is the dual step size and  $\Lambda = (\Lambda_1, \Lambda_2)$  is the dual variable.

The optimization procedure for this 3 block ADMM is in algorithm 4.2.2.

*Algorithm 4.2.2* (Inner loop: 3 block ADMM). The starting points are  $P^0 = P_0$ ,  $\Lambda_1^0 = \mathbf{0}$ ,  $\Lambda_2^0 = \mathbf{0}$ . The dual step size is  $\beta > 0$ . For  $k \geq 0$ , the iteration steps are

$$\begin{aligned} W^{k+1} &= \arg \min_W \mathcal{L}_\beta(W, Z^k, P^k, \Lambda_1^k, \Lambda_2^k) = \arg \min_W h_1(W) + \frac{\beta}{2} \|W - P^k + \frac{\Lambda_1^k}{\beta}\|_F^2, \\ Z^{k+1} &= \arg \min_Z \mathcal{L}_\beta(W^k, Z, P^k, \Lambda_1^k, \Lambda_2^k) = \arg \min_Z h_2(Z) + \frac{\beta}{2} \|Z - P^k + \frac{\Lambda_2^k}{\beta}\|_F^2, \\ P^{k+1} &= \arg \min_P \mathcal{L}_\beta(W^k, Z^k, P, \Lambda_1^k, \Lambda_2^k) \\ &= \arg \min_P \|P - P_0\|_F^2 + \frac{\beta}{2} (\|W^{k+1} - P + \frac{\Lambda_1^k}{\beta}\|_F^2 + \|Z^{k+1} - P + \frac{\Lambda_2^k}{\beta}\|_F^2), \\ \Lambda_1^{k+1} &= \Lambda_1^k + \beta(W^{k+1} - P^{k+1}), \\ \Lambda_2^{k+1} &= \Lambda_2^k + \beta(Z^{k+1} - P^{k+1}). \end{aligned} \quad (4.2.10)$$

Note that when  $h_1(\cdot)$  comes from a constraint function, the update step for  $W$  is a projection

step. Analogous result holds for  $h_2(\cdot)$ .

Usually, 3-block ADMM is enough for solving most of the problems encountered in statistics, including our two examples, as  $m$  in (4.2.7) is usually not very large. In the case that 3-block ADMM is not enough (i.e.  $m \geq 2$ ), the reason for  $m \geq 2$  is that the number of constraints is large. Then a natural way is to do recursive ADMM. For example, if we have  $g = 0$  and four constraints  $C_1, C_2, C_3, C_4$ , we can do a 3-block ADMM for  $\arg \min_P \|P - P_0\|_F^2 + \mathfrak{T}\{P \in C_1 \cap C_2\} + \mathfrak{T}\{P \in C_3 \cap C_4\}$ , where in each projection step, say on  $C_1 \cap C_2$ , we can do another 3-block ADMM. This could be costly, but do-able.

Another remark is that in some cases, Optimization Problem (4.2.7) can be reduced to 2-block ADMM. But less blocks sometimes may lead to worse performance (Lin et al., 2018) and it's not generalizable to more blocks, we rest with 3-block ADMM.

#### 4.2.3. Optimization Convergence Analysis

In this section we give theoretical analysis for the algorithm-template we introduced in Section 4.2.2.

For outer loop, we have the convergence result for inexact proximal gradient descent in Theorem 4.2.1.

**Theorem 4.2.1** (Inexact Proximal Gradient Descent). *We take the inexact proximal gradient descent algorithm 4.2.1. Suppose the inner loop (approximation of the proximal) satisfies*

$$\left| \widetilde{Prox}_{\eta(g(x)+\mathfrak{T}\{x \in C_1 \cap C_2 \cap \dots \cap C_J\})}(X) - Prox_{\eta(g(x)+\mathfrak{T}\{x \in C_1 \cap C_2 \cap \dots \cap C_J\})}(X) \right| \leq \delta_0 \quad (4.2.11)$$

*for all  $X \in \mathcal{R}(\delta_0, C_1 \cap C_2 \cap \dots \cap C_J)$ . Suppose on  $\mathcal{R}(\delta_0, C_1 \cap C_2 \cap \dots \cap C_J)$ ,  $f$  is  $L$  smooth and  $L_f$  Lipschitz, and  $g$  is  $L_g$  Lipschitz. We let step size  $\eta \leq \frac{1}{L}$ . Suppose  $\tilde{X}$  has the smallest  $f(X) + g(X)$  value among  $X_0, X_1, \dots, X_K$ , the starting point and the results of first  $K$*

iterations. Then we have

$$\begin{aligned} f(\tilde{X}) + g(\tilde{X}) - f(X^*) - g(X^*) &\leq \frac{1}{2K\eta} \|X_0 - X^*\|^2 + \\ &\quad (L_f + L_g)\delta_0 + \frac{L}{2}\delta_0^2 + \frac{\delta_0 D}{\eta} + \frac{\delta_0^2}{2\eta}, \end{aligned} \quad (4.2.12)$$

where  $D$  is the diameter of  $C_1 \cap C_2 \cap \dots \cap C_J$ .

*Remark 4.2.3.* Schmidt et al. (2011) studied the convergence of inexact proximal gradient descent when the non-smooth part is finite. But in the presence of constraints, although function  $g$  is finite, the optimization problem (4.2.7) in our setting is always infinite.

*Remark 4.2.4.* The Lipschitz conditions needed for  $f$  and  $g$  are natural conditions satisfied in most setting. For most non-smooth penalties,  $g$  satisfies Lipschitz condition on the entire space. For most problems, the constraint set is compact (or contained in a compact set), thus the smoothness and convexity of  $f$  dictates Lipschitz condition.

Now we turn to the convergence of the inner loop, 3-block ADMM.

Let  $W^*, Z^*, P^*$  be true primal variables and  $\Lambda^* = (\Lambda_1, \Lambda_2)$  be the true dual variable, i.e. solution to the optimization problem

$$\max_{\Lambda_1, \Lambda_2} \min_{W, Z, P} \mathcal{L}_\beta(W, Z, P, \Lambda_1, \Lambda_2).$$

We have Proposition 4.2.1 for the convergence rate of the 3-block ADMM.

**Proposition 4.2.1** (3 block ADMM convergence rate). *Suppose we take algorithm 4.2.2 with dual step size  $\beta \leq \frac{6}{17}$ , suppose  $\bar{P}^t = \frac{1}{t} \sum_{j=1}^t P^j$ , then we have*

$$\|\bar{P}^t - P^*\|^2 \leq \frac{1}{2\beta t} \left( \beta^2 \|Z^1 - P^*\|^2 + 2\beta^2 \|P^1 - P^*\|^2 + \|\Lambda^1 - \Lambda^*\|^2 + \frac{20}{3}\beta^2 \|P^1 - P_0\|^2 \right). \quad (4.2.13)$$

*Remark 4.2.5.* In general, convergence for multi-block ADMM with more than two blocks does not hold (Chen et al., 2016). Convergence in some specific settings has been studied. But to our knowledge, no convergence rate has been established for direct 3-block ADMM applied to our setting. In the most closely related literature, Cai et al. (2017) does not

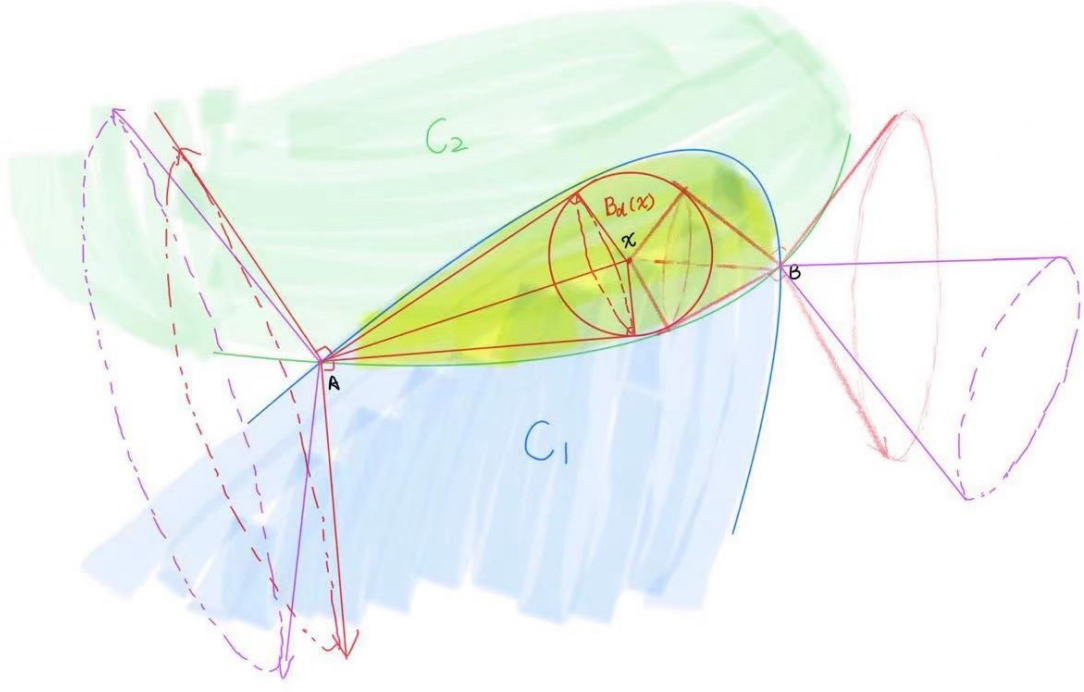


Figure 4.1: Illustration of geometry of dual variable

have convergence rate; the requirement on constraints in Lin et al. (2018) or Hong and Luo (2017) does not fit our setting; Lin et al. (2016) has strict requirement on dual step size and slower rate based on their requirement.

Note that  $\|\mathbf{\Lambda}^1 - \mathbf{\Lambda}^*\|$  is involved in the convergence rate.  $\mathbf{\Lambda}^1$  depends explicitly on  $\beta$ ,  $P_0$ ,  $h_1(\cdot)$  and  $h_2(\cdot)$ , which can usually be easily studied and bounded, and it's usually relatively small in our setting.  $\mathbf{\Lambda}^*$ , however, can be very large (in terms of norm) and depends implicitly on the geometry of  $h_1(\cdot)$  and  $h_2(\cdot)$ , which is dimension-dependent. But optimization literature does not deal with it, as it is considered as a constant for a single optimization problem. This issue is not particular to 3-block ADMM. 2-block ADMM also involves true dual variable in the convergence rate, which is treated as constant in the literature.

We bound  $\|\mathbf{\Lambda}^*\|$ , a geometry related quantity, by easy-to-compute geometry quantities.

To understand the involvement of geometry intuitively, figure 4.1 takes the projection on the intersection of two convex sets as an example for illustration. If the point to be taken projection on, say  $P_0$ , satisfies  $\text{Proj}_{C_1 \cap C_2}(P_0) = A$ , the number of iterations needed to get enough close to  $C_1 \cap C_2$  would be relatively large, as  $P_k$  can stay far from  $C_1 \cap C_2$  while it's already close to both  $C_1$  and  $C_2$  separately. On the other hand, when  $\text{Proj}_{C_1 \cap C_2}(P_0) = B$ , it would take less iterations to get enough close to  $B$ . Simple calculation show that  $-\Lambda_1^*$  and  $-\Lambda_2^*$  are subgradients for  $\mathfrak{T}\{X \in C_1\}$  and  $\mathfrak{T}\{X \in C_2\}$  at  $\text{Proj}_{C_1 \cap C_2}(P_0)$ , satisfying  $-\Lambda_1^* - \Lambda_2^* = 2(P_0 - P^*)$ . In figure 4.1, the purple cones at  $A$  and  $B$  show the region  $\Lambda_1^*$  and  $\Lambda_2^*$  can take value in at  $A$  and  $B$  respectively. We find bound for  $\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2$  by finding bound for “the maximum angle” the purple cones. The purple cone (smaller cone) at  $A$  can be considered as polar cone (Chandrasekaran and Jordan, 2013) of the smallest cone containing  $C_1 \cap C_2$  with  $A$  considered as origin, which at least contains the ball  $B_d(x)$ . Thus we can bound the purple cone by red cone. Same logic applies to purple cone (smaller cone) at  $B$ . Lemma 4.2.1 gives the precise description of this intuition.

**Lemma 4.2.1** (Geometry Bound). *We define the generalized polar cone of convex set  $C$  at point  $P$  to be  $N_C(P) = \{\mathbf{a} : \langle \mathbf{a}, P - x \rangle \geq 0 \text{ for all } x \in C\}$ . Define the maximum angle of two convex sets  $C_1$  and  $C_2$  to be*

$$\theta(C_1, C_2) = \sup_{P \in \partial(C_1 \cap C_2)} \sup_{\lambda_1 \in N_{C_1}(P), \lambda_2 \in N_{C_2}(P)} \arccos(\langle \lambda_1, \lambda_2 \rangle),$$

where  $\partial(C_1 \cap C_2)$  is the boundary of  $C_1 \cap C_2$ . We define a quantity based on maximum angle of  $C_1$  and  $C_2$  to be  $C(C_1, C_2) = \frac{1}{2 \cos^2(\frac{\theta(C_1, C_2)}{2})}$ , then we have

$$C(C_1, C_2) \leq \frac{D^2}{2d^2},$$

where  $D = \sup_{x, y \in C_1 \cap C_2} \|x - y\|_2^2$ ,  $d = \sup\{d : \exists x \in C_1 \cap C_2 \text{ such that } B_d(x) \subset C_1 \cap C_2\}$ . Further, suppose  $\Lambda^*$  and  $P^*$  are the true dual variable and primal variable of the Augmented Lagrange function (4.2.9). Then when  $h_1(W) = \mathfrak{T}\{W \in C_1\}$  and  $h_2(Z) = \mathfrak{T}\{Z \in C_2\}$ , we

have

$$\|\mathbf{\Lambda}^*\|_2^2 \leq \max\{4, 4C(C_1, C_2)\} \|P_0 - P^*\|^2.$$

#### 4.2.4. Remark

With the statistical-optimization interplay, algorithm-template, and optimization analysis, we are ready to provide theoretically guaranteed algorithm for a large class of estimator for a wide class of problems, and produce a precise analysis of how running time affects statistical accuracy.

### 4.3. Application to 1 Bit Matrix Completion

In this section we apply the framework introduced in Section 4.2 to the 1 bit matrix completion example we introduced in Section 4.1.1, which yields novel results and also further illustrates our framework.

#### 4.3.1. Statistical-Optimization Interplay

Suppose a solution to optimization problem (4.1.4) is  $X^*$ . The approximation optimization conditions (4.1.6) of the computed estimator  $\tilde{X}$  in 1 bit matrix completion setting becomes

$$\begin{aligned} -\mathcal{L}_{\Omega, Y}(\tilde{X}) &\leq -\mathcal{L}_{\Omega, Y}(X^*) + \delta, \\ \|\tilde{X}\|_\infty &\leq \alpha + \delta_1, \|\tilde{X}\|_* \leq \alpha\sqrt{rd_1d_2} + \delta_2, \inf_{Z \in C_1 \cap C_2} \|Z - \tilde{X}\|_2 \leq \delta_0, \end{aligned} \tag{4.3.1}$$

where  $C_1 = [-\alpha, \alpha]^{d_1 \times d_2}$  and  $C_2 = \{M \in \mathbb{R}^{d_1 \times d_2} \mid \|M\|_* \leq \alpha\sqrt{rd_1d_2}\}$ .

Our goal is to understand the statistical behavior of  $\tilde{X}$ . Applying statistical-optimization interplay step of our framework to the statistical analysis in Davenport et al. (2014), where  $X^*$  is  $\hat{M}$  and  $\tilde{X}$  is  $\tilde{M}$ , gives Theorem 4.3.1, which describes how optimization-induced error affects the statistical accuracy before solving the optimization problem.

**Theorem 4.3.1.** *Consider 1 bit matrix completion problem introduced in Example 4.1.1.*

Let  $\hat{M}$  be a solution to optimization problem (4.1.4). Suppose  $\tilde{M}$  satisfies  $-\mathcal{L}_{\Omega,Y}(\tilde{M}) \leq -\mathcal{L}_{\Omega,Y}(\hat{M}) + \delta$ ,  $\|\tilde{M}\|_* \leq \alpha\sqrt{rd_1d_2} + \delta_2$ ,  $\|\tilde{M}\|_\infty \leq \alpha + \delta_1$ . Recall that  $D(A\|B)$  is the average KL divergence between matrix  $A$  and  $B$ . Denote

$$L_\gamma = \sup_{|x| \leq \gamma} \frac{|l'(x)|}{l(x)(1-l(x))} \quad (4.3.2)$$

for  $\gamma > 0$  such that  $l(x) \in (0, 1)$  for  $|x| \leq \gamma$ . Then we have, with probability at least  $1 - \frac{c_1}{d_1+d_2}$ ,

$$D(l(M)\|l(\tilde{M})) \leq c_0 L_{\alpha+\delta_1} (\alpha\sqrt{rd_1d_2} + \delta_2) \sqrt{\frac{d_1+d_2}{nd_1d_2}} \sqrt{1 + \frac{(d_1+d_2) \log d_1d_2}{n}} + \frac{\delta}{n}, \quad (4.3.3)$$

$c_0, c_1$  are absolute constants that can be explicitly written out.

*Remark 4.3.1.* Note that in the formulation of example 4.1.1 we require link function  $l$  to be twice differentiable in addition to mere differentiability in the original work (Davenport et al., 2014) for fitting into our framework. But for statistical-optimization interplay, twice differentiability is not necessary, as Theorem 4.3.1 still holds with only differentiability.

*Remark 4.3.2.* Note that when  $\delta = 0$ ,  $\delta_1 = 0$  and  $\delta_2 = 0$ ,  $\tilde{M}$  in Theorem 4.3.1 is exactly the target estimator and the rate is of the same order with that in Davenport et al. (2014). In the view of approximate optimization, the target exact solution is a special case.

*Remark 4.3.3.* 1 bit matrix completion is a representative example for constrained M-estimator, or more precisely, constrained maximum likelihood estimator with no penalty term or optimization-wise smooth penalty term. Other constrained M-estimator includes constrained kernel ridge regression and constrained version of sparse principle component analysis.

#### 4.3.2. Optimization Algorithm

Note that in Davenport et al. (2014), they use a heuristic method without theoretical guarantee. Here we apply our optimization template algorithm to 1 bit matrix completion and give results on its convergence in terms of the approximate optimization conditions.

Note that the proximal operator  $\text{Prox}_{\eta(g(X) + \mathfrak{T}\{X \in C_1 \cap C_2 \cap \dots \cap C_J\})}(\cdot)$  in optimization template algorithm becomes projection operator  $\text{Proj}_{C_1 \cap C_2}(\cdot)$  for 1 bit matrix completion, which gives the outer loop in Algorithm 4.3.1.

*Algorithm 4.3.1* (1-bit Matrix Completion Outer Loop: Inexact Projected Gradient Descent). Starting point is  $X_0 = \mathbf{0}$ . Step size  $\eta > 0$ . For  $k \geq 0$ , the iteration steps are

$$X_{k+0.5} = X_k - \eta \nabla(-\mathcal{L}_{\Omega, Y}(X_k)), \quad X_{k+1} = \widetilde{\text{Proj}}_{C_1 \cap C_2}(X_{k+0.5}), \quad (4.3.4)$$

where  $\widetilde{\text{Proj}}_{C_1 \cap C_2}(X_{k+0.5})$  is a close approximation of *projection point*

$$\begin{aligned} \text{Proj}_{C_1 \cap C_2}(X_{k+0.5}) = \\ \arg \min_X (\|X - X_{k+0.5}\|_F^2 + \mathfrak{T}\{X \in C_1 \cap C_2\}). \end{aligned} \quad (4.3.5)$$

To compute approximate projection point  $\widetilde{\text{Proj}}_{C_1 \cap C_2}(P_0)$ , we apply the template algorithm inner loop. We know that the Augmented Lagrange Function for this 3-block ADMM is:

$$\begin{aligned} \mathcal{L}_\beta(W, Z, P, \Lambda_1, \Lambda_2) = & \mathfrak{T}\{W \in C_1\} + \mathfrak{T}\{Z \in C_2\} + \|P - P_0\|_F^2 + \\ & \frac{\beta}{2} (\|W - P + \frac{\Lambda_1}{\beta}\|_2^2 + \|Z - P + \frac{\Lambda_2}{\beta}\|_2^2), \end{aligned} \quad (4.3.6)$$

where  $\Lambda_1$  and  $\Lambda_2$  are dual variables and  $\beta$  is the dual update step size.

Applying the inner loop template algorithm, Algorithm 4.2.2, to 1 bit matrix completion, gives the inner loop steps for 1 bit matrix completion in Algorithm 4.3.2.

*Algorithm 4.3.2* (1-bit Matrix Completion Inner Loop: 3-block ADMM). The starting points



are  $P^0 = P_0, \Lambda_1^0 = \mathbf{0}, \Lambda_2^0 = \mathbf{0}$ . For  $k \geq 0$ , the iterative steps are

$$\begin{aligned} W^{k+1} &= \text{Proj}_{C_1}(P^k - \frac{1}{\beta}\Lambda_1^k), Z^{k+1} = \text{Proj}_{C_2}(P^k - \frac{1}{\beta}\Lambda_2^k), \\ P^{k+1} &= \frac{1}{\beta+1} \left( P_0 + \Lambda_1^k + \Lambda_2^k + \frac{\beta}{2}(W^{k+1} + Z^{k+1}) \right), \\ \Lambda_1^{k+1} &= \Lambda_1^k + \beta \left( W^{k+1} - P^{k+1} \right), \\ \Lambda_2^{k+1} &= \Lambda_2^k + \beta \left( Z^{k+1} - P^{k+1} \right). \end{aligned} \tag{4.3.7}$$

Take the average  $\bar{P}^k = \frac{1}{k} \sum_{i=1}^k P^i$  for the output if we end it at k-th iteration.

### 4.3.3. Optimization Convergence

In this section, we establish convergence rate for optimization algorithm introduced in Section 4.3.2 in terms of the approximate optimization conditions. We apply results in Section 4.2.3 to 1 bit matrix completion setting with appropriate modifications.

In this section, we need the assumption that the link function  $l$  for 1-bit matrix completion is twice differentiable, as introduced in section 4.1.1. So in addition to Lipschitz constant defined in (4.3.2), we have well defined smoothness constant for 1 bit matrix completion example, defined as

$$\tilde{L}_\gamma = \sup_{|x| \leq \gamma} \left\{ \frac{|l''(x)l(x) - (l'(x))^2|}{l(x)^2}, \frac{|l''(x)(1-l(x)) + (l'(x))^2|}{(1-l(x))^2} \right\}, \tag{4.3.8}$$

for  $\gamma > 0$  such that  $l(x) \in (0, 1)$  for  $|x| \leq \gamma$ .

For the convergence of the outer loop, we apply Theorem 4.2.1 to 1 bit matrix completion setting, which gives Proposition 4.3.1.

**Proposition 4.3.1** (Outer loop for 1 bit matrix completion). *Suppose we take projected gradient descent, Algorithm 4.3.1, for outer loop, and the projection error in all steps satisfies*

$$\|\widetilde{\text{Proj}}_{C_1 \cap C_2}(X) - \text{Proj}_{C_1 \cap C_2}(X)\| \leq \delta_0$$

. Suppose the link function  $l(x)$  is twice differentiable. Let  $\tilde{L}_\gamma$  be defined in (4.3.8). Suppose  $\tilde{L}_{\alpha+\delta_0} \leq L$ . Let  $L_\gamma$  be defined in (4.3.2). Let  $X^*$  be a solution of optimization problem (4.1.4). Take step size  $\eta = \frac{1}{L}$ , we have

$$\min_{0 \leq k \leq K} -\mathcal{L}_{\Omega, Y}(X_k) \leq -\mathcal{L}_{\Omega, Y}(X^*) + \frac{\alpha^2 L d_1 d_2}{2K} + \delta_0(2\alpha L \sqrt{d_1 d_2} + L_{\alpha+\delta_0} + L\delta_0). \quad (4.3.9)$$

To investigate the convergence for inner loop, we apply Proposition 4.2.1 and Lemma 4.2.1 in the general framework to 1 bit matrix completion example. Proposition 4.3.2 gives the convergence for inner loop for 1 bit matrix completion.

**Proposition 4.3.2** (Convergence of inner loop for 1 bit matrix completion). *Suppose  $P^* = \text{Proj}_{C_1 \cap C_2}(P_0)$ . Taking Algorithm 4.3.2, with dual step size  $\beta \leq \frac{6}{17}$ , we have*

$$\|\bar{P}^t - P^*\|^2 \leq \frac{1}{2\beta t} \left( 7\beta^2 + \max\{4, 8C(C_1, C_2)\} + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2} \right) \|P_0 - P^*\|^2, \quad (4.3.10)$$

where  $C(C_1, C_2) \leq \frac{d_1 d_2}{2}$ .

Combing the inner loop result, Proposition 4.3.2, and outer loop result, Proposition 4.3.1, we have that Theorem 4.3.2 showing the overall optimization convergence in terms of approximate conditions.

**Theorem 4.3.2** (Optimization: 1 bit matrix completion). *Suppose we take projected gradient descent, Algorithm 4.3.1, for outer loop, and 3-block ADMM, Algorithm 4.3.2, for inner loop, where  $P_0$  in the inner loop is  $X_{k+0.5}$  in the outer loop. Let  $L_\alpha$  be defined in (4.3.2). Let  $\tilde{L}_\alpha$  be defined in (4.3.8). If we take step size  $\eta = \frac{1}{2\tilde{L}_\alpha}$ , dual step size  $\beta \leq \frac{6}{17}$ , the number of iterations of inner loop  $t \geq t_0$ , and take  $T$  iterations for outer loop, then*

$\tilde{X} = \arg \min_{X \in \{X_0, X_1, \dots, X_T\}} -\mathcal{L}_\Omega(X)$  satisfies the approximate conditions (4.3.1) with

$$\begin{aligned} \delta &\leq \frac{\alpha^2 \tilde{L}_\alpha d_1 d_2}{T} + (4\alpha \tilde{L}_\alpha \sqrt{d_1 d_2} + 2L_\alpha) \sqrt{\frac{1}{t}} \sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} + 2\tilde{L}_\alpha \frac{1}{t} \left( q(\beta) + \frac{2d_1 d_2}{\beta} \right), \\ \max\{\delta_1, \delta_2, \delta_0\} &\leq \sqrt{\frac{1}{t}} \sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}}, \end{aligned} \tag{4.3.11}$$

where  $q(\beta) = \frac{7\beta}{2} + \frac{10}{3} \frac{\beta^3}{(\beta+1)^2}$ ,  $u_0 = \max\{u : L_{\alpha+u} \leq 2L_\alpha, \tilde{L}_{\alpha+u} \leq 2\tilde{L}_\alpha\}$ , and  $t_0 = \frac{1}{2\beta} \left( 7\beta^2 + 4d_1 d_2 + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2} \right) \left( 1 + \frac{L_\alpha}{u_0 \tilde{L}_\alpha} + \frac{L_\alpha}{\tilde{L}_\alpha} \right)^2$ .

#### 4.3.4. Overall Result

In this section, we are ready to show how the running time affects the statistical accuracy, as shown in Theorem 4.3.3.

**Theorem 4.3.3.** *For 1 bit matrix completion introduced in Section 4.1.1, suppose the link function  $l(x)$  is twice differentiable. Let  $\tilde{L}_\alpha$  be define in (4.3.8). Let  $L_\alpha$  be defined in (4.3.2). Suppose we take projected gradient descent, Algorithm 4.3.1, for outer loop with step size  $\eta = \frac{1}{2\tilde{L}_\alpha}$  and  $T$  iterations, and 3-block ADMM, Algorithm 4.3.2, for inner loop, where  $P_0$  in the inner loop is  $X_{k+0.5}$  in the outer loop. For inner loop, Algorithm 4.3.1, we take dual step size  $\beta \leq \frac{6}{17}$  and iteration number  $t \geq t_0$ , where  $t_0$  is specified later. Let  $\tilde{M}$  be among the starting point and resulting points in first  $T$  iterations of the outer loop,  $\{X_0, X_1, \dots, X_T\}$ , such that it has the smallest  $-\mathcal{L}_{\Omega, Y}(\cdot)$  value. Then with probability at least  $1 - \frac{c_1}{d_1+d_2}$ , we have*

$$\begin{aligned} &D(l(M) \| l(\tilde{M})) \\ &\leq 2c_0 L_\alpha \left( \alpha \sqrt{r d_1 d_2} + \sqrt{\frac{1}{t}} \sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} \right) \sqrt{\frac{d_1 + d_2}{n d_1 d_2}} \sqrt{1 + \frac{(d_1 + d_2) \log(d_1 d_2)}{n}} \\ &\quad + \frac{\alpha^2 \tilde{L}_\alpha d_1 d_2}{T n} + \frac{4\alpha \tilde{L}_\alpha \sqrt{d_1 d_2} + 2L_\alpha}{n} \sqrt{\frac{1}{t}} \sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} + \frac{2\tilde{L}_\alpha}{n} \frac{1}{t} \left( q(\beta) + \frac{2d_1 d_2}{\beta} \right). \end{aligned} \tag{4.3.12}$$

where  $c_0, c_1$  are absolute constants, and  $q(\beta), t_0$  is defined as follows.

$$q(\beta) = \frac{7\beta}{2} + \frac{10}{3} \frac{\beta^3}{(\beta+1)^2}, u_0 = \max\{u : L_{\alpha+u} \leq 2L_\alpha, \tilde{L}_{\alpha+u} \leq 2\tilde{L}_\alpha\},$$

$$t_0 = \frac{1}{2\beta} \left( 7\beta^2 + 4d_1d_2 + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2} \right) \left( 1 + \frac{L_\alpha}{u_0\tilde{L}_\alpha} + \frac{L_\alpha}{\tilde{L}_\alpha} \right)^2.$$

Note that, when the computing resource in terms of running time is unlimited, meaning  $t \rightarrow \infty$  and  $T \rightarrow \infty$ , the rate is the same with that established in Davenport et al. (2014). Also note that Theorem 4.3.3 gives better understanding of the roles the iteration number  $T$  and  $t$  play. The running-time-induced statistical error is of the order  $O\left(\sqrt{\frac{1}{t}} \cdot \left(\frac{L_\alpha}{\sqrt{\min\{d_1, d_2\}}} + \alpha\tilde{L}_\alpha\right)\right) + O(\frac{\alpha^2\tilde{L}_\alpha}{T})$ . The running time for inner loop plays a crucial role, which is reasonable as the inner-loop-error propagates down the outer loop.

There are flexibility in the choice of step sizes  $\eta$ , similar results can be given for other legitimate choices of step sizes. The heuristic algorithm in Davenport et al. (2014) is a 2-block ADMM. Our framework can also be adapted to 2-block ADMM, the change in the down-stream-convergence-analysis is to replace the 3-block convergence rate with 2-block convergence rate and analyze the dimension-dependent geometric quantity involved there with the insights provided by Lemma 4.2.1.

#### 4.4. Application to Causal Inference for Panel Data

In this section, we apply our framework to the causal inference for panel data. Athey et al. (2021) proposed an estimator of the general form (4.1.1) for causal inference for panel data. Their statistical analysis, however, is not tight, and they do not have an optimization procedure targeting their estimator. We provide an improved statistical analysis and apply our framework based on our improved analysis, resulting in a theoretically guaranteed algorithm with precise quantification of the statistical accuracy after certain running time of user's choice.

We take the statistical model in the work by Athey et al. (2021). The model is for panel

data. There are  $N$  items, which can stand for companies. The time period is  $T$ . For each item  $i$ , there is an adoption time  $t_i$ , after which item  $i$  is treated all the way to time  $T$ , and this adoption time is set to  $T$  if never treated. They take Rubin's potential outcome framework. And the complete potential outcome matrix when all are assigned to the control group is  $Y^{full}$ ,

$$Y^{full} = \mathbf{L}^* + \boldsymbol{\varepsilon}, \quad \text{where } \mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{L}^*) = 0. \quad (4.4.1)$$

The assumptions on  $\boldsymbol{\varepsilon}$  are as follows.  $\boldsymbol{\varepsilon}$  is independent from  $\mathbf{L}^*$  and the elements of  $\boldsymbol{\varepsilon}$  are  $\sigma$ -sub-Gaussian and independent of each other.

$\mathcal{O}$  is the observation-pair set indicating whether a unit (an item at a certain time) is treated. If we let  $W$  to be defined as

$$W_{it} = \begin{cases} 1, & \text{for } (i, t) \notin \mathcal{O} \\ 0, & \text{for } (i, t) \in \mathcal{O} \end{cases}. \quad (4.4.2)$$

The assumptions for  $\mathcal{O}$  and thus  $W$  are as follows. For each row, suppose row  $i$ , there is an adoption time  $t_i$ , such that  $W_{it} = 1$  for all  $t_i < t \leq T$ ,  $t_i = T$  if the unit never adopt the treatment. The rows of  $W$  are independent. Condition on  $\mathbf{L}^*$ , the adoption time  $t_i$  are independent of each other and  $\boldsymbol{\varepsilon}$ . Also,  $|\mathbf{L}^*|_\infty \leq L_{max}$ , where  $L_{max}$  is a positive real number.

Then under this model, the observed controls are  $Y_{it} = Y_{it}^{full}, (i, t) \in \mathcal{O}$ . For treated elements, i.e.  $(i, t) \notin \mathcal{O}$ ,  $Y_{it}^{full}$  is missing and we let  $Y_{it} = 0$ . The goal is to estimate  $\mathbf{L}^*$ .

We introduce some quantities here. For item  $i$ , the probability that it's not treated through out is  $\pi_T^{(i)} = \mathbb{E}(\mathbb{I}\{t_i = T\})$ . The minimum of this "probability of control" over  $N$  items is  $p_c = \min_{1 \leq i \leq N} \pi_T^{(i)}$ . We use  $\mathbf{P}_{\mathcal{O}}$  to denote an operator mapping  $N$  by  $T$  matrix to  $N$  by  $T$  matrix, with each elements defined as  $\mathbf{P}_{\mathcal{O}}(B)_{(i,t)} = B_{(i,t)}$  if  $(i, t) \in \mathcal{O}$ , and 0 if  $(i, t) \notin \mathcal{O}$ .

Note that in this setting, the matrix  $W$  do not have independence for columns, which renders

RIP condition and restricted strong convexity invalid. The targeted estimator (Athey et al., 2021) is

$$\hat{\mathbf{L}} = \arg \min_{\|\mathbf{L}\|_\infty \leq L_{\max}} \left\{ \frac{1}{|\mathcal{O}|} \|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L})\|_F^2 + \lambda \|\mathbf{L}\|_* \right\} \quad (4.4.3)$$

So causal inference for panel data example fits our general framework (4.1.1). The smooth convex function  $f$ , the convex-but-not-necessarily-smooth function  $g$  and the constraint set in the general framework become follows in causal panel data setting.

$$f(\mathbf{L}) = \frac{1}{|\mathcal{O}|} \|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L})\|_F^2, g(\mathbf{L}) = \lambda \|\mathbf{L}\|_*, C_1 = [-L_{\max}, L_{\max}]^{N \times T}. \quad (4.4.4)$$

Applying our framework to it are two sub-problem as follows.

The first sub-problem is to investigating the statistical behavior of an estimator  $\tilde{\mathbf{L}}$  satisfying conditions (4.4.5).

$$\begin{aligned} \frac{1}{|\mathcal{O}|} \|\mathbf{P}_{\mathcal{O}}(Y - \tilde{\mathbf{L}})\|_F^2 + \lambda \|\tilde{\mathbf{L}}\|_* &\leq \frac{1}{|\mathcal{O}|} \|\mathbf{P}_{\mathcal{O}}(Y - \hat{\mathbf{L}})\|_F^2 + \lambda \|\hat{\mathbf{L}}\|_* + \delta, \\ \|\tilde{\mathbf{L}}\|_\infty &\leq L_{\max} + \delta_1, \end{aligned} \quad (4.4.5)$$

where  $\hat{\mathbf{L}}$  is defined in (4.4.3).

The second sub-problem is developing theoretically guaranteed algorithm finding an  $\tilde{\mathbf{L}}$  satisfying (4.4.5) and analyzing its convergence rate in terms of  $\delta$  and  $\delta_1$  in (4.4.5). Athey et al. (2021) does not have an algorithm for  $\hat{\mathbf{L}}$  in (4.4.3) and the heuristic algorithm used there is for another target estimator.

#### 4.4.1. Statistical-Optimization Interplay

We start with the first sub-problem.

The statistical property of the approximate estimator  $\tilde{\mathbf{L}}$  satisfying (4.4.5) is shown in Theorem 4.4.1, which describes how optimization induced error affects statistical error before

solving the optimization problem.

**Theorem 4.4.1.** *Consider statistical model for causal inference of panel data. Suppose the true parameter matrix  $\mathbf{L}^*$  has rank at most  $R$ , and the penalty parameter*

$$\lambda = \frac{13\sigma \max\{\sqrt{N \log(N+T)}, 8\sqrt{T} \log^{\frac{3}{2}}(N+T)\}}{|\mathcal{O}|}$$

. Let  $\hat{\mathbf{L}}$  be defined in (4.4.3). Suppose the computed estimator  $\tilde{\mathbf{L}}$  satisfies  $f(\tilde{\mathbf{L}}) + g(\tilde{\mathbf{L}}) \leq f(\hat{\mathbf{L}}) + g(\hat{\mathbf{L}}) + \delta$  and  $|\tilde{\mathbf{L}}|_\infty \leq L_{\max} + \delta_1$ . Then with probability at least  $1 - \frac{2}{(N+T)^2}$ , we have

$$\begin{aligned} \frac{\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2}{NT} \leq & \max \left\{ q_0 \frac{R\sigma^2 (N+T) \log^3(N+T)}{p_c^2 NT} + \frac{72}{p_c} \delta + q_1 \frac{\delta(L_{\max} + \delta_1)}{\sigma p_c} \frac{1}{NT} \right. \\ & + q_2 \frac{R(L_{\max} + \delta_1)^2 (N+T)}{p_c^2 NT}, \\ & \left. \frac{132(L_{\max} + \delta_1)^2 \log(N+T)}{p_c N} \right\}, \end{aligned} \quad (4.4.6)$$

where  $q_0, q_1, q_2$  are constants that can be explicitly written out.

*Remark 4.4.1.* Note that when  $\delta = 0$  and  $\delta_1 = 0$ , the estimator becomes the original exact estimator (i.e.  $\hat{\mathbf{L}}$  in (4.4.3)), and our rate becomes of order

$$\max\left\{\sigma^2 R \left(\frac{N+T}{NT}\right) \log^3(N+T) \frac{1}{p_c}, L_{\max} \frac{\log(N+T)}{N}\right\}.$$

This is a faster rate than that in Athey et al. (2021), which is because we sharpen the statistical analysis of the original estimator and we apply our framework to our own analysis of the statistical performance of the original exact estimator. If we apply our framework directly to the analysis in Athey et al. (2021), we expect the same rate when  $\delta$  and  $\delta_1$  are set to 0.

*Remark 4.4.2.* causal inference for panel data is a representation for constrained penalized M-estimator, or more precisely, constrained penalized maximum likelihood estimator, where the penalty term is not smooth (optimization wise). Other constrained non-smoothly-penalized M-estimator includes Lasso with constraints, Danzig selector, elastic net, SVM,

sparse principle component analysis in the penalized form, neural network with Relu activation function.

#### 4.4.2. Optimization Algorithm

In this Section, we apply our algorithm template to causal inference of panel data, which gives theoretically guaranteed optimization algorithm for causal inference of panel data.

To standardize the optimization problem for fitting into our optimization template better, the target optimization problem can be written as

$$\min_{\mathbf{L}} \frac{1}{2} \|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L})\|_F^2 + \frac{1}{2} \lambda |\mathcal{O}| \|\mathbf{L}\|_* + \mathfrak{T}\{|\mathbf{L}|_{\infty} \leq L_{\max}\}. \quad (4.4.7)$$

Applying general outer loop, Algorithm 4.2.1, to causal inference for panel data gives Algorithm 4.4.1.

*Algorithm 4.4.1* (Causal Inference for Panel Data Outer Loop: Inexact Proximal Gradient Descent). Start from point  $\mathbf{L}_0 = \mathbf{0}$ . Step size is  $\eta > 0$ . For  $k \geq 0$ ,

$$\begin{aligned} \mathbf{L}_{k+0.5} &= \mathbf{L}_k - \eta \nabla (\|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L}_k)\|_F^2), \\ \mathbf{L}_{k+1} &= \widetilde{\text{Prox}}_{\eta(\frac{1}{2}\lambda|\mathcal{O}|\|\mathbf{L}\|_* + \mathfrak{T}\{|\mathbf{L}|_{\infty} \leq L_{\max}\})}(\mathbf{L}_{k+0.5}), \end{aligned} \quad (4.4.8)$$

where  $\widetilde{\text{Prox}}$  is an approximate proximal algorithm aiming at finding the proximal of  $\mathbf{L}_{k+0.5}$ ,

$$\begin{aligned} &\text{Prox}_{\eta(\frac{1}{2}\lambda|\mathcal{O}|\|\mathbf{L}\|_* + \mathfrak{T}\{|\mathbf{L}|_{\infty} \leq L_{\max}\})}(\mathbf{L}_{k+0.5}) = \\ &\arg \min_{\mathbf{L}} \left( \frac{1}{2} \|\mathbf{L} - \mathbf{L}_{k+0.5}\|^2 + \eta \left( \frac{\lambda|\mathcal{O}|\|\mathbf{L}\|_*}{2} + \mathfrak{T}\{|\mathbf{L}|_{\infty} \leq L_{\max}\} \right) \right). \end{aligned} \quad (4.4.9)$$

We abbreviate the approximate proximal and proximal in equation (4.4.8) and (4.4.9) as  $\widetilde{\text{Prox}}_{\eta}(\mathbf{L}_{k+0.5})$  and  $\text{Prox}_{\eta}(\mathbf{L}_{k+0.5})$ , respectively, when there is no confusion.

For the inner loop (i.e. computing approximate proximal point  $\widetilde{\text{Prox}}_{\eta}(\mathbf{L}_{k+0.5})$ ), we apply



the template-algorithm, Algorithm 4.2.2.

In this setting, the Augmented Lagrange Function for 3-block ADMM with dual step size  $\beta$  and  $\mathbf{L}_{k+0.5}$  replaced by  $P_0$  is

$$\mathcal{L}_\beta(W, Z, P) = \mathfrak{T}\{W \in C_1\} + \|Z\|_* \lambda |\mathcal{O}| + \|P - P_0\|_F^2 + \frac{\beta}{2} (\|W - P + \frac{\Lambda_1}{\beta}\|_2^2 + \|Z - P + \frac{\Lambda_2}{\beta}\|_2^2), \quad (4.4.10)$$

where  $\Lambda_1$  and  $\Lambda_2$  are dual variables.

The template inner loop, Algorithm 4.2.2, in this setting becomes Algorithm 4.4.2.

*Algorithm 4.4.2* (3 block ADMM for causal inference for panel data). The starting points are  $P^0 = P_0$ ,  $\Lambda_1^0 = \mathbf{0}$ ,  $\Lambda_2^0 = \mathbf{0}$ . Dual step size is  $\beta > 0$ . For  $k \geq 0$ , the iterative steps are

$$\begin{aligned} W^{k+1} &= \text{Proj}_{C_1}(P^k - \frac{1}{\beta}\Lambda_1^k), Z^{k+1} = \text{thresh}(P^k - \frac{1}{\beta}\Lambda_2^k, \frac{\lambda|\mathcal{O}|}{\beta}), \\ P^{k+1} &= \frac{1}{\beta+1} \left( P_0 + \Lambda_1^k + \Lambda_2^k + \frac{\beta}{2}(W^{k+1} + Z^{k+1}) \right), \\ \Lambda_1^{k+1} &= \Lambda_1^k + \beta \left( W^{k+1} - P^{k+1} \right), \\ \Lambda_2^{k+1} &= \Lambda_2^k + \beta \left( Z^{k+1} - P^{k+1} \right), \end{aligned} \quad (4.4.11)$$

where  $\text{thresh}(P, b)$  is defined as follows. Suppose the Singular value decomposition of  $P$  is  $P = UDV$ , then  $\text{thresh}(P, b) = U(D - \text{diag}(b))_+ V$ . We take the average  $\bar{P}^k = \frac{1}{k} \sum_{i=1}^k P^i$  for the output if we end it at  $k$ -th iteration.

#### 4.4.3. Optimization Convergence

In this section, we establish convergence rate for our optimization algorithm introduced in Section 4.4.2 in terms of approximate optimization conditions. We apply results in Section 4.2.3 to our causal inference for panel data setting with appropriate modifications.

Applying theorem 4.2.1 to causal inference for panel data, we have Proposition 4.4.1.

**Proposition 4.4.1** (outer loop for causal inference for panel data). *Suppose we take the gradient proximal algorithm, Algorithm 4.4.1, for outer loop with  $\eta = 1$ . Suppose the*

proximal error satisfies

$$|\widetilde{Prox}_{\frac{\lambda|\mathcal{O}|}{2}\|\mathbf{L}\|_*+\mathfrak{T}\{\mathbf{L}\in C_1\}}(X) - Prox_{\frac{\lambda|\mathcal{O}|}{2}+\mathfrak{T}\{\mathbf{L}\in C_1\}}(X)| \leq \delta_0$$

for all  $X \in \mathcal{R}(\delta_0, C_1)$ .  $C_1$  is defined in (4.4.4) and  $\delta_0$  is a positive real number. Let  $\hat{\mathbf{L}}$  be the target estimator define in (4.4.3). Then we have

$$\begin{aligned} \min_{0 \leq k \leq K} \frac{1}{2} \|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L}_k)\|_F^2 + \frac{\lambda|\mathcal{O}|}{2} \|\mathbf{L}_k\|_* &\leq \frac{1}{2} \|\mathbf{P}_{\mathcal{O}}(Y - \hat{\mathbf{L}})\|_F^2 + \frac{\lambda|\mathcal{O}|}{2} \|\hat{\mathbf{L}}\|_* \\ &+ \frac{1}{2K} \|\mathbf{L}_0 - \hat{\mathbf{L}}\|^2 + \delta_0^2 + 2\delta_0 L_{max} \sqrt{NT} + C(Y)\delta_0 + \min\{\sqrt{N}, \sqrt{T}\} \frac{\lambda|\mathcal{O}|}{2} \delta_0, \end{aligned} \quad (4.4.12)$$

where  $C(Y) = \sup_{\mathbf{L} \in C_1} \|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L})\|$ .

For the inner loop, we have the convergence result in Proposition 4.4.2.

**Proposition 4.4.2** (Convergence of inner loop for causal inference for panel data). *Taking algorithm 4.4.2, with dual step size  $\beta \leq \frac{6}{17}$ , after  $k$  iterations, we have*

$$\begin{aligned} \|\bar{P}^k - P^*\|^2 &\leq \frac{1}{\beta k} \left( (3\beta^2 + 8) \|P_0 - P^*\|^2 + \left( 5 + \frac{8}{3} \left( \frac{\beta}{1+\beta} \right)^2 \right) (\lambda|\mathcal{O}|)^2 \min\{N, T\} \right. \\ &\quad \left. + \left( \beta^2 + \frac{8}{3} \left( \frac{\beta^2}{1+\beta} \right)^2 \right) \|P_0 - Proj_{C_1}(P_0)\|^2 \right). \end{aligned} \quad (4.4.13)$$

Combing the inner loop result, Proposition 4.4.2, and outer loop result, Proposition 4.4.1, we have Theorem 4.4.2 showing the overall convergence in terms of approximate conditions.

**Theorem 4.4.2** (optimization : causal inference for panel data). *Suppose we take proximal gradient descent, Algorithm 4.4.1 with  $\eta = 1$ , for outer loop, and 3-block ADMM algorithm 4.4.2 with dual step size  $\beta \leq \frac{6}{17}$  for inner loop, where  $P_0$  in the inner loop is  $\mathbf{L}_{k+0.5}$  in the outer loop. Define four constants depending on  $\beta$  only,  $q_0(\beta), q_1(\beta), q_2(\beta), q_3(\beta)$ , which we will explicitly write out later. Suppose the number of iterations for inner loop  $k \geq q_0(\beta)$ . Suppose we take  $K$  iterations for outer loop and  $\tilde{\mathbf{L}} = \arg \min_{0 \leq i \leq K} \{ \frac{1}{|\mathcal{O}|} \|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L}_i)\|_F^2 +$*

$\lambda \|\mathbf{L}_i\|_*\}$ . Define a quantity  $\delta(k)$  as

$$\delta(k) = \sqrt{\frac{q_1(\beta)(\lambda|\mathcal{O}|)^2 \min\{N, T\} + q_2(\beta)C(Y)^2 + q_3(\beta)(\|Y\|^2 + 2(NT - |\mathcal{O}|)L_{\max}^2)}{k - q_0(\beta)}}. \quad (4.4.14)$$

Then we have  $\tilde{\mathbf{L}}$  satisfies the polluted conditions (4.4.5) with

$$\begin{aligned} \delta_1 &\leq \delta(k), \\ \delta &\leq \frac{NTL_{\max}^2}{|\mathcal{O}|K} + \frac{2\delta(k)^2}{|\mathcal{O}|} + \delta(k) \left( \frac{4L_{\max}\sqrt{NT}}{|\mathcal{O}|} + \frac{2C(Y)}{|\mathcal{O}|} + \min\{\sqrt{N}, \sqrt{T}\}\lambda \right), \end{aligned} \quad (4.4.15)$$

where  $C(Y) = \sup_{\mathbf{L} \in C_1} \|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L})\|$ . The  $\beta$  dependent constants are

$$\begin{aligned} q_0(\beta) &= \left( \frac{1}{\beta} \left( 6\beta^2 + 16 + 2\beta^2 + \frac{16}{3} \left( \frac{\beta^2}{1+\beta} \right)^2 \right) \right), q_3(\beta) = \frac{1}{\beta} (3\beta^2 + 8), \\ q_1(\beta) &= \frac{1}{\beta} \left( 5 + \frac{8}{3} \left( \frac{\beta}{1+\beta} \right)^2 \right), q_2(\beta) = \beta \left( 2 + \frac{16}{3} \left( \frac{\beta}{1+\beta} \right)^2 \right). \end{aligned}$$

#### 4.4.4. Overall Results

In this section, we are ready to show how the running time influences the statistical accuracy, as shown in Theorem 4.4.3.

**Theorem 4.4.3.** Suppose  $\mathbf{L}^*$  has rank at most  $R$ , and the penalty parameter

$$\lambda = \frac{13\sigma \max\{\sqrt{N \log(N+T)}, 8\sqrt{T} \log^{\frac{3}{2}}(N+T)\}}{|\mathcal{O}|}.$$

Suppose we take proximal gradient descent, Algorithm (4.4.1) with  $\eta = 1$ , for outer loop and 3-block ADMM, Algorithm 4.4.2, with dual step size  $\beta \leq \frac{6}{17}$ , for inner loop, where  $P_0$  in the inner loop is  $\mathbf{L}_{k+0.5}$  in the outer loop. There are constants depending on  $\beta$  only, namely,  $q_0(\beta)$ ,  $\widetilde{q_1(\beta)}$ ,  $\widetilde{q_2(\beta)}$ ,  $\widetilde{q_3(\beta)}$  such that for iteration number of inner loop  $k > q_0(\beta)$ , the error

for inner loop is upper bounded by

$$\delta(k) = \sqrt{\frac{\widetilde{q_1(\beta)}\sigma^2 NT \log^3(N+T) + \widetilde{q_2(\beta)}\|Y\|^2 + \widetilde{q_3(\beta)}NTL_{\max}^2}{\textcolor{red}{k} - q_0(\beta)}}. \quad (4.4.16)$$

Denote  $\tilde{\mathbf{L}}$  to be the outcome in  $K$  iterations in outer loop that has the minimum  $f(\tilde{\mathbf{L}}) + g(\tilde{\mathbf{L}})$ .

There are absolute constants  $q_0, q_1, q_2$  such that with probability at least  $1 - \frac{2}{(N+T)^2}$ ,

$$\begin{aligned} & \frac{\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2}{NT} \\ & \leq \max \left\{ q_0 \frac{\sigma^2 R}{p_c^2} \left( \frac{N+T}{NT} \right) \log^3(N+T) + \left[ \frac{NTL_{\max}^2}{|\mathcal{O}|\textcolor{red}{K}} + \frac{2\delta(\textcolor{red}{k})^2}{|\mathcal{O}|} + \right. \right. \\ & \quad \left. \left. \delta(\textcolor{red}{k}) \left( \frac{8L_{\max}\sqrt{NT} + 2\|Y\|}{|\mathcal{O}|} + \min\{\sqrt{N}, \sqrt{T}\}\lambda \right) \right] \left( \frac{72}{p_c} + q_1 \frac{(L_{\max} + \delta(\textcolor{red}{k}))}{\sigma p_c NT} \right) \right. \\ & \quad \left. + q_2 \left( \frac{N+T}{NT} \right) \frac{\sqrt{R}(L_{\max} + \delta(\textcolor{red}{k}))^2}{p_c^2}, \right. \\ & \quad \left. 132(L_{\max} + \delta(\textcolor{red}{k}))^2 \frac{\log(N+T)}{Np_c} \right\}. \end{aligned} \quad (4.4.17)$$

Note that the optimization error induced statistical error increase is of the order  $O(\frac{L_{\max}^2}{K}) + O(\frac{L_{\max} + \sigma}{\sqrt{k}})$ , meaning that inner loop can be the bottle neck in terms of convergence rate to the limit statistical accuracy. Also, note that when the computing resource is infinity, i.e.  $k \rightarrow \infty$  and  $K \rightarrow \infty$ , our results is stronger than that in the work by Athey et al. (2021). This is because our statistical analysis is stricter and we apply our framework based on our analysis. Our framework can also be applied directly to the problem in terms of the part of statistical analysis of the approximate estimator (i.e. statistical-optimization interplay) based on their original work (Athey et al., 2021), then it would lead to the same rate in the case of infinity computing resource as that in Athey et al. (2021).

## 4.5. Application to Linear Regression (LASSO)

Our framework is designed for problems considering general matrices with constraints, but it is also applicable to vector setting without constraints, which can be considered as a degenerate case. In this section, we show that linear regression with LASSO is such a setting.

We show that analysis and template optimization algorithm in our framework are applicable to (high dimensional sparse) linear regression with LASSO. The optimization algorithm converges to the target LASSO estimator and we give a quantification of how iteration number affects the statistical accuracy of the computed estimator. Further, under restricted strong convexity condition, which holds with high probability and is considered by Loh and Wainwright (2015), our template algorithm applied to LASSO actually has linear convergence rate in a certain range, which matches the optimization rate in Loh and Wainwright (2015). Compared with Loh and Wainwright (2015), we pose less conditions, our optimization algorithm is fully convergent to the target estimator (theirs is not), and in the range that their optimization method performs well, ours is equally well.

Consider the linear model

$$y = \mathbf{X}\theta^* + w, \quad (4.5.1)$$

where we observe the vector-matrix pair  $(y, \mathbf{X}) \in \mathbb{R}^n \times \mathbb{R}^{n \times d}$ .  $d$ -dimensional vector  $\theta^*$  is the unknown true parameter and  $w$  is the noise vector. Each row of  $\mathbf{X}$ ,  $x_i$ , is i.i.d. drawn from  $N(\mathbf{0}, \Sigma)$ . Noise  $w$  is independent of  $\mathbf{X}$ . Each element of  $w$ ,  $w_i$ , is i.i.d drawn from  $N(0, \sigma^2)$ . The goal is to estimate  $\theta^*$ . LASSO estimator is given by

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 + \lambda_n \|\theta\|_1, \quad (4.5.2)$$

for a chosen  $\lambda_n$ .

Under our framework (4.1.1), the smooth convex function  $f(\cdot)$  is  $f(\theta) = \|y - \mathbf{X}\theta\|_2^2$ , the

convex-not-necessarily-smooth function  $g(\cdot)$  is  $g(\theta) = \lambda_n \|\theta\|_1$ . And we do not have constraints.

The first sub-problem becomes investigating the statistical behavior of  $\tilde{\theta}$  satisfying

$$\frac{1}{2n} \|y - \mathbf{X}\tilde{\theta}\|_2^2 + \lambda_n \|\tilde{\theta}\|_1 \leq \frac{1}{2n} \|y - \mathbf{X}\hat{\theta}\|_2^2 + \lambda_n \|\hat{\theta}\|_1 + \delta. \quad (4.5.3)$$

And the second sub-problem is the optimization problem shown in (4.5.2). Our optimization template algorithm in Section 4.2.2 degenerates into the ordinary proximal gradient descent algorithm.

#### 4.5.1. Statistical-Optimization Interplay

LASSO has been intensively analyzed in the literature and the statistical behavior of  $\hat{\theta}$  in Equation (4.5.2) is well understood. The analysis procedures of  $\hat{\theta}$  is consistent with our observation of the analysis of estimators following the general form (4.1.1), specifically summarized as follows. Those analysis start with

$$\frac{1}{2n} \|y - \mathbf{X}\hat{\theta}\|_2^2 + \lambda_n \|\hat{\theta}\|_1 \leq \frac{1}{2n} \|y - \mathbf{X}\theta^*\|_2^2 + \lambda_n \|\theta^*\|_1. \quad (4.5.4)$$

Then with proper conditions on  $\lambda_n$ , this inequality can be easily reduced to

$$0 \leq \frac{1}{2n} \|\mathbf{X}(\hat{\theta} - \theta^*)\|_2^2 \leq \frac{\lambda_n}{2} (\|\hat{\theta} - \theta^*\|_1 + 2\|\theta^*\|_1 - 1\|\hat{\theta}\|_1). \quad (4.5.5)$$

Given that the middle part is essentially a quadratic form of  $\hat{\theta} - \theta^*$  and the right hand side is essentially of linear order for  $\hat{\theta} - \theta^*$ , Inequality (4.5.5) implies  $\|\hat{\theta} - \theta^*\|$  is upper bounded. This is the key idea in the analysis of LASSO estimator. A careful reflection on this procedure gives the key observation that the additive nature of the inequality (4.5.4) is never touched throughout the analysis, which is in align with the mechanism of our framework, meaning that analysis of LASSO estimator can be relatively easily carried to

its approximate version solution, i.e.  $\tilde{\theta}$  satisfying (4.5.3).

Theorem 4.5.1 describes the statistical behavior of  $\tilde{\theta}$ , where we can see how the optimization-induced error affects statistical error before solving the optimization problem.

**Theorem 4.5.1.** *Let  $\rho^2(\Sigma)$  be the maximum diagonal entry of the covariance matrix  $\Sigma$ . Under the linear regression model (4.5.1), for any sparse index set  $S$  such that the cardinal of  $S$ ,  $|S| = s$ , denote  $\theta_{S^c}^*$  to be the vector keeping elements not in  $S$  the same and setting those in  $S$  to be 0. Suppose  $c_1\kappa \geq 64s \cdot c_2\rho^2(\Sigma)\frac{\log d}{n}$ , where  $c_1, c_2$  are constants and can be taken as  $c_1 = 1/8, c_2 = 50$ , and  $\kappa$  is the smallest singular value of  $\Sigma$ . For  $\lambda_n \geq 4\sigma\rho(\Sigma)\sqrt{1 + \frac{\log d}{n}}\sqrt{\frac{\log 2(n+d)}{n}}$ ,  $\tilde{\theta}$  satisfying (4.5.3) satisfies the following inequality with probability at least  $1 - \frac{\exp(-n/32)}{1 - \exp(-n/32)} - \exp(-\frac{n}{2}) - \frac{1}{2(n+d)}$ .*

$$\|\tilde{\theta} - \theta^*\|_2 < \frac{\delta}{2\lambda_n\sqrt{s}} + \frac{\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}. \quad (4.5.6)$$

*Remark 4.5.1.* The error bound in Theorem 4.5.1 has three terms. The first corresponds to optimization error. The second corresponds to approximation error (how different from an s sparse vector). The third term corresponds to estimation error associated with  $s$  unknown coefficients. Till now, we do not need an optimization algorithm that guarantee  $\|\tilde{\theta} - \theta^*\|$  or  $\delta$  in Inequality (4.5.3) to be small. All we need is Inequality (4.5.3) for some  $\delta$ . So the optimization convergence rate for  $\delta$  in Inequality (4.5.3) is possibly faster than general optimization convergence with additional strong convexity or restricted strong convexity conditions. We will show that this is indeed the case, which shows that the first two parts of our framework (i.e. statistical-optimization interplay and optimization template algorithm) automatically adapts to additional stronger conditions.

#### 4.5.2. Optimization Algorithm and Convergence

In the absence of the constraints, our template optimization method degenerates into the ordinary proximal gradient descent as shown in Algorithm 4.5.1.

*Algorithm 4.5.1.* Starting point is  $\theta_0 = \mathbf{0}$ . Step size is  $\eta > 0$ . For  $k \geq 0$ ,

$$\begin{aligned}\theta_{k+0.5} &= \theta_k - \eta \nabla_{\theta} \left( \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 \right), \\ \theta_{k+1} &= \arg \min_{\theta} \left( \frac{1}{2} \|\theta - \theta_{k+0.5}\|^2 + \eta \lambda_n \|\theta\|_1 \right).\end{aligned}\tag{4.5.7}$$

Note that  $\theta_{k+1} = \arg \min_{\theta} \left( \frac{1}{2} \|\theta - \theta_{k+0.5}\|^2 + \eta \lambda_n \|\theta\|_1 \right)$  has explicit expression: the  $i$ -th element of  $\theta_{k+1}$  is  $(\theta_{k+1})_i = \text{sign}((\theta_{k+0.5})_i) \cdot (|(\theta_{k+0.5})_i| - \eta \lambda_n)_+$ , where  $\text{sign}(x) = -1$  for  $x < 0$ ,  $\text{sign}(x) = 0$  for  $x = 0$  and  $\text{sign}(x) = 1$  for  $x > 0$ .

From the convergence results of our template optimization method, i.e. Theorem 4.2.1, we have the optimization convergence rate for Algorithm 4.5.1 in Theorem 4.5.2.

**Theorem 4.5.2** (Optimization Convergence Rate). *Let  $\|\frac{\mathbf{X}^T \mathbf{X}}{n}\|_s$  be the spectral norm of  $\frac{\mathbf{X}^T \mathbf{X}}{n}$ . Let step size  $\eta \leq \|\frac{n}{\mathbf{X}^T \mathbf{X}}\|_s$  for Algorithm 4.5.1. Suppose  $\tilde{\theta}$  is among  $\theta_0, \theta_1, \dots, \theta_T$  and has the smallest  $\frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 + \lambda_n \|\theta\|_1$  value. Then we have that*

$$\frac{1}{2n} \|y - \mathbf{X}\tilde{\theta}\|_2^2 + \lambda_n \|\tilde{\theta}\|_1 \leq \frac{1}{2n} \|y - \mathbf{X}\hat{\theta}\|_2^2 + \lambda_n \|\hat{\theta}\|_1 + \frac{1}{2T\eta} \|\hat{\theta}\|^2, \tag{4.5.8}$$

where  $\hat{\theta}$  is defined in (4.5.2).

Theorem 4.5.2 gives fully converging sub-linear convergence rate, which does not require strong convexity of any form.

Loh and Wainwright (2015) exploits restricted strong convexity, which holds with high probability in high dimensional sparse linear regression, and gives an algorithm with linear convergence rate in certain region. But their convergence result is not fully converging, i.e. optimization error does not converge to 0. We show that, under restricted strong convexity condition, our fully converging optimization algorithm also has linear convergence rate in certain region. Theorem 4.5.3 shows how our optimization algorithm performs under different conditions.

**Theorem 4.5.3.** *Under the linear regression model (4.5.1), let  $S$  be an index set with  $s$*



elements. Suppose  $\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty$ , and

$$\frac{\|\mathbf{X}\theta\|_2^2}{n} \geq a_1\|\theta\|_2^2 - a_2\|\theta\|_1^2, \text{ for all } \theta \in \mathbb{R}^d, \quad (4.5.9)$$

with  $a_2 \leq \frac{1}{64s}a_1$ . Set the step size  $\eta = \frac{n}{\|\mathbf{X}^T \mathbf{X}\|_s}$  in Algorithm 4.5.1. Denote  $F(\theta) = \frac{1}{2n}\|\mathbf{X}\theta\|_2^2 + \lambda_n\|\theta\|_1$ . Suppose,  $F(\theta_K) - F(\hat{\theta}) \leq \varepsilon_K$ , where  $\hat{\theta}$  is defined in Equation (4.5.2). Then we have for  $k \geq K$ ,

$$\begin{aligned} F(\theta_k) - F(\hat{\theta}) &\leq \left(1 - \frac{a_1}{8\frac{\|\mathbf{X}^T \mathbf{X}\|_s}{n}}\right)_+^{k-K} \varepsilon_K + 128a_2s \cdot \left(\frac{2\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2 \\ &\quad + 8a_2\frac{\varepsilon_K^2}{\lambda_n^2}, \end{aligned} \quad (4.5.10)$$

where  $\|\cdot\|_s$  is spectral norm,  $\kappa$  is the smallest singular value of  $\Sigma$ , and  $\theta_{S^c}^*$  is  $\theta^*$  taking only elements in  $S^c$  to be the same and setting others to 0.

Without above conditions except for step size  $\eta = \frac{n}{\|\mathbf{X}^T \mathbf{X}\|_s}$  in Algorithm 4.5.1 and using the same notation, we have for  $k \geq 1$ ,

$$\varepsilon_k \leq \frac{\frac{\|\mathbf{X}^T \mathbf{X}\|_s}{n}}{2k} \|\hat{\theta}\|_2^2. \quad (4.5.11)$$

Inequality (4.5.10) in Theorem 4.5.3 has similar form with Theorem 3 in Loh and Wainwright (2015), but our optimization procedure is unconstrained and does not require a pre-specified bound for  $\|\theta^*\|_1$ . We explain the results in details in remarks. In addition to Inequality (4.5.10), we have Inequality (4.5.11), a fully converging convergence result without restricted strong convexity requirement, which parallels Theorem (4.5.2).

*Remark 4.5.2.* Note that Inequality (4.5.10) is only meaningful for  $\varepsilon_K < \frac{\lambda_n^2}{8a_2}$ . This means the algorithm needs to start with a close enough initial point or the algorithm can get into this region after some iterations. Similar issue exists for that considered in Loh and Wainwright (2015). Loh and Wainwright (2015) dealt with it by posing hard constraints on

$\|\theta\|_1$ , which leads to a constrained optimization. However, this constraint is not necessary for Lasso. As shown in Inequality (4.5.11) in Theorem 4.5.3,  $\varepsilon_K$  goes to zero with a rate at least  $\frac{1}{K}$ , so the algorithm will get into the region  $\varepsilon_K < \frac{\lambda_n^2}{8a_2}$  after some iterations. Also, without the knowledge of  $\|\theta^*\|_1$ , hand-choosing constraint will likely miss the target.

*Remark 4.5.3.* Note that the right hand side Inequality (4.5.10) is larger than or equal to  $128a_2s \cdot \left( \frac{2\|\theta_{Sc}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa} \right)^2$ . Hence this convergence result has a limit and does not go to 0 with iteration number going to  $\infty$ . It also implies another requirement for Inequality (4.5.10) to be meaningful:  $\varepsilon_K > 128a_2s \cdot \left( \frac{2\|\theta_{Sc}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa} \right)^2$ . So Inequality (4.5.10) does not show fully convergence of the algorithm. Result in Loh and Wainwright (2015) has similar issue, and they established that this optimization limit is smaller than the statistical limit as  $n$  is relatively large. Similar logic applies to our case. This optimization limit highly depends on  $a_2$ . In fact, condition (4.5.9) holds with high probability for  $a_1 = c_1\kappa$  and  $a_2 = c_2\rho^2(\Sigma) \frac{\log d}{n}$ . The optimization limit in our case is also a shrinking quantity (with respect to  $n$ ) times the statistical accuracy. We will see this more clearly in Theorem 4.5.4. We now examine how large a region Inequality (4.5.10) applies to. We need  $\varepsilon_K$  to satisfy

$$128a_2s \cdot \left( \frac{2\|\theta_{Sc}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa} \right)^2 \leq \varepsilon_K \leq \frac{\lambda_n^2}{8a_2}. \quad (4.5.12)$$

Note that  $\lambda_n$  in Theorem 4.5.3 needs to satisfy a lower bound condition (i.e.  $\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty$ ). In fact, for  $\lambda_n \sim \rho(\Sigma)\sigma\sqrt{\frac{\log(n+d)}{n}}$ , the lower bound holds with high probability. As  $\lambda_n \sim \rho(\Sigma)\sigma\sqrt{\frac{\log(n+d)}{n}}$ ,  $a_2 \sim \rho^2(\Sigma) \frac{\log d}{n}$ , we have (4.5.13), which shows that the left hand side of Inequality (4.5.12) is significantly smaller than the right hand side of Inequality (4.5.12) when the dimension is not extremely high.

$$128a_2s \cdot \left( \frac{2\|\theta_{Sc}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa} \right)^2 \sim \max\left\{ \frac{\log d}{n} \|\theta_{Sc}^*\|_1, \frac{s^2(\log d)^2}{n^2} \frac{\rho^2(\Sigma)}{\kappa} \right\} \frac{\lambda_n^2}{8a_2}. \quad (4.5.13)$$

*Remark 4.5.4.* Inequality (4.5.10) in Theorem 4.5.3 implies the block-wise linear convergence

rate within range  $[k_0, k_1]$ , where

$$\varepsilon_{k_0} \leq \frac{\lambda_n^2}{48a_2} \text{ and } \varepsilon_{k_1} \geq 6 \cdot 128a_2s \cdot \left( \frac{2\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa} \right)^2.$$

If  $a_1 < 8\|\mathbf{X}^T\mathbf{X}/n\|_s$ , for  $k \geq k_0$ , let  $T_k = \lfloor (k - k_0) / \lceil \frac{\log 1/6}{\log(1 - \frac{a_1}{8\|\mathbf{X}^T\mathbf{X}/n\|_s})} \rceil \rfloor$ . If  $a_1 \geq 8\|\mathbf{X}^T\mathbf{X}/n\|_s$ , for  $k \geq k_0$ , let  $T_k = k - k_0$ . We have

$$F(\theta_k) - F(\hat{\theta}) \leq \max\{2^{-T_k}\varepsilon_{k_0}, 6 \cdot 128a_2s \cdot \left( \frac{2\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa} \right)^2\}. \quad (4.5.14)$$

A more detailed proof of this statement is given in the proof of Theorem 4.5.4. In fact, Theorem 4.5.3 implies the conventional linear convergence within the range discussed in (4.5.12) with properly chosen decay factor. But that involves much more tedious details without giving additional insight, so we do not make that a formal assertion here.

#### 4.5.3. Overall Results

With Theorem 4.5.1 and optimization convergence results in Theorem 4.5.3, we have Theorem 4.5.4 describing how iteration number affects the statistical accuracy.

**Theorem 4.5.4.** *Let  $\rho^2(\Sigma)$  be the maximum diagonal entry of the covariance matrix  $\Sigma$ . Under the linear regression model (4.5.1), for any sparse index set  $S$  such that the cardinal of  $S$ ,  $|S| = s$ , denote  $\theta_{S^c}^*$  to be the vector keeping elements not in  $S$  the same and setting those in  $S$  to be 0. Suppose  $c_1\kappa \geq 64s \cdot c_2\rho^2(\Sigma) \frac{\log d}{n}$ , where  $c_1, c_2$  are constants and can be taken as  $c_1 = 1/8, c_2 = 50$ , and  $\kappa$  is the smallest singular value of  $\Sigma$ . Suppose  $\lambda_n \geq 4\rho(\Sigma)\sqrt{1 + \frac{\log d}{n}}\sqrt{\frac{\log 2(n+d)}{n}}$ . Use Algorithm 4.5.1 with step size  $\eta = \frac{\|\mathbf{X}^T\mathbf{X}\|_s}{n}$ . Let*

$$K_0 = \lceil \frac{48c_2\rho^2(\Sigma) \frac{\log d}{n} \left( \|\theta^*\|_2 + \frac{\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa} \right)^2 \|\mathbf{X}^T\mathbf{X}/n\|_s}{2\lambda_n^2} \rceil. \quad (4.5.15)$$

Let

$$T_k = \begin{cases} \lfloor (k - k_0) / \lceil \frac{\log 1/6}{\log(1 - \frac{c_1 \kappa}{8\|\mathbf{X}^T \mathbf{X}/n\|_s})} \rceil \rfloor, & \text{when } c_1 \kappa < 8\|\mathbf{X}^T \mathbf{X}/n\|_s \\ k - k_0, & \text{otherwise} \end{cases}. \quad (4.5.16)$$

Let

$$\begin{aligned} \delta_k = & \min \left\{ \frac{\|\mathbf{X}^T \mathbf{X}/n\|_s}{2k} \left( \frac{\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa} + \|\theta^*\|_2 \right)^2, \right. \\ & \max \left\{ 2^{-T_k} \frac{\lambda_n^2}{48c_2 \rho^2(\Sigma) \frac{\log d}{n}}, \rho^2(\Sigma) \frac{\log d}{n} s \cdot \left( \frac{2\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa} \right)^2 \cdot 768c_2 \right\} \\ & \left. + \mathbb{1}\{k \leq K_0\} \frac{\|y\|_2^2}{2n} \right\}. \end{aligned} \quad (4.5.17)$$

Then with probability at least  $1 - \frac{\exp(-n/32)}{1 - \exp(-n/32)} - \exp(-\frac{n}{2}) - \frac{1}{2(n+d)}$ , the following statements holds.

$$\|\theta_k - \theta^*\|_2 < \frac{\delta_k}{2\lambda_n \sqrt{s}} + \frac{\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa}. \quad (4.5.18)$$

*Remark 4.5.5.* Theorem 4.5.4 shows how the number of iteration affects the statistical accuracy of the computed estimator. It shows that the error caused by optimization goes to zero with the iteration number goes to infinity. Recall that  $\lambda_n \sim \sqrt{\frac{\log(n+d)}{n}}$  when  $n \geq \log d$ , which is satisfied as we do not consider extreme high dimensional case. Note that when the computation resource is infinity,  $\|\theta_k - \theta^*\|_2 \sim \frac{\|\theta_{S^c}^*\|_1}{\sqrt{s}} + \sqrt{s} \sqrt{\frac{\log(n+d)}{n}}$ . When the true vector  $\theta^*$  is indeed  $s$ -sparse,  $\|\theta_k - \theta^*\|_2 \sim \sqrt{s} \sqrt{\frac{\log(n+d)}{n}}$ , which is the optimal rate for high dimensional linear regression.

*Remark 4.5.6.* From the expression of  $\delta_k$  in Inequality (4.5.17) and the role of  $\delta_k$  on statistical accuracy shown in Inequality (4.5.18), the convergence rate of error caused by opti-

mization,  $\frac{F(\theta_k) - F(\hat{\theta})}{2\lambda_n\sqrt{s}}$ , has convergence rate  $\sim \frac{1}{k}$  when

$$\frac{F(\theta_k) - F(\hat{\theta})}{2\lambda_n\sqrt{s}} \geq \frac{\lambda_n}{\sqrt{s} \cdot 96c_2\rho^2(\Sigma)\frac{\log d}{n}} \sim \frac{\sigma}{\rho(\Sigma)} \frac{\sqrt{n \log(n+d)}}{\sqrt{s} \log d},$$

or when

$$\begin{aligned} \frac{F(\theta_k) - F(\hat{\theta})}{2\lambda_n\sqrt{s}} &\leq \frac{768c_2\rho^2(\Sigma)\frac{\log d}{n}s}{2\lambda_n\sqrt{s}} \left( \frac{2\|\theta_{Sc}^*\|_1}{\sqrt{s}} + \left(2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}\right) \frac{\lambda_n}{c_1\kappa} \right)^2 \\ &\sim \rho^2(\Sigma) \frac{\log d}{n} s \left( \frac{\|\theta_{Sc}^*\|_1^2}{s\sqrt{s}\lambda_n} + \sqrt{s} \frac{\lambda_n}{\kappa^2} \right). \end{aligned}$$

Otherwise, the optimization algorithm has linear convergence rate. Considering the case where  $\theta_{Sc}^* = \mathbf{0}$ , which is the conventional setting in high dimensional sparse linear regression, we have that the upper and lower bound for the range where  $\frac{F(\theta_k) - F(\hat{\theta})}{2\lambda_n\sqrt{s}}$  has linear convergence are of the order  $\frac{n}{s \log d} \frac{\kappa}{\rho^2(\Sigma)} \Delta_{stat}$  and  $\frac{s \log d}{n} \frac{\rho^2(\Sigma)}{\kappa} \Delta_{stat}$  respectively, where  $\Delta_{stat} = (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa}$  is the limit statistical accuracy. Therefore, our algorithm performs as well as that in literature (e.g. Loh and Wainwright (2015)) under the classical setting, and is fully convergent in general or in special cases (i.e. sparsity and RSC conditions), which is not shown in Loh and Wainwright (2015) for any cases. This shows that our framework, including statistical-optimization interplay and the template algorithm, automatically adapts to the special cases that has simpler setting admitting stronger assumptions. The optimization convergence results for the general framework, however, need to be further crafted when additional conditions are satisfied.

*Remark 4.5.7.* Note that the results has  $\mathbf{X}$ ,  $y$ , and  $\theta^*$  involved.  $\mathbf{X}$  and  $y$  are observable, so we can adjust iteration number accordingly to guarantee the desired accuracy in terms of  $\theta^*$ . For  $\theta^*$ , usually we can have a conservative upper bound for  $\|\theta^*\|_2$ , hence we adjust our iteration number accordingly for the guaranteed accuracy.

## 4.6. Discussion

In the present work, we proposed a framework for considering the influence of the running time on the statistical accuracy and applied the framework to three examples: 1-bit matrix completion and causal inference for panel data and high dimensional sparse linear regression. We get novel interesting novel results for the first two examples and show that our framework adapts to the degenerate case in the third example. Our backbone statistical analysis for causal panel data is also sharper than that in the literature. It would be interesting to see what results can be derived when our framework is applied to other applicable problems, like kernel ridge regression, SVM, network analysis, neural network, and more intensively studied problems like Danzig selector and elastic net to see how the results compare.

Our framework focuses on estimators that are matrices (and vectors as a special case), but our way of integrating optimization consideration into statistical accuracy before solving the optimization problem can be easily carried to tensors. It would be interesting to see how a parallel tensor version framework performs.

Our framework provides a new perspective of the relationship between computational cost and statistical accuracy, where we quantify the value of computing resource in terms of how much statistical accuracy it can buy, precisely and on a continuous scale. This perspective makes it possible to be used in equilibrium in economic problems, e.g. the computing resource invested is the cost and statistical accuracy generates revenue. It would be interesting to see how it works in those equilibrium and it would also be interesting to further investigate the interplay along this perspective.

Our optimization template algorithm can fill in the blank of theoretically guaranteed optimization algorithm for estimators in a large class of statistical problems that fit in the general form of our framework.

The optimization convergence analysis in our framework provides a pipeline for analyzing an optimization problem to the level meeting statistical needs. It would be interesting to investigate the unanalyzed heuristic algorithms or finer the analysis of other statistic-induced

optimization problem to make the constants free from dimension or other statistically important quantities. Also, for our inner loop, we exploited and analyzed the convergence rate of 3-block ADMM, which usually meets the need for statistical problems encountered and can serve as building stone for more blocks, but it would be interesting to investigate the convergence rate for direct multi-block ADMM or its variant under reasonable assumptions.

## APPENDIX

In the appendix, we give the proofs of the theorems, propositions and lemmas in the dissertation. We also give detailed simulation results and detailed discussions of what is briefed in the dissertation.

### A.1. Proofs of the Results in Chapter 2

This section presents the proofs of all the main results given in Chapter 2 except Theorems 2.2.1 and 2.2.2.

#### A.1.1. Notation, Lemmas and Basic Properties

We begin with introducing and recollecting notation that will be frequently used in the proofs.

Note that  $Y_l$ ,  $Y_s$ , and  $Y_e$  are defined on the same probability space. We use  $\mathbb{E}_s$  to denote the expectation with respect to the distribution of  $Y_s$  and so on. We denote by  $i_j^*$  the index for the subinterval at level  $j$  that contains the minimizer  $Z(f)$  and by  $\tilde{j}$  the index for the level where the chosen interval is at least two blocks away from the subinterval containing the minimizer, i.e.,

$$\begin{aligned} i_j^* &= \max\{i : Z(f) \in [t_{j,i-1}, t_{j,i}]\}, \\ \tilde{j} &= \min\{j : |\hat{i}_j - i_j^*| \geq 2\}. \end{aligned} \tag{A.1.1}$$

It is easy to see that  $\tilde{j} \geq 2$ , and  $\tilde{j}$  only depends on  $Y_l$ . In addition, we let

$$j^* = \min\{j : m_j \leq \frac{\rho_z(\varepsilon; f)}{4}\}. \tag{A.1.2}$$



Then by definition  $\frac{\rho_z(\varepsilon; f)}{8} < m_{j^*} \leq \frac{\rho_z(\varepsilon; f)}{4}$ . Furthermore,  $\mu_{j,i}$  denotes the average of  $f$  on interval  $[t_{j,i-1}, t_{j,i}]$ , i.e.,

$$\mu_{j,i} = \frac{1}{m_j} \int_{t_{j,i-1}}^{t_{j,i}} f(t) dt. \quad (\text{A.1.3})$$

Now we give a list of notations that will be used throughout the proofs of theorems in Section 2.3, in case readers get lost in the mid of reading a proof.

$$\begin{aligned} i_j^* &= \max\{i : Z(f) \in [t_{j,i-1}, t_{j,i}]\}, \\ \tilde{j} &= \min\{j : |\hat{i}_j - i_j^*| \geq 2\}, \\ \mu_{j,i} &= \frac{1}{m_j} \int_{t_{j,i-1}}^{t_{j,i}} f(t) dt, \\ j^* &= \min\{j : m_j \leq \frac{\rho_z(\varepsilon; f)}{4}\}, \\ \mathcal{E}_{j,i} &= \frac{1}{\sqrt{m_j}} (W_2(t_{j,i}) - 2W_2(t_{j,i-1}) + W_2(t_{j,i-2})), \\ j^w &= \min\{j : |\hat{i}_j - i_j^*| \geq 5\}, \\ \hat{f} &= \frac{1}{m_{\hat{j}}} \int_{t_{\hat{j}+\Delta-1}}^{t_{\hat{j}+\Delta}} f(t) dt, \\ \Delta &= 2 \left( \mathbb{1}\{\tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}+6} - \tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}+5} \leq 2\sigma_j\} - \mathbb{1}\{\tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}-6} - \tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}-5} \leq 2\sigma_j\} \right). \end{aligned} \quad (\text{A.1.4})$$

For the white noise model, we obtain in the data splitting step three independent copies of the observations  $Y_l$ ,  $Y_s$ , and  $Y_e$ . In our construction, they have the same variance  $3\varepsilon^2$ ; however, this is not necessary. To better show how the results depend on the variance so that similar results can be easily derived for modified splitting procedures, in the supplement we denote the variances for  $Y_l$ ,  $Y_s$ , and  $Y_e$  to be  $c_l^2\varepsilon^2$ ,  $c_s^2\varepsilon^2$  and  $c_e^2\varepsilon^2$  respectively.

For the regression model, the splitting procedure can also be changed and the variances of the copies of observations for locating strategy, stopping rule and additional estimation and inference procedures does not have to be the same, we denote  $\gamma_l, \gamma_s, \gamma_e$  to be the

scaling factors for the three copies respectively:  $\text{Var}(y_{l,i}) = \gamma_l^2 \sigma^2$ ,  $\text{Var}(y_{s,i}) = \gamma_s^2 \sigma^2$ , and  $\text{Var}(y_{e,i}) = \gamma_e^2 \sigma^2$ , for all  $i$ .

For the regression model, we have similar notion of the length of subinterval, the index of the interval in which the minimizer lies, etc. The following notation will be used in the proofs of the results for regression model.

$$\begin{aligned}
m_j &= \frac{2^{J-j}}{n}, \\
\mathfrak{t}_{j,i} &= i \cdot m_j - \frac{1}{n}, \\
\mathfrak{i}_j^* &= \max\{i : Z(f) \in [\mathfrak{t}_{j,i-1} + \frac{1}{2n}, \mathfrak{t}_{j,i} + \frac{1}{2n}]\}, \\
\tilde{\mathfrak{j}} &= \min\{\min\{j : |\hat{\mathfrak{i}}_j - \mathfrak{i}_j^*| \geq 2\}, \infty\}, \\
\mathfrak{j}^* &= \min\{j : m_j \leq \frac{\rho_z(\frac{\sigma}{\sqrt{n}}; f)}{4}\}, \\
\mathfrak{j}^{\mathfrak{w}} &= \min\{j : |\hat{\mathfrak{i}}_j - \mathfrak{i}_j^*| \geq 5\}, \\
Y_x &= \{y_{x,0}, y_{x,1}, \dots, y_{x,n}\}, \text{ for } x = l, s, e, \\
\text{ave}_f(j, i) &= \frac{1}{2^{J-j}} \sum_{k=2^{J-j}(i-1)}^{2^{J-j} \cdot i - 1} f(x_k), \\
\mathfrak{E}_{j,i,x} &= Y_{j,i,x} - \text{ave}_f(j, i) \cdot 2^{J-j}, \\
\hat{\mathfrak{f}} &= \text{ave}_f(\hat{\mathfrak{j}}, \tilde{\mathfrak{i}}_{\hat{\mathfrak{j}}}).
\end{aligned} \tag{A.1.5}$$

For a better logic flow, some additional notation for non-parametric regression are introduced in Section A.1.9.

We also recall some of the basic properties that will be frequently used in the proofs. The proofs will be deferred to the next section of the supplement, as all the other supporting lemmas. We first revisit a basic property for convex functions.

**Lemma A.1.1.** *For a convex function  $f$ , and any  $0 \leq x_1 < x_2 < x_3 \leq 1$ , we have*

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_1)}{x_3 - x_1} \leq \frac{f(x_3) - f(x_2)}{x_3 - x_2}.$$

Next we introduce the following lemma that helps with detailed calculation.

**Lemma A.1.2.** *For  $x > 6^{1/3}$ , we have*

$$\frac{2x\Phi(2 - (2x)^{\frac{3}{2}}\sqrt{2/3})}{x\Phi(2 - \sqrt{2/3}x^{3/2})} < 0.008,$$

where  $\Phi$  is the CDF of a standard normal distribution.

We further introduce two quantities that will be often used in the proofs of the theorems in Section 2.3 of the main paper. Let

$$Q = \sup_{x \geq 0} x^2 \Phi(-x) \quad \text{and} \quad V = \sup_{x \geq 0} x^2 \Phi(2 - x), \quad (\text{A.1.6})$$

for which we have the following results.

**Lemma A.1.3.**

$$Q = \sup_{x \geq 0} x^2 \Phi(-x) \leq 0.169, \quad V = \sup_{x \geq 0} x^2 \Phi(2 - x) < 2.0555. \quad (\text{A.1.7})$$

### A.1.2. Proof of Proposition 2.2.1

We start with proving for  $f \in \mathcal{F}$

$$c \leq \frac{\rho_m(c\varepsilon; f)}{\rho_m(\varepsilon; f)} \leq c^{\frac{2}{3}}. \quad (\text{A.1.8})$$

*Proof.* Without loss of generality, we assume  $M(f) = 0$ . We first prove the left hand side.

Write

$$g_\beta(t) = \max\{f(t), \rho_m(\beta\varepsilon; f)\},$$

and it is not hard to see that

$$\|g_1 - f\|^2 = \varepsilon^2, \quad \|g_c - f\|^2 = c^2\varepsilon^2. \quad (\text{A.1.9})$$

Write  $\tilde{g}(t) = \max\{f(t), c\rho_m(\varepsilon; f)\}$ , and suppose  $t_{l,m} = \min\{t : \tilde{g} \geq f\}$ ,  $t_{r,m} = \max\{t : \tilde{g} \geq f\}$ . It holds that  $[t_{l,m}, t_{r,m}] \subset \{t : f(t) \leq \rho_m(\varepsilon; f)\}$ . We have

$$\|\tilde{g} - f\|^2 = \int_{t_{l,m}}^{t_{r,m}} (c\rho_m(\varepsilon; f) - f(t))^2 dt \quad (\text{A.1.10})$$

$$\leq \int_{t_{l,m}}^{t_{r,m}} c^2 (\rho_m(\varepsilon; f) - f(t))^2 dt \quad (\text{A.1.11})$$

$$\leq c^2 \|g_1 - f\|^2 = c^2 \varepsilon^2. \quad (\text{A.1.12})$$

Therefore,  $\tilde{g} \leq g_c$  at all the points.  $c\rho_m(\varepsilon; f) \leq \rho_m(c\varepsilon; f)$ .

Now we turn to the right hand side, for which we are interested in

$$\inf_{f \in \mathcal{F}} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)}.$$

Define the left side and right side of the “water area” with “water level”  $\rho_m(c\varepsilon; f)$  to be

$$x_{l,m} = \min\{x : g_c(x) \geq f(x)\}, x_{r,m} = \max\{x : g_c(x) \geq f(x)\}. \quad (\text{A.1.13})$$

We then divide the rest of the proof into four steps.

- The first step is to show that taking the infimum of  $\frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)}$  over  $\mathcal{F}$  is the same as over the function class

$$\mathcal{F}_l = \{f \in \mathcal{F} : f|_{[0, x_{l,m}]}, f|_{[x_{l,m}, 1]} \text{ are linear functions}\}.$$

- The second step will show that it is further the same as over the function class

$$\mathcal{F}_U = \{f \in \mathcal{F} : f|_{[0, Z(f)]}, f|_{[Z(f), 1]} \text{ are piece-wise linear functions with at most two pieces, } f|_{[0, x_{l,m}]}, f|_{[x_{r,m}, 1]} \text{ are linear functions} \}.$$

- In the third step, we define two extended function spaces

$$\begin{aligned} \tilde{\mathcal{F}}_c &= \{f \text{ is convex function with unique minimizer on } (-\infty, \infty) : \\ &\quad f|_{(-\infty, 0]}, f|_{[1, \infty)} \text{ are linear functions, } f|_{[0, 1]} \in \mathcal{F}\}, \\ \tilde{\mathcal{F}}_U &= \{f \in \tilde{\mathcal{F}}_c : f|_{(-\infty, Z(f)]} \text{ and } f|_{[Z(f), \infty)} \text{ are piece-wise linear functions with at most three pieces} \}. \end{aligned}$$

Also, we define two extended geometric indexes  $\tilde{\rho}_z(\varepsilon; f)$ ,  $\tilde{\rho}_m(\varepsilon; f)$  for  $f \in \tilde{\mathcal{F}}_c$ :

$$\tilde{\rho}_z(\varepsilon; f) = \max\{|t - Z(f)| : f(t) \leq \mu(\varepsilon; f)\}, \quad \tilde{\rho}_m(\varepsilon; f) = \mu(\varepsilon; f) - M(f),$$

where  $\mu(\varepsilon; f)$  satisfies

$$\|\max\{\mu(\varepsilon; f), f\} - f\|^2 = \varepsilon^2.$$

We will show in the third step that

$$\inf_{f \in \mathcal{F}_U} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)} \geq \inf_{f \in \tilde{\mathcal{F}}_U} \frac{\tilde{\rho}_m(\varepsilon; f)}{\tilde{\rho}_m(c\varepsilon; f)}.$$

- Finally, in the fourth step, we will show that

$$\inf_{f \in \tilde{\mathcal{F}}_U} \frac{\tilde{\rho}_m(\varepsilon; f)}{\tilde{\rho}_m(c\varepsilon; f)} \geq \inf_{f \in \tilde{\mathcal{F}}_L} \frac{\tilde{\rho}_m(\varepsilon; f)}{\tilde{\rho}_m(c\varepsilon; f)} = c^{-\frac{2}{3}},$$

where  $\tilde{\mathcal{F}}_L = \{f \in \tilde{\mathcal{F}}_U : f|_{(-\infty, Z(f)]} \text{ and } f|_{[Z(f), \infty)} \text{ are linear functions}\}.$

**Step 1** Define a functional  $L_1$ ,

$$\begin{aligned}
L_1 : \mathcal{F} &\longrightarrow \mathcal{F}_l \\
f &\longmapsto L_1(f),
\end{aligned} \tag{A.1.14}$$

where  $L_1(f)$  is defined as follows. Define the *right slope on the left* and the *left slope on the right* to be

$$Ls(f) = \lim_{\eta \rightarrow 0^+} \frac{f(x_{l,m} + \eta) - f(x_{l,m})}{\eta}, \quad Rs(f) = \lim_{\eta \rightarrow 0^+} \frac{f(x_{r,m}) - f(x_{r,m} - \eta)}{\eta}.$$

Both of the limits exist due to convexity. Let

$$(L_1(f))(t) = \begin{cases} f(x_{l,m}) + Ls \cdot (t - x_{l,m}) & 0 \leq t < x_{l,m} \\ f(t) & t \in [x_{l,m}, x_{r,m}] \\ f(x_{r,m}) + Rs \cdot (t - x_{r,m}) & 1 \geq t > x_{r,m} \end{cases}.$$

Without loss of generality, we assume  $M(f) = 0$ . It is clear that

$$\rho_m(c\varepsilon; f) = \rho_m(c\varepsilon; L_1(f)), \quad M(L_1(f)) = 0, \quad L_1(f)(t) \leq f(t) \forall t \in [0, 1].$$

In what follows we will prove

$$\rho_m(\varepsilon; f) \geq \rho_m(\varepsilon; L_1(f)). \tag{A.1.15}$$

Let  $\tilde{L}_1(f) = \max\{L_1(f), \rho_m(\varepsilon; f)\}$ , then we have

$$\begin{aligned}
\|\tilde{L}_1(f) - L_1(f)\|^2 &= \int_0^1 ((\rho_m(\varepsilon; f) - L_1(f)(t))_+)^2 dt \\
&\geq \int_0^1 ((\rho_m(\varepsilon; f) - f(t))_+)^2 dt \\
&= \varepsilon^2.
\end{aligned} \tag{A.1.16}$$

Inequality (A.1.15) then follows. Therefore, we have

$$\inf_{f \in \mathcal{F}} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)} \geq \inf_{f \in \mathcal{F}} \frac{\rho_m(\varepsilon; L_1(f))}{\rho_m(c\varepsilon; L_1(f))} \geq \inf_{f \in \mathcal{F}_l} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)}. \quad (\text{A.1.17})$$

Since  $\mathcal{F}_l \subset \mathcal{F}$ , we also have

$$\inf_{f \in \mathcal{F}} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)} \leq \inf_{f \in \mathcal{F}_l} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)}. \quad (\text{A.1.18})$$

This gives

$$\inf_{f \in \mathcal{F}} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)} = \inf_{f \in \mathcal{F}_l} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)}. \quad (\text{A.1.19})$$

**Step 2** Define a functional  $L_2$ ,

$$\begin{aligned} L_2 : \mathcal{F}_l &\longrightarrow \mathcal{F}_l \\ f &\longmapsto L_2(f), \end{aligned} \quad (\text{A.1.20})$$

where  $L_2(f)$  is defined as follows. We first introduce two quantities:

$$l(\delta; f) = \min\{t : f(t) \leq \delta + M(f)\}, \quad r(\delta; f) = \max\{t : f(t) \leq \delta + M(f)\}.$$

When there is no confusion, we will omit  $f$ , resulting in  $l(\delta), r(\delta)$ . Now we define four functions  $l_1(t), l_{2,\delta}(t), l_{3,\delta}(t), l_4(t)$ . Recall the definition of  $x_{l,m}$  and  $x_{r,m}$  in (A.1.13).

$$\begin{aligned}
l_1(t) &= \frac{f(x_{l,m}) - f(0)}{x_{l,m}}t + f(0), \text{ when } x_{l,m} > 0, \\
l_1(t) &= (t - x_{l,m}) \lim_{s \rightarrow 0^+} \frac{f(x_{l,m} + s) - f(x_{l,m})}{s} + f(x_{l,m}), \text{ when } Z(f) > x_{l,m} = 0, \\
l_1(t) &= M(f), \text{ when } Z(f) = x_{l,m} = 0, \\
l_{2,\delta}(t) &= \frac{\delta}{l(\delta) - Z(f)}(t - Z(f)) + M(f), \text{ when } Z(f) > 0, \\
l_{2,\delta}(t) &= M(f), \text{ when } Z(f) = 0, \\
l_{3,\delta}(t) &= \frac{\delta}{r(\delta) - Z(f)}(t - Z(f)) + M(f), \text{ when } Z(f) < 1, \\
l_{3,\delta}(t) &= M(f), \text{ when } Z(f) = 1, \\
l_4(t) &= \frac{f(1) - f(x_{r,m})}{1 - x_{r,m}}(t - x_{r,m}) + f(x_{r,m}), \text{ when } x_{r,m} < 1, \\
l_4(t) &= (x_{r,m} - t) \lim_{s \rightarrow 0^+} \frac{f(x_{r,m} - s) - f(x_{r,m})}{s} + f(x_{r,m}), \text{ when } Z(f) < x_{r,m} = 1, \\
l_4(t) &= M(f), \text{ when } Z(f) = x_{r,m} = 1.
\end{aligned} \tag{A.1.21}$$

With these four functions, we can define a new function  $h(\delta; f)$ , for  $0 \leq t \leq 1$ :

$$(h(\delta; f))(t) = \max\{l_1(t), l_{2,\delta}(t), l_{3,\delta}(t), l_4(t)\}.$$

When there is no confusion, we will write it as  $h(\delta)$ . It's obvious that

$$\begin{aligned}
\rho_m(c\varepsilon; h(\rho_m(c\varepsilon; f))) &\geq \rho_m(c\varepsilon; f), & \lim_{\delta \rightarrow 0^+} \rho_m(c\varepsilon; h(\delta)) &\leq \rho_m(c\varepsilon; f), \\
f(t) &\geq (h(\delta))(t), \text{ for } 0 \leq t < l(\delta), & f(t) &\leq (h(\delta))(t), \text{ for } 1 \geq t > r(\delta),
\end{aligned}$$



and that  $\rho_m(c\varepsilon; h(\delta))$  increases as  $\delta$  increases. Therefore,  $\exists \delta_0 \in (0, \rho_m(c\varepsilon; f)]$  such that  $\rho_m(c\varepsilon; h(\delta_0)) = \rho_m(c\varepsilon; f)$ . We define  $L_2(f)$  to be  $h(\delta_0)$ . Since  $\delta_0 \leq \rho_m(c\varepsilon; f)$ , we have

$$h(\delta_0)|_{[0,1]/[x_{l,m}, x_{r,m}]} = f|_{[0,1]/[x_{l,m}, x_{r,m}]},$$

and since  $\rho_m(c\varepsilon; h(\delta_0)) = \rho_m(c\varepsilon; f)$ , we have

$$\{t : h(\delta_0) \leq \rho_m(c\varepsilon; f)\} = [x_{l,m}, x_{r,m}].$$

Therefore,

$$\begin{aligned}
0 &= \|f - g_c\|^2 - \|h(\delta_0) - g_c\|^2 \\
&= \int_{x_{l,m}}^{x_{r,m}} \left( (f(t) - g_c(t))^2 - (h(\delta_0)(t) - g_c(t))^2 \right) dt \\
&= \int_{x_{l,m}}^{x_{r,m}} (h - f)(2g_c - f - h) dt \\
&= \int_{(x_{l,m}, l(\delta_0)) \cup (r(\delta_0), x_{r,m})} (h - f)_+ (2g_c - f - h) dt \\
&\quad + \int_{[l(\delta_0), r(\delta_0)]} (h - f)_- (2g_c - f - h) dt \\
&\leq \int_{(x_{l,m}, l(\delta_0)) \cup (r(\delta_0), x_{r,m})} 2(h - f)(\rho_m(c\varepsilon; f) - \delta_0) dt \\
&\quad + \int_{[l(\delta_0), r(\delta_0)]} 2(h - f)(\rho_m(c\varepsilon; f) - \delta_0) dt \\
&\leq 2(\rho_m(c\varepsilon; f) - \delta_0) \int_0^1 (h - f) dt.
\end{aligned} \tag{A.1.22}$$

It then follows that

$$\begin{aligned}
& \|h - g_1\|^2 - \|f - g_1\|^2 \\
&= \int_{x_{l,m}}^{x_{r,m}} \left( (h - g_1)^2 - (f - g_1)^2 \right) dt \\
&= \int_{x_{l,m}}^{x_{r,m}} (h - f)(f + h - 2g_1) dt \\
&= \int_{x_{l,m}}^{x_{r,m}} (h - f)(f + h - 2g_c) dt + \int_{x_{l,m}}^{x_{r,m}} 2(h - f)(g_c - g_1) dt \\
&= \|h - g_c\|^2 - \|f - g_c\|^2 + 2(g_c - g_1) \int_0^1 (h - f) dt \geq 0.
\end{aligned} \tag{A.1.23}$$

As a result,  $\rho_m(\varepsilon; h) \leq \rho_m(\varepsilon; f)$ , which is  $\rho_m(\varepsilon; L_2(f)) \leq \rho_m(\varepsilon; f)$ . Finally, as we have  $L_2(f) \in \mathcal{F}_l$ , we know that

$$\inf_{f \in \mathcal{F}_l} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)} \geq \inf_{f \in \mathcal{F}_l} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)}.$$

**Step 3** Since functions in  $\tilde{\mathcal{F}}_c$  have unique minimizer, we know that  $\tilde{\rho}_z(\varepsilon; f)$  and  $\tilde{\rho}_m(\varepsilon; f)$  exist for all  $\varepsilon > 0$ . As  $\tilde{\mathcal{F}}_l \subset \tilde{\mathcal{F}}_c$ ,  $\tilde{\rho}_z(\varepsilon; f)$  and  $\tilde{\rho}_m(\varepsilon; f)$  also exist for functions in  $\tilde{\mathcal{F}}_l$ . Now for each  $f \in \mathcal{F}_l$ , define a class of functions  $L_3(f) = \{\tilde{f}_{\delta_1, \delta_2} \in \tilde{\mathcal{F}}_l : \delta_1 > 0, \delta_2 > 0\}$  such that

$$\tilde{\rho}_m(\varepsilon; \tilde{f}_{\delta_1, \delta_2}) \leq \rho_m(\varepsilon; f), \quad \liminf_{\max\{\delta_1, \delta_2\} \rightarrow 0^+} \tilde{\rho}_m(c\varepsilon; \tilde{f}_{\delta_1, \delta_2}) \geq \rho_m(c\varepsilon; f).$$

Furthermore, define function  $\tilde{f}_{\delta_1, \delta_2}$  by defining its values on three intervals  $(-\infty, 0)$ ,  $[0, 1]$ ,  $(1, \infty)$ . Specifically, for  $t \in [0, 1]$ ,

$$\tilde{f}_{\delta_1, \delta_2}(t) = f(t),$$

for  $t \in (-\infty, 0)$ ,

$$\tilde{f}_{\delta_1, \delta_2}(t) = \begin{cases} f(0) + \frac{f(x_{l,m}) - f(0)}{x_{l,m}} t, & x_{l,m} > 0 \\ f(0) + \min\{-\delta_1^{-1}, \lim_{s \rightarrow 0^+} \frac{f(s) - f(0)}{s}\} t, & x_{l,m} = 0 \end{cases},$$

and for  $t \in (1, \infty)$ ,

$$\tilde{f}_{\delta_1, \delta_2}(t) = \begin{cases} f(1) + \frac{f(x_{r,m}) - f(1)}{x_{r,m} - 1}(t - 1), & x_{r,m} < 1 \\ f(1) + \max\{\delta_r^{-1}, \lim_{s \rightarrow 0^+} \frac{f(1) - f(1-s)}{s}\}(t - 1), & x_{l,m} = 1 \end{cases}.$$

Then for any  $\rho_m(c\varepsilon; f) > \xi > 0$ ,

$$\begin{aligned} & \lim_{\max\{\delta_1, \delta_2\} \rightarrow 0^+} \|\max\{\tilde{f}_{\delta_1, \delta_2}, M(f) + \rho_m(c\varepsilon; f) - \xi\} - \tilde{f}_{\delta_1, \delta_2}\| \\ &= \|\max\{f, M(f) + \rho_m(c\varepsilon; f) - \xi\} - f\| < c\varepsilon. \end{aligned} \tag{A.1.24}$$

Therefore,

$$\liminf_{\max\{\delta_1, \delta_2\} \rightarrow 0^+} \tilde{\rho}_m(c\varepsilon; \tilde{f}_{\delta_1, \delta_2}) \geq \rho_m(c\varepsilon; f) - \xi.$$

Since it holds for any  $\rho_m(c\varepsilon; f) > \xi > 0$ , we have

$$\liminf_{\max\{\delta_1, \delta_2\} \rightarrow 0^+} \tilde{\rho}_m(c\varepsilon; \tilde{f}_{\delta_1, \delta_2}) \geq \rho_m(c\varepsilon; f).$$

For any  $\delta_1, \delta_2 > 0$ ,

$$\|\max\{\tilde{f}_{\delta_1, \delta_2}, M(f) + \rho_m(\varepsilon; f)\} - \tilde{f}_{\delta_1, \delta_2}\| \geq \|\max\{f, M(f) + \rho_m(\varepsilon; f)\} - f\| \geq \varepsilon,$$

which yields that

$$\tilde{\rho}_m(\varepsilon; \tilde{f}_{\delta_1, \delta_2}) \leq \rho_m(\varepsilon; f).$$

Since  $L_3(f) \subset \tilde{\mathcal{F}}_U$ , we get

$$\inf_{f \in \mathcal{F}_U} \frac{\rho_m(\varepsilon; f)}{\rho_m(c\varepsilon; f)} \geq \inf_{f \in \tilde{\mathcal{F}}_U} \frac{\tilde{\rho}_m(\varepsilon; f)}{\tilde{\rho}_m(c\varepsilon; f)}.$$

**Step 4** Now we define several sets of functions such that  $\tilde{\mathcal{F}}_l$  is the disjoint union of them.

Let

$$\begin{aligned} \tilde{G}(k_1, k_2) = \{f \in \tilde{\mathcal{F}}_l : f|_{(-\infty, Z(f))} \text{ is } k_1\text{-piece linear function,} \\ f|_{(Z(f), \infty)} \text{ is } k_2\text{-piece linear function}\}. \end{aligned} \quad (\text{A.1.25})$$

Then

$$\tilde{\mathcal{F}}_l = \bigcup_{1 \leq k_1, k_2 \leq 3} \tilde{G}(k_1, k_2).$$

It's easy to see that  $\tilde{\mathcal{F}}_L = \tilde{G}(1, 1)$  and that

$$\frac{\tilde{\rho}_m(\varepsilon; f)}{\tilde{\rho}_m(c\varepsilon; f)} = c^{-\frac{2}{3}}, \quad \forall f \in \tilde{\mathcal{F}}_L.$$

We are left to prove that

$$\inf_{f \in \tilde{\mathcal{F}}_l} \frac{\tilde{\rho}_m(\varepsilon; f)}{\tilde{\rho}_m(c\varepsilon; f)} \geq \inf_{f \in \tilde{\mathcal{F}}_L} \frac{\tilde{\rho}_m(\varepsilon; f)}{\tilde{\rho}_m(c\varepsilon; f)}.$$

Let

$$G(k) = \bigcup_{k_1 + k_2 = k} \tilde{G}(k_1, k_2), \quad \text{for } k = 2, 3, 4, 5, 6.$$

It suffices to prove that for  $k \geq 3$

$$\inf_{f \in G(k)} \frac{\tilde{\rho}_m(\varepsilon; f)}{\tilde{\rho}_m(c\varepsilon; f)} \geq \inf_{f \in G(k-1)} \frac{\tilde{\rho}_m(\varepsilon; f)}{\tilde{\rho}_m(c\varepsilon; f)},$$

by proving which we will finish the final step.

Suppose the set of the turning points is  $S_t$ , then  $|S_t / \{Z(f)\}| = k - 2 \geq 1$ . Suppose

$$x^* = \max\{x \in S_t : f(x) = \max\{f(t) : t \in S_t\}\}, \quad t_l = \min S_t, \quad t_r = \max S_t.$$

Apparently  $x^* \neq Z(f)$ . Without loss of generality, assume  $x^* > Z(f)$ . Then by definition

of  $x^*$ ,  $f|_{[x^*, \infty)}$  is a linear function. We define a function  $L_4(f) \in G(k-1)$ ,

$$(L_4(f))(t) = \begin{cases} f(t), & t < x^* \\ f(x^*) + \lim_{s \rightarrow 0^+} \frac{f(x^*) - f(x^* - s)}{s} (t - x^*), & t \geq x^* \end{cases}. \quad (\text{A.1.26})$$

When  $f(x^*) \geq M(f) + \tilde{\rho}_m(c\varepsilon; f)$ , we have

$$\tilde{\rho}_m(c\varepsilon; L_4(f)) = \tilde{\rho}_m(c\varepsilon; f), \quad \tilde{\rho}_m(\varepsilon; L_4(f)) \leq \tilde{\rho}_m(\varepsilon; f).$$

When  $f(x^*) < M(f) + \tilde{\rho}_m(c\varepsilon; f)$ , we have  $f(t_l) \leq f(x^*) < M(f) + \tilde{\rho}_m(c\varepsilon; f)$ . Denote  $p_l, p_r$  to be the left and right root of  $f(t) = M(f) + \tilde{\rho}_m(\varepsilon; f)$ . Then  $p_l < x_{l,m} < t_l \leq Z(f) < x^* < x_{r,m} < p_r$ . We have

$$\begin{aligned} & \|\max\{L_4(f), M(f) + \tilde{\rho}_m(\varepsilon; f)\} - L_4(f)\|^2 \\ &= \int_{p_l}^{x^*} (M(f) + \tilde{\rho}_m(\varepsilon; f) - L_4(f))^2 dt + \\ & \quad \frac{1}{3} \lim_{s \rightarrow 0^+} \frac{s}{f(x^*) - f(x^* - s)} (\tilde{\rho}_m(\varepsilon; f) + M(f) - f(x^*))^3. \end{aligned} \quad (\text{A.1.27})$$

Furthermore,

$$\begin{aligned} & \int_{p_l}^{x^*} (M(f) + \tilde{\rho}_m(\varepsilon; f) - L_4(f))^2 dt \\ &= \int_{p_l}^{x^*} (M(f) + \tilde{\rho}_m(\varepsilon; f) - f)^2 dt \\ &= \int_{p_l}^{t_l} (M(f) + \tilde{\rho}_m(\varepsilon; f) - f)^2 dt + \int_{t_l}^{x^*} (M(f) + \tilde{\rho}_m(\varepsilon; f) - f)^2 dt. \end{aligned} \quad (\text{A.1.28})$$

Similarly,  $\|\max\{L_4(f), M(f) + \tilde{\rho}_m(c\varepsilon; f)\} - L_4(f)\|^2$  can be split into 3 parts as well.

$$\begin{aligned} & \|\max\{L_4(f), M(f) + \tilde{\rho}_m(c\varepsilon; f)\} - L_4(f)\|^2 \\ &= \int_{x_{l,m}}^{t_l} (M(f) + \tilde{\rho}_m(c\varepsilon; f) - f)^2 dt + \int_{t_l}^{x^*} (M(f) + \tilde{\rho}_m(c\varepsilon; f) - f)^2 dt + \\ & \quad \frac{1}{3} \lim_{s \rightarrow 0^+} \frac{s}{f(x^*) - f(x^* - s)} (\tilde{\rho}_m(c\varepsilon; f) + M(f) - f(x^*))^3. \end{aligned} \quad (\text{A.1.29})$$

For  $\|\max\{f, M(f) + \tilde{\rho}_m(c\varepsilon; f)\} - f\|^2$  and  $\|\max\{f, M(f) + \tilde{\rho}_m(\varepsilon; f)\} - f\|^2$ , they can be split into 3 parts as well.

$$\begin{aligned} & \|\max\{f, M(f) + \tilde{\rho}_m(c\varepsilon; f)\} - f\|^2 \\ &= \int_{x_{l,m}}^{t_l} (M(f) + \tilde{\rho}_m(c\varepsilon; f) - f)^2 dt + \int_{t_l}^{x^*} (M(f) + \tilde{\rho}_m(c\varepsilon; f) - f)^2 dt + \\ & \quad \frac{1}{3} \lim_{s \rightarrow 0^+} \frac{s}{f(x^* + s) - f(x^*)} (\tilde{\rho}_m(c\varepsilon; f) + M(f) - f(x^*))^3. \end{aligned} \quad (\text{A.1.30})$$

$$\begin{aligned} & \|\max\{f, M(f) + \tilde{\rho}_m(\varepsilon; f)\} - f\|^2 \\ &= \int_{p_l}^{t_l} (M(f) + \tilde{\rho}_m(\varepsilon; f) - f)^2 dt + \int_{t_l}^{x^*} (M(f) + \tilde{\rho}_m(\varepsilon; f) - f)^2 dt + \\ & \quad \frac{1}{3} \lim_{s \rightarrow 0^+} \frac{s}{f(x^* + s) - f(x^*)} (\tilde{\rho}_m(\varepsilon; f) + M(f) - f(x^*))^3. \end{aligned} \quad (\text{A.1.31})$$

Since we have

$$\begin{aligned} \frac{\int_{p_l}^{t_l} (M(f) + \tilde{\rho}_m(\varepsilon; f) - f)^2 dt}{\int_{x_{l,m}}^{t_l} (M(f) + \tilde{\rho}_m(c\varepsilon; f) - f)^2 dt} &= \left( \frac{M(f) + \tilde{\rho}_m(\varepsilon; f) - f(t_l)}{M(f) + \tilde{\rho}_m(c\varepsilon; f) - f(t_l)} \right)^3 \\ &\leq \left( \frac{\tilde{\rho}_m(\varepsilon; f) + M(f) - f(x^*)}{\tilde{\rho}_m(c\varepsilon; f) + M(f) - f(x^*)} \right)^3, \end{aligned} \quad (\text{A.1.32})$$

$$\begin{aligned} \frac{\int_{t_l}^{x^*} (M(f) + \tilde{\rho}_m(\varepsilon; f) - f)^2 dt}{\int_{t_l}^{x^*} (M(f) + \tilde{\rho}_m(c\varepsilon; f) - f)^2 dt} &\leq \left( \frac{\tilde{\rho}_m(\varepsilon; f) + M(f) - f(x^*)}{\tilde{\rho}_m(c\varepsilon; f) + M(f) - f(x^*)} \right)^2 \\ &< \left( \frac{\tilde{\rho}_m(\varepsilon; f) + M(f) - f(x^*)}{\tilde{\rho}_m(c\varepsilon; f) + M(f) - f(x^*)} \right)^3, \end{aligned} \quad (\text{A.1.33})$$

and

$$\begin{aligned} & \frac{\frac{1}{3} \lim_{s \rightarrow 0^+} \frac{s}{f(x^* + s) - f(x^*)} (\tilde{\rho}_m(\varepsilon; f) + M(f) - f(x^*))^3}{\frac{1}{3} \lim_{s \rightarrow 0^+} \frac{s}{f(x^* + s) - f(x^*)} (\tilde{\rho}_m(c\varepsilon; f) + M(f) - f(x^*))^3} \\ &= \left( \frac{\tilde{\rho}_m(\varepsilon; f) + M(f) - f(x^*)}{\tilde{\rho}_m(c\varepsilon; f) + M(f) - f(x^*)} \right)^3, \end{aligned} \quad (\text{A.1.34})$$

we know that

$$\frac{1}{c^2} = \frac{\|\max\{f, M(f) + \tilde{\rho}_m(\varepsilon; f)\} - f\|^2}{\|\max\{f, M(f) + \tilde{\rho}_m(c\varepsilon; f)\} - f\|^2} \leq \left( \frac{\tilde{\rho}_m(\varepsilon; f) + M(f) - f(x^*)}{\tilde{\rho}_m(c\varepsilon; f) + M(f) - f(x^*)} \right)^3.$$

Given that

$$\lim_{s \rightarrow 0^+} \frac{s}{f(x^* + s) - f(x^*)} \leq \lim_{s \rightarrow 0^+} \frac{s}{f(x^*) - f(x^* - s)},$$

we have

$$\frac{\|\max\{L_4(f), M(f) + \tilde{\rho}_m(\varepsilon; f)\} - L_4(f)\|^2}{\|\max\{L_4(f), M(f) + \tilde{\rho}_m(c\varepsilon; f)\} - L_4(f)\|^2} \geq \frac{1}{c^2}.$$

Define function  $(L_5(f))(t)$

$$(L_5(f))(t) = M(f) + ((L_4(f))(t) - M(f)) \frac{c\varepsilon}{\|\max\{L_4(f), M(f) + \tilde{\rho}_m(c\varepsilon; f)\} - L_4(f)\|}. \quad (\text{A.1.35})$$

Then

$$\tilde{\rho}_m(c\varepsilon; L_5(f)) = \tilde{\rho}_m(c\varepsilon; f), \quad \tilde{\rho}_m(\varepsilon; L_5(f)) \leq \tilde{\rho}_m(\varepsilon; f), \quad L_5(f) \in G(k-1).$$

Thus the statement is proved.  $\square$

Now let's turn to the proof of the geometric property of the minimizer, namely, for  $f \in \mathcal{F}$ ,

$$\max\{(c/2)^{\frac{2}{3}}, c\} \leq \frac{\rho_z(c\varepsilon; f)}{\rho_z(\varepsilon; f)} \leq 1. \quad (\text{A.1.36})$$

*Proof.* The right hand side of the inequality is straightforward. For the left hand side, we prove a stronger version,

$$c^{-2} \geq \frac{3}{4} \left( \frac{\rho_z(\varepsilon; f)}{\rho_z(c\varepsilon; f)} \right)^2 + \frac{1}{4} \left( \frac{\rho_z(\varepsilon; f)}{\rho_z(c\varepsilon; f)} \right)^3. \quad (\text{A.1.37})$$

Similar to Step 3 in the previous proof for the minimum, for any  $f \in \mathcal{F}$ , we have a class of

functions  $\{\tilde{f}_{\delta_1, \delta_2} : \delta_1, \delta_2\}$ , but with a bit of abuse of notation, we define  $\tilde{f}_{\delta_1, \delta_2}$  here as

$$\tilde{f}_{\delta_1, \delta_2}(t) = \begin{cases} f(t), & t \in [0, 1] \\ f(0) + \min\{-\delta_1^{-1}, \lim_{s \rightarrow 0^+} \frac{f(s) - f(0)}{s}\}t, & t \in (-\infty, 0) \cdot \\ f(1) + \max\{\delta_2^{-1}, \lim_{s \rightarrow 0^+} \frac{f(1) - f(1-s)}{s}\}(t - 1), & t \in (1, \infty) \end{cases}$$

Similarly , we have

$$\lim_{\max\{\delta_1, \delta_2\} \rightarrow 0^+} \tilde{\rho}_z(\varepsilon; \tilde{f}_{\delta_1, \delta_2}) = \rho_z(\varepsilon; f), \quad \lim_{\max\{\delta_1, \delta_2\} \rightarrow 0^+} \tilde{\rho}_z(c\varepsilon; \tilde{f}_{\delta_1, \delta_2}) = \rho_z(c\varepsilon; f).$$

Hence

$$\sup_{f \in \mathcal{F}} \frac{\rho_z(\varepsilon; f)}{\rho_z(c\varepsilon; f)} \leq \sup_{f \in \tilde{\mathcal{F}}_c} \frac{\tilde{\rho}_z(\varepsilon; f)}{\tilde{\rho}_z(c\varepsilon; f)}.$$

Similar to the proof of the minimum, for  $f \in \tilde{\mathcal{F}}_c$ , denote  $p_l, p_r$  to be the two roots of  $f(t) = M(f) + \tilde{\rho}_m(\varepsilon; f)$ , and denote  $q_l, q_r$  to be the two roots of  $f(t) = M(f) + \tilde{\rho}_m(c\varepsilon; f)$ .

Without loss of generality, we can assume  $p_r = Z(f) + \tilde{\rho}_z(\varepsilon; f)$ . We define four quantities:

$$\begin{aligned} \Delta_1 &= \int_{p_l}^{Z(f)} (\tilde{\rho}_m(\varepsilon; f) + M(f) - f)^2 dt, \\ \Delta_2 &= \int_{q_l}^{Z(f)} (\tilde{\rho}_m(c\varepsilon; f) + M(f) - f)^2 dt, \\ \Delta_3 &= \int_{Z(f)}^{q_r} (\tilde{\rho}_m(c\varepsilon; f) + M(f) - f)^2 dt, \\ \Delta_4 &= \int_{Z(f)}^{p_r} (\tilde{\rho}_m(\varepsilon; f) + M(f) - f)^2 dt. \end{aligned} \tag{A.1.38}$$

Then we know that

$$\varepsilon^2 = \|\max\{f, M(f) + \tilde{\rho}_m(\varepsilon; f)\} - f\|^2 = \Delta_1 + \Delta_4, \tag{A.1.39}$$

and that

$$c^2 \varepsilon^2 = \|\max\{f, M(f) + \tilde{\rho}_m(c\varepsilon; f)\} - f\|^2 = \Delta_2 + \Delta_3. \tag{A.1.40}$$



We also have

$$\frac{\Delta_1}{\Delta_2} \geq \left( \frac{\tilde{\rho}_m(\varepsilon; f)}{\tilde{\rho}_m(c\varepsilon; f)} \right)^2 \geq \left( \frac{p_r - Z(f)}{q_r - Z(f)} \right)^2 \geq \left( \frac{\tilde{\rho}_z(\varepsilon; f)}{\tilde{\rho}_z(c\varepsilon; f)} \right)^2. \quad (\text{A.1.41})$$

Next we will prove that

$$\frac{\Delta_4}{\Delta_3} \geq \left( \frac{p_r - Z(f)}{q_r - Z(f)} \right)^3 \geq \left( \frac{\tilde{\rho}_z(\varepsilon; f)}{\tilde{\rho}_z(c\varepsilon; f)} \right)^3.$$

For the ease of notation, we introduce four quantities  $w_1 = p_r - Z(f) = \tilde{\rho}_z(\varepsilon; f)$ ,  $w_2 = q_r - Z(f) \leq \tilde{\rho}_z(c\varepsilon; f)$ ,  $v_1 = \tilde{\rho}_m(\varepsilon; f)$ ,  $v_2 = \tilde{\rho}_m(c\varepsilon; f)$ . Then we have

$$\begin{aligned} \frac{\Delta_4}{\Delta_3} &= \frac{\int_0^{w_1} (v_1 + M(f) - f(p_r - t))^2 dt}{\int_0^{w_2} (v_2 + M(f) - f(q_r - t))^2 dt} \\ &= \frac{w_1 \int_0^1 (v_1 + M(f) - f(p_r - w_1 \cdot t))^2 dt}{w_2 \int_0^1 (v_2 + M(f) - f(q_r - w_2 \cdot t))^2 dt}. \end{aligned} \quad (\text{A.1.42})$$

We also have

$$\begin{aligned} M(f) + v_1 - f(p_r - w_1 \cdot t) &= f(p_r) - f(p_r - w_1 \cdot t) \\ &= \frac{f(p_r) - f(p_r - w_1 \cdot t)}{w_1 \cdot t} w_1 \cdot t \\ &\geq \frac{f(q_r) - f(q_r - w_2 \cdot t)}{w_2 \cdot t} w_1 \cdot t \\ &= \frac{w_1}{w_2} (f(q_r) - f(q_r - w_2 \cdot t)), \end{aligned} \quad (\text{A.1.43})$$

where the inequality follows from convexity of  $f$  as well as the fact that  $p_r > q_r$ ,  $p_r - w_1 \cdot t \geq q_r - w_2 \cdot t$ . Continuing with inequality (A.1.42), we have

$$\frac{\Delta_4}{\Delta_3} \geq \frac{w_1 \int_0^1 \left( \frac{w_1}{w_2} (f(q_r) - f(q_r - w_2 \cdot t)) \right)^2 dt}{w_2 \int_0^1 (f(q_r) - f(q_r - w_2 \cdot t))^2 dt} = \left( \frac{w_1}{w_2} \right)^3. \quad (\text{A.1.44})$$

In addition, we have

$$\frac{\Delta_3}{\Delta_2} \geq \frac{1}{3} \frac{w_2}{\tilde{\rho}_z(c\varepsilon; f)} \quad (\text{A.1.45})$$

Therefore,

$$\begin{aligned}
c^{-2} &= \frac{\Delta_1 + \Delta_4}{\Delta_2 + \Delta_3} \geq \frac{\left(\frac{w_1}{w_2}\right)^2 \Delta_2 + \Delta_3 \left(\frac{w_1}{w_2}\right)^3}{\Delta_2 + \Delta_3} \\
&\geq \frac{1 + \frac{1}{3} \frac{w_2}{\tilde{\rho}_z(c\varepsilon; f)} \frac{w_1}{w_2}}{1 + \frac{1}{3} \frac{w_2}{\tilde{\rho}_z(c\varepsilon; f)}} \left(\frac{w_1}{w_2}\right)^2 \geq \frac{1 + \frac{1}{3} \frac{\tilde{\rho}_z(\varepsilon; f)}{\tilde{\rho}_z(c\varepsilon; f)}}{\frac{4}{3}} \left(\frac{\tilde{\rho}_z(\varepsilon; f)}{\tilde{\rho}_z(c\varepsilon; f)}\right)^2 \\
&= \frac{3}{4} \left(\frac{\tilde{\rho}_z(\varepsilon; f)}{\tilde{\rho}_z(c\varepsilon; f)}\right)^2 + \frac{1}{4} \left(\frac{\tilde{\rho}_z(\varepsilon; f)}{\tilde{\rho}_z(c\varepsilon; f)}\right)^3.
\end{aligned} \tag{A.1.46}$$

Since this holds for all  $f \in \mathcal{F}$ , we have

$$c^{-2} \geq \frac{3}{4} \left( \sup_{f \in \mathcal{F}} \frac{\tilde{\rho}_z(\varepsilon; f)}{\tilde{\rho}_z(c\varepsilon; f)} \right)^2 + \frac{1}{4} \left( \sup_{f \in \mathcal{F}} \frac{\tilde{\rho}_z(\varepsilon; f)}{\tilde{\rho}_z(c\varepsilon; f)} \right)^3.$$

□

### A.1.3. Proof of Proposition 2.2.2

We first show that the local modulus of continuity  $\omega_z(\varepsilon; f)$  can be lower bounded by  $\rho_z(\varepsilon; f)$ . Given  $f$  and  $\varepsilon$ , define  $u_\varepsilon = \sup\{u : \|f - f_u\|_2 \leq \varepsilon\}$ . Let  $t_\ell$  and  $t_r$  ( $t_\ell < Z(f) < t_r$ ) be the two end points of the interval  $\{t : f(t) \leq u_\varepsilon\}$ , and without loss of generality let's assume that  $|t_r - Z(f)| \geq |t_\ell - Z(f)|$ . This means that  $\rho_z(\varepsilon; f) = t_r - Z(f)$ . For  $\delta \in (0, t_r - t_\ell)$ , consider function

$$g_\delta(t) = \max \left\{ f(t), u_\varepsilon - \frac{u_\varepsilon - f(t_r - \delta)}{t_r - t_\ell - \delta} (t - t_\ell) \right\}. \tag{A.1.47}$$

It is easy to verify that  $g$  is convex, with its minimum point at  $t_r - \delta$ , and that  $\|f - g_\delta\| \leq \|f - f_{u_\varepsilon}\| \leq \varepsilon$ . See a graphical illustration in Figure A.1. Therefore, taking  $\delta \rightarrow 0$  we have

$$\omega_z(\varepsilon; f) \geq \lim_{\delta \rightarrow 0} (t_r - \delta) = \rho_z(\varepsilon; f).$$

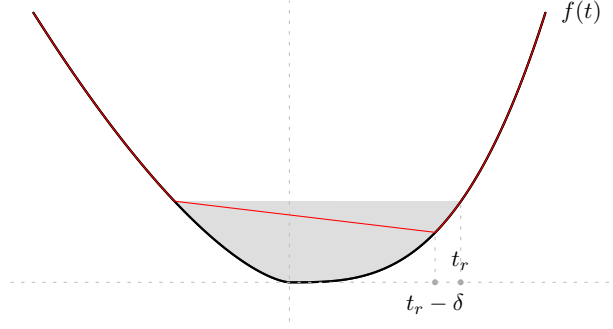


Figure A.1: Illustration of construction of  $g_\delta$ , colored red in the plot

Let's switch to upper bound. Suppose  $g$  is a function such that  $\|f - g\| \leq \varepsilon$ , with minimum point at  $Z(g) > Z(f)$ . We will use proof by contradiction.

If  $Z(g) > Z(f) + 3\rho_z(\varepsilon; f)$ , then  $1 \geq Z(f) + 3\rho_z(\varepsilon; f)$ . Recycling our notation, write  $t_\ell(u_\varepsilon) = \inf\{t : f(t) \leq u_\varepsilon\}$  and  $t_r(u_\varepsilon) = \sup\{t : f(t) \leq u_\varepsilon\}$ . Since  $f$  is a convex function, it is continuous, hence  $f(t_r(u_\varepsilon)) = u_\varepsilon$ . We have two cases: 1,  $g(t_r(u_\varepsilon)) > u_\varepsilon$ , 2,  $g(t_r(u_\varepsilon)) \leq u_\varepsilon$ .

For case 1, we know  $g(t) > u_\varepsilon$  for  $t_\ell(u_\varepsilon) \leq t \leq t_r(u_\varepsilon)$ , so

$$\|f - g\|^2 > \int_{t_\ell(u_\varepsilon)}^{t_r(u_\varepsilon)} (u_\varepsilon - f(t))^2 dt = \varepsilon^2.$$

For case 2, we know  $g(t) \leq u_\varepsilon$  for  $t_r(u_\varepsilon) \leq t \leq t_r(u_\varepsilon) + 2\rho_z(\varepsilon; f)$ , so

$$\begin{aligned} \|f - g\|^2 &\geq \int_{t_r(u_\varepsilon)}^{t_r(u_\varepsilon) + 2\rho_z(\varepsilon; f)} \left( \frac{u_\varepsilon}{t_r(u_\varepsilon) - Z(f)} (t - t_r(u_\varepsilon)) \right)^2 dt \\ &\geq \frac{u^2}{(t_r(u_\varepsilon) - Z(f))^2} \frac{8\rho_z(\varepsilon; f)^3}{3} = \frac{8}{3}\rho_z(\varepsilon; f)u^2 \geq \frac{4\varepsilon^2}{3} \end{aligned}$$

Either case, there is a contradiction. Therefore,  $Z(g) \leq Z(f) + 3\rho_z(\varepsilon; f)$ .

Let's now turn to  $\omega_m(\varepsilon; f)$  and show firstly that  $\omega_m(\varepsilon; f) \geq \rho_m(\varepsilon; f)$ . In fact, if we take the

convex function  $g_\delta$  as defined in (A.1.47), we have that  $\|f - g_\delta\| \leq \varepsilon$  and that

$$\lim_{\delta \rightarrow 0^+} \min_t g_\delta(t) - Z(f) = \rho_m(\varepsilon; f),$$

which completes the proof.

Next, we will show that  $\omega_m(\varepsilon; f)$  can be upper bounded by  $\rho_m(\varepsilon; f)$  up to a constant factor of 3.

For any  $g \in \mathcal{F}$  such that  $\|f - g\| \leq \varepsilon$ , we can immediately obtain

$$M(g) - M(f) \leq \rho_m(\varepsilon; f).$$

Otherwise, if  $M(g) - M(f) > \rho_m(\varepsilon; f)$ , then  $g(t) > \rho_m(\varepsilon; f) + M(f)$  for all  $t$ , and hence  $\varepsilon^2 \geq \|f - g\|^2 \geq (M(g) - M(f) - \rho_m(\varepsilon; f))^2(t_r(u_\varepsilon) - t_l(u_\varepsilon)) + \|f_{u_\varepsilon} - f\|^2 > \varepsilon^2$ .

On the other hand, we need to show the minimum value of  $g$  cannot be too small compared to  $M(f)$ . For the ease of presentation, we assume that  $M(f) = 0$  only for this part. As in the previous parts, we write  $t_\ell = \inf\{t : f(t) \leq u_\varepsilon\}$ ,  $t_r = \sup\{t : f(t) \leq u_\varepsilon\}$ , and  $v_\varepsilon = t_r - t_\ell$ . Graphically,  $v_\varepsilon$  is the width of the water-filling surface. Suppose that  $M(g) = -\alpha u_\varepsilon$  for some  $\alpha > 0$ . Consider the width of the set  $\{t : g(t) \leq 0\}$ , which we denote as  $\gamma v_\varepsilon$  for some  $\gamma > 0$ . From Figure A.2, we see that the integral  $\|f - g\|_2^2$  has to contain the  $\ell_2$  area of the three shaded triangles (the two triangles on the side might not exist). Given that  $M(g) = -\alpha u_\varepsilon$  and  $|\{t : g(t) \leq 0\}| = \gamma v_\varepsilon$ , some calculation shows that

$$\begin{aligned} \|f - g\|^2 &\geq u_\varepsilon^2 v_\varepsilon \cdot \frac{1}{3} \alpha^2 \gamma \left( 1 + \left( \frac{1}{\gamma} - \frac{\alpha + 1}{\alpha} \right)^3 \vee 0 \right) \\ &\geq \varepsilon^2 \cdot \frac{1}{3} \alpha^2 \gamma \left( 1 + \left( \frac{1}{\gamma} - \frac{\alpha + 1}{\alpha} \right)^3 \vee 0 \right) \end{aligned}$$

where the second inequality follows from  $u_\varepsilon^2 v_\varepsilon \geq \varepsilon^2$ . Fixing  $\alpha$  and minimizing over  $\gamma$ , we

have that if  $\alpha > 3$ ,  $\|f - g\|^2 > \varepsilon^2$ , which is contradictory. Therefore, we have

$$M(f) - M(g) \leq 3\rho_m(\varepsilon; f).$$

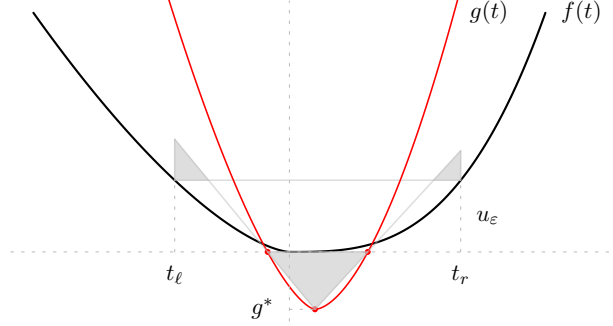


Figure A.2: Illustration of upper bound proof

#### A.1.4. Proof of Theorem 2.2.3

We will first introduce two propositions, which we will prove later. Based on these two propositions, we will finish proving the theorem.

**Proposition A.1.1** (Penalty for super-efficiency in estimation of the minimizer). *For any estimator  $\hat{Z}$ , if  $\exists f \in \mathcal{F}$  such that  $\mathbb{E}_f |\hat{Z} - Z(f)| \leq cR_z(\varepsilon; f)$ , then  $\exists f_1 \in \mathcal{F}$ , such that*

$$\mathbb{E}_{f_1} (|\hat{Z} - Z(f_1)|) \geq h_z(c)R_z(\varepsilon; f_1),$$

for  $0 < c < \frac{2}{15}$ .  $h_z(c) \geq \mathbb{1}\{0.0042 \leq c < \frac{2}{15}\} 0.111 (1 - \Phi(1 + \Phi^{-1}(3c))) + \mathbb{1}\{0 < c < 0.0042\} \max\{\frac{1}{6} (\frac{3}{27})^{\frac{1}{3}} \Phi^{-1}(1 - 3c)^{\frac{2}{3}}, 0.111 (1 - \Phi(1 + \Phi^{-1}(3c)))\}$ .

**Proposition A.1.2** (Penalty for super-efficiency in estimation of the minimum). *For any estimator  $\hat{M}$ , if  $\exists f \in \mathcal{F}$  such that  $\mathbb{E} |\hat{M} - M(f)| \leq cR_m(\varepsilon; f)$ , then  $\exists f_1 \in \mathcal{F}$ , such that*

$$\mathbb{E}_{f_1} |\hat{M} - M(f_1)| \geq h_m(c)R_m(\varepsilon; f_1),$$

for  $0 < c < 0.1$ .  $h_m(c) \geq \mathbb{1}\{0.1 > c \geq \frac{\Phi(-1)}{2.06}\}0.208118 + \mathbb{1}\{0 < c < \frac{\Phi(-1)}{2.06}\}z_{2.06c}^{\frac{2}{3}}/4.12$ .

In the statement of the theorem in the main paper, we use  $\gamma$  in the place of  $c$ , but since we save  $\gamma$  for other usage in the proofs of the proposition and lemmas, we take  $c$  in the place of the  $\gamma$  in the statement of the theorem in the main paper.

From Proposition A.1.1 we know that

$$h_z(c) \geq 0.286z_{3c}^{\frac{2}{3}}, \text{ for } c < 0.0042.$$

Suppose  $\bar{h}_z(c) = 0.111(1 - \Phi(1 - z_{3c}))$ . Then we know that  $\bar{h}_z(c)$  decreases with  $c$  increasing, and since we also know  $\log\left(\frac{1}{c}\right)^{\frac{1}{3}}$  decreases with  $c$  increasing when  $c \in (0, 0.1)$ , we know that

$$\begin{aligned} \inf_{c \in [0.0042, 0.1]} \frac{\bar{h}_z(c)}{\log\left(\frac{1}{c}\right)^{\frac{1}{3}}} &\geq \min_{4 \leq k \leq 99} \inf_{c \in [\frac{k}{1000}, \frac{k+1}{1000}]} \frac{\bar{h}_z(c)}{\log\left(\frac{1}{c}\right)^{\frac{1}{3}}} \\ &\geq \min_{4 \leq k \leq 99} \frac{\bar{h}_z(\frac{k+1}{1000})}{\log\left(\frac{1000}{k}\right)^{\frac{1}{3}}} \geq 0.0266 > \frac{1}{38}. \end{aligned}$$

From Proposition A.1.2 we know that

$$\inf_{c \in [\frac{\Phi(-1)}{2.06}, 0.1]} \frac{h_m(c)}{\log\left(\frac{1}{c}\right)^{\frac{1}{3}}} \geq \frac{0.208118}{\log\left(\frac{2.06}{\Phi(-1)}\right)^{\frac{1}{3}}} \geq 0.1520614 > \frac{1}{7}.$$

Therefore, we are only left to see the relationships between  $z_{2.06c}^{\frac{2}{3}}$ ,  $z_{3c}^{\frac{2}{3}}$  with  $\log\left(\frac{1}{c}\right)^{\frac{1}{3}}$ . We have the following lemma that we will prove in Section A.2 on page 201.

**Lemma A.1.4.** For  $\alpha < 0.08$ ,  $z_{2.06\alpha} \geq 0.61\sqrt{\log 1/\alpha}$ . For  $\alpha < 0.005$ ,  $z_{3\alpha} \geq 0.599\sqrt{\log 1/\alpha}$ .

Since  $0.08 > \frac{\Phi(-1)}{2.06}$ , we have for  $c < 0.1$ ,

$$h_m(c) \geq \min\left\{\frac{1}{7}, 0.61^{\frac{2}{3}}/4.12\right\} \left(\log \frac{1}{c}\right)^{\frac{1}{3}} = \frac{1}{7} \left(\log \frac{1}{c}\right)^{\frac{1}{3}}.$$

For  $h_z(c)$ , we have, for  $c < 0.1$ ,

$$h_z(c) \geq \min\left\{\frac{1}{38}, 0.599^{\frac{2}{3}} \frac{1}{6} \left(\frac{3}{27}\right)^{\frac{1}{3}} \left(\log \frac{1}{c}\right)^{\frac{1}{3}} = \frac{1}{38} \left(\log \frac{1}{c}\right)^{\frac{1}{3}}\right\}.$$

Now we start proving the propositions.

*Proof of Proposition A.1.1.* We have the following two lemmas, which we will prove in Section A.2 on page 203 and 209.

**Lemma A.1.5.** *For any estimator  $\hat{Z}$ , if  $\exists f \in \mathcal{F}$  such that  $\mathbb{E}_f |\hat{Z} - Z(f)| \leq c\rho_z(\varepsilon; f)$ , then  $\exists f_1 \in \mathcal{F}$ , such that*

$$\mathbb{E}_{f_1} (|\hat{Z} - Z(f_1)|) \geq \tilde{h}_z(c) \rho_z(\varepsilon; f_1)$$

for  $c < 1$ . For  $0 < c < 0.063$ ,  $\tilde{h}_z(c) \geq \frac{1}{4} \left(\frac{3}{27}\right)^{\frac{1}{3}} \Phi^{-1}(1 - 2c)^{\frac{2}{3}}$ .

**Lemma A.1.6.** *For any estimator  $\hat{Z}$ , if  $\exists f \in \mathcal{F}$  such that  $\mathbb{E}_f |\hat{Z} - Z(f)| \leq c\rho_z(\varepsilon; f)$ , then  $\exists f_1 \in \mathcal{F}$ , such that*

$$\mathbb{E}_{f_1} (|\hat{Z} - Z(f_1)|) \geq \tilde{h}_z(c) \rho_z(\varepsilon; f_1)$$

for  $c < 1$ . For  $0 < c < 0.2$ ,  $\tilde{h}_z(c) \geq 0.1666 (1 - \Phi(1 + \Phi^{-1}(2c)))$ .

Recall that, by Lemma 2.6.3,  $0.308\rho_z(\varepsilon; f) \leq R_z(\varepsilon; f) \leq \frac{3}{2}\rho_z(\varepsilon; f)$ . Therefore, for any estimator  $\hat{Z}$ , if  $\exists f \in \mathcal{F}$  such that  $\mathbb{E}_f |\hat{Z} - Z(f)| \leq cR_z(\varepsilon; f)$ , then  $\exists f_1 \in \mathcal{F}$ , such that

$$\mathbb{E}_{f_1} (|\hat{Z} - Z(f_1)|) \geq h_z(c) R_z(\varepsilon; f_1),$$

for  $c < \frac{2}{15}$ .

$$h_z(c) \geq \mathbb{1}\{0.0042 \leq c < \frac{2}{15}\} 0.111 (1 - \Phi(1 + \Phi^{-1}(3c))) + \mathbb{1}\{c < 0.0042\} \max\left\{\frac{1}{6} \left(\frac{3}{27}\right)^{\frac{1}{3}} \Phi^{-1}(1 - 3c)^{\frac{2}{3}}, 0.111 (1 - \Phi(1 + \Phi^{-1}(3c)))\right\}.$$

□

*Proof of Proposition A.1.2.* Again we introduce a lemma and prove it in Section A.2 on page 210.

**Lemma A.1.7.** *For any estimator  $\hat{M}$ , if  $\exists f \in \mathcal{F}$  such that  $\mathbb{E}_f |\hat{M} - M(f)| \leq c\rho_m(\varepsilon; f)$ , then  $\exists f_1 \in \mathcal{F}$ , such that*

$$\mathbb{E}_{f_1} (|\hat{M} - M(f_1)|) \geq \tilde{h}_m(c)\rho_m(\varepsilon; f_1)$$

for  $c < 1$ . For  $c \leq 0.103$ ,  $\tilde{h}_m(c) \geq \mathbb{1}\{0.103 \geq c \geq \frac{\Phi(-1)}{2}\}0.214362 + \mathbb{1}\{c < \frac{\Phi(-1)}{2}\}z_{2c}^{\frac{2}{3}}/4$ .

According to Lemma 2.6.2, we have  $R_m(\varepsilon; f) \leq 1.03\rho_m(\varepsilon; f)$ . Therefore, we have, for any estimator  $\hat{M}$ , if  $\exists f \in \mathcal{F}$  such that  $\mathbb{E}_f |\hat{M} - M(f)| \leq cR_m(\varepsilon; f)$ , then  $\exists f_1 \in \mathcal{F}$ , such that

$$\mathbb{E}_{f_1} |\hat{M} - M(f_1)| \geq h_m(c)R_m(\varepsilon; f_1)$$

for  $c < 1$ . For  $c < 0.1$ ,  $h_m(c) \geq \mathbb{1}\{0.1 > c \geq \frac{\Phi(-1)}{2.06}\}0.208118 + \mathbb{1}\{c < \frac{\Phi(-1)}{2.06}\}z_{2.06c}^{\frac{2}{3}}/4.12$ .

□

### A.1.5. Proof of Theorem 2.3.1

Recall that  $\tilde{j}$  is defined in Equation (A.1.1) and only depends on  $Y_l$ . We have

$$\begin{aligned} \mathbb{E}(|\hat{Z} - Z(f)|) &= \mathbb{E}(\mathbb{1}\{\hat{j} < \tilde{j}\}|\hat{Z} - Z(f)|) + \mathbb{E}(\mathbb{1}\{\hat{j} \geq \tilde{j}\}|\hat{Z} - Z(f)|) \\ &\leq \mathbb{E}(\mathbb{1}\{\hat{j} < \tilde{j}\}1.5m_{\hat{j}}) + \mathbb{E}(\mathbb{1}\{\hat{j} \geq \tilde{j}\}|\hat{Z} - Z(f)|). \end{aligned} \tag{A.1.48}$$

We begin with bounding the first term.



$$\begin{aligned}
& \mathbb{E}_{l,s}(\mathbb{1}\{\hat{j} < \tilde{j}\}m_{\hat{j}}) \\
&= \sum_{j_1=3}^{j^*-1} m_{j_1} \mathbb{E}_{l,s}(\mathbb{1}\{\hat{j} < \tilde{j}, \hat{j} = j_1\}) + \sum_{j_1=j^*}^{\infty} m_{j_1} \mathbb{E}_{l,s}(\mathbb{1}\{\hat{j} < \tilde{j}, \hat{j} = j_1\}) \\
&\leq \sum_{j_1=3}^{j^*-1} 2^{j^*-j_1} m_{j^*} \mathbb{E}_{l,s}(\mathbb{1}\{j_1 < \tilde{j}, \hat{j} = j_1\}) + \sum_{j_1=j^*}^{\infty} m_{j^*} \mathbb{E}_{l,s}(\mathbb{1}\{\hat{j} < \tilde{j}, \hat{j} = j_1\}) \\
&\leq \sum_{j_1=3}^{j^*-1} 2^{j^*-j_1} m_{j^*} \mathbb{E}_{l,s}(\mathbb{1}\{T_{j_1} \leq 2\sqrt{6}\varepsilon\sqrt{m_{j_1}}, j_1 < \tilde{j}\}) + m_{j^*} P(j^* \leq \hat{j} < \tilde{j}) \\
&\leq \sum_{j_1=3}^{j^*-1} 2^{j^*-j_1} m_{j^*} \mathbb{E}_{l,s}(\mathbb{1}\{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5} \leq 2\sqrt{6}\varepsilon\sqrt{m_{j_1}}, j_1 < \tilde{j}\}) + \\
&\quad \sum_{j_1=3}^{j^*-1} 2^{j^*-j_1} m_{j^*} \mathbb{E}_{l,s}(\mathbb{1}\{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5} \leq 2\sqrt{6}\varepsilon\sqrt{m_{j_1}}, j_1 < \tilde{j}\}) + \\
&\quad m_{j^*} P(j^* \leq \hat{j} < \tilde{j}) \\
&= \sum_{j_1=3}^{j^*-1} 2^{j^*-j_1} m_{j^*} \mathbb{E}_l \left( \mathbb{E}_s(\mathbb{1}\{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5} \leq 2\sqrt{6}\varepsilon\sqrt{m_{j_1}}, j_1 < \tilde{j}\} | Y_l) \right) + \\
&\quad \sum_{j_1=3}^{j^*-1} 2^{j^*-j_1} m_{j^*} \mathbb{E}_l \left( \mathbb{E}_s(\mathbb{1}\{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5} \leq 2\sqrt{6}\varepsilon\sqrt{m_{j_1}}, j_1 < \tilde{j}\} | Y_l) \right) + \\
&\quad m_{j^*} P(j^* \leq \hat{j} < \tilde{j}).
\end{aligned} \tag{A.1.49}$$

As bounding the expectations in first two terms of the right hand side of Inequality (A.1.49) takes similar steps, we will walk through the steps for the first term. Note that only when  $\hat{i}_{j_1} + 6 \leq 2^{j_1}$  the indicator function in the expectation can take 1, so in the following we take  $\mathbb{1}\{\hat{i}_{j_1} + 6 \leq 2^{j_1}\}$  as an indicator function in the expectation without writing it out.

We introduce the following quantity for the (partly standardized) noise part of the statistic defined in stopping-rule Section 2.3.1.

$$\mathcal{E}_{j,i} = \frac{1}{\sqrt{m_j}} (W_2(t_{j,i}) - 2W_2(t_{j,i-1}) + W_2(t_{j,i-2})), \tag{A.1.50}$$

where  $W_2$  is define in Equation (2.3.2).

Then for  $2 \leq i \leq 2^j$ , we have

$$\mathcal{E}_{j,i} \sim N(0, 6\varepsilon^2).$$

Hence for  $j_1 \leq j^* - 1$  we have

$$\begin{aligned} & \mathbb{E}_l(\mathbb{E}_s(\mathbb{1}\{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5} \leq 2\sqrt{6}\varepsilon\sqrt{m_{j_1}}, j_1 < \tilde{j}\}|Y_l)) \\ &= \mathbb{E}_l(\mathbb{E}_s(\mathbb{1}\{(\mu_{j_1, \hat{i}_{j_1}+6} - \mu_{j_1, \hat{i}_{j_1}+5})\sqrt{m_{j_1}} - 2\sqrt{6}\varepsilon \leq -\mathcal{E}_{j_1, \hat{i}_{j_1}+6}\}|Y_l)\mathbb{1}\{j_1 < \tilde{j}\})) \quad (\text{A.1.51}) \\ &\leq \mathbb{E}_l(\mathbb{E}_s(\mathbb{1}\{(\mu_{j_1, i_{j_1}^*+5} - \mu_{j_1, i_{j_1}^*+4})\sqrt{m_{j_1}} - 2\sqrt{6}\varepsilon \leq -\mathcal{E}_{j_1, \hat{i}_{j_1}+6}\}|Y_l)\mathbb{1}\{j_1 < \tilde{j}\})). \end{aligned}$$

Further, for  $(\mu_{j_1, i_{j_1}^*+5} - \mu_{j_1, i_{j_1}^*+4})\sqrt{m_{j_1}}$ , we have

$$\begin{aligned} & (\mu_{j_1, i_{j_1}^*+5} - \mu_{j_1, i_{j_1}^*+4})\sqrt{m_{j_1}} \geq \left(\frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} m_{j_1}\right)\sqrt{m_{j_1}} \\ &= \rho_m(\varepsilon; f)\sqrt{\rho_z(\varepsilon; f)} 2^{\frac{3}{2}(j^*-j_1)} \left(\frac{m_{j^*}}{\rho_z(\varepsilon; f)}\right)^{\frac{3}{2}} \quad (\text{A.1.52}) \\ &\geq \frac{1}{\sqrt{2}} \varepsilon 2^{\frac{3}{2}(j^*-j_1)} \left(\frac{m_{j^*}}{\rho_z(\varepsilon; f)}\right)^{\frac{3}{2}} \geq \frac{1}{\sqrt{2}} \varepsilon 2^{\frac{3}{2}(j^*-j_1)} 2^{-\frac{9}{2}}, \end{aligned}$$

where the first inequality is due to Inequality (2.6.9), and the second inequality is due to the definition of  $j^*$  in Equation (A.1.2). We will use both the last and second to last quantity in Inequality (A.1.52) later. Continuing with Inequality (A.1.51), we have

$$\begin{aligned} & \mathbb{E}_l(\mathbb{E}_s(\mathbb{1}\{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5} \leq 2\sqrt{6}\varepsilon\sqrt{m_{j_1}}, j_1 < \tilde{j}\}|Y_l)) \\ &\leq \mathbb{E}_l\left(\mathbb{E}_s\left(\mathbb{1}\left\{\frac{1}{\sqrt{2}}\varepsilon 2^{\frac{3}{2}(j^*-j_1)} 2^{-\frac{9}{2}} \left(\frac{8m_{j^*}}{\rho_z(\varepsilon; f)}\right)^{\frac{3}{2}} - 2\sqrt{6}\varepsilon \leq -\mathcal{E}_{j_1, \hat{i}_{j_1}+6}\right\}|Y_l\right)\mathbb{1}\{j_1 < \tilde{j}\}\right) \quad (\text{A.1.53}) \\ &= \mathbb{E}_l\left(\Phi\left(2 - 2^{\frac{3}{2}(j^*-j_1-3)-1} \frac{1}{\sqrt{3}} \left(\frac{8m_{j^*}}{\rho_z(\varepsilon; f)}\right)^{\frac{3}{2}}\right)\mathbb{1}\{j_1 < \tilde{j}\}\right) \\ &\leq \mathbb{E}_l\left(\Phi\left(2 - 2^{\frac{3}{2}(j^*-j_1-3)-1} \frac{1}{\sqrt{3}}\right)\mathbb{1}\{j_1 < \tilde{j}\}\right). \end{aligned}$$

Note that for  $j_1 \leq j^* - 5$ ,  $2 - 2^{\frac{3}{2}(j^*-j_1-3)-1} \frac{1}{\sqrt{3}} \leq 0$ . We have similar results for the expectation in the second term of Inequality (A.1.49). Plugging them in, we have

$$\begin{aligned}
& \mathbb{E}_{l,s}(\mathbb{1}\{\hat{j} < \tilde{j}\}m_{\tilde{j}}) \\
& \leq m_{j^*}P(j^* \leq \hat{j} < \tilde{j}) + \sum_{j_1=3}^{j^*-5} 2^{j^*-j_1}m_{j^*}\Phi(2 - 2^{\frac{3}{2}(j^*-j_1-3)-1} \frac{1}{\sqrt{3}})\mathbb{E}_l(\mathbb{1}\{j_1 < \tilde{j}\}) \times 2 + \\
& \quad \sum_{j_1=j^*-4}^{j^*-1} 2^{j^*-j_1}m_{j^*}\Phi(2 - 2^{\frac{3}{2}(j^*-j_1-3)-1} \frac{1}{\sqrt{3}} \left( \frac{8m_{j^*}}{\rho_z(\varepsilon; f)} \right)^{\frac{3}{2}})\mathbb{E}_l(\mathbb{1}\{j_1 < \tilde{j}\}) \times 2 \\
& \leq m_{j^*} \times 2^5 \sum_{j_1=3}^{j^*-6} 2^{j^*-j_1-4}\Phi(2 - 2^{\frac{3}{2}(j^*-j_1-4)} \cdot \sqrt{\frac{2}{3}}) + 2\Phi(2 - 2^{\frac{3}{2}} \cdot \sqrt{\frac{2}{3}})2^5m_{j^*} \\
& \quad + \frac{\rho_z(\varepsilon; f)}{8} \times 16 \times \left( \sum_{k=0}^3 \frac{2^{-k}8m_{j^*}}{\rho_z(\varepsilon; f)}\Phi(2 - \sqrt{\frac{2}{3}} \left( \frac{2^{-k}8m_{j^*}}{\rho_z(\varepsilon; f)} \right)^{\frac{3}{2}}) \right) + m_{j^*} \\
& \leq m_{j^*} \times 2^5 \times \frac{4\Phi(2 - 8 \times \sqrt{2/3})}{1 - 0.008} + 24.3m_{j^*} + 2\rho_z(\varepsilon; f)(2 + 1 + 0.5 + 0.25) + m_{j^*} \\
& < 24.4m_{j^*} + 8\rho_z(\varepsilon; f) \leq 14.1\rho_z(\varepsilon; f).
\end{aligned} \tag{A.1.54}$$

The detailed calculations of the third inequality are based on Lemma A.1.2.

Now we can proceed with bounding the second term in Inequality (A.1.48).

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\hat{j} \geq \tilde{j}\}|\hat{Z} - Z(f)|) \\
& \leq \sum_{j=1}^{j^*-4} \mathbb{E}(\mathbb{1}\{\hat{j} \geq j, \tilde{j} = j\}|\hat{Z} - Z(f)|) + \sum_{j=j^*-3}^{\infty} \mathbb{E}(\mathbb{1}\{\hat{j} \geq j, \tilde{j} = j\}|\hat{Z} - Z(f)|) \\
& \leq \sum_{j=1}^{j^*-4} 2^{j^*-j}m_{j^*}\mathbb{E}(\mathbb{1}\{\hat{j} \geq j\}(5\mathbb{1}\{X_{j,i_j^*-3} \leq X_{j,i_j^*-1}\} + 4\mathbb{1}\{X_{j,i_j^*-2} \leq X_{j,i_j^*-1}\} \\
& \quad + 4\mathbb{1}\{X_{j,i_j^*+2} \leq X_{j,i_j^*+1}\} + 5\mathbb{1}\{X_{j,i_j^*+3} \leq X_{j,i_j^*+1}\} \\
& \quad + 6\mathbb{1}\{X_{j,i_j^*+4} \leq X_{j,i_j^*+1}\} + 6\mathbb{1}\{X_{j,i_j^*-4} \leq X_{j,i_j^*-1}\})) + 6 \times 8 \times m_{j^*}
\end{aligned}$$

$$\begin{aligned}
&\leq 2 \sum_{j=1}^{j^*-4} 2^{j^*-j} m_{j^*} \left( 4\Phi\left(-\frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} m_j \frac{\sqrt{m_j}}{\sqrt{6\varepsilon}}\right) + 5\Phi\left(-2\frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} m_j \frac{\sqrt{m_j}}{\sqrt{6\varepsilon}}\right) + \right. \\
&\quad \left. 6\Phi\left(-3\frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} m_j \frac{\sqrt{m_j}}{\sqrt{6\varepsilon}}\right) \right) + 48m_{j^*} \\
&\leq 48m_{j^*} + 2 \sum_{j=1}^{j^*-4} 2^{j^*-j} m_{j^*} \left( 4\Phi\left(-\sqrt{2/3}\left(\frac{1}{8}\right)^{\frac{3}{2}} \times 2^{\frac{3}{2}(j^*-j-1)}\right) + \right. \\
&\quad \left. 5\Phi\left(-\sqrt{2/3} \cdot 2\left(\frac{1}{8}\right)^{\frac{3}{2}} \times 2^{\frac{3}{2}(j^*-j-1)}\right) + 6\Phi\left(-\sqrt{2/3} \cdot 3\left(\frac{1}{8}\right)^{\frac{3}{2}} \times 2^{\frac{3}{2}(j^*-j-1)}\right) \right) \\
&\leq 48m_{j^*} + 2m_{j^*} \left( 16\Phi\left(-\sqrt{\frac{2}{3}}\right) \times 4 + 32\Phi\left(-4\sqrt{\frac{1}{3}}\right) \frac{1}{1 - 2 \times \frac{\Phi(-8\sqrt{\frac{2}{3}})}{\Phi(-4/\sqrt{3})}} \times 4 + \right. \\
&\quad \left. 5 \times \Phi\left(-2\sqrt{\frac{2}{3}}\right) \times 16 \times \frac{1}{1 - 2 \times \frac{\Phi(-8/\sqrt{3})}{\Phi(-2\sqrt{2/3})}} + 6 \times \Phi(-\sqrt{6}) \times \frac{1}{1 - 2 \times \frac{\Phi(-4\sqrt{3})}{\Phi(-\sqrt{6})}} \right) \\
&< 20.9\rho_z(\varepsilon; f).
\end{aligned}$$

The third to last inequality is due to the fact that  $\frac{\Phi(-2\sqrt{2}x)}{\Phi(-x)}$  decreases with  $x > 0$  increasing.

Putting the two parts together, we have

$$\mathbb{E}(|\hat{Z} - Z(f)|) < 35\rho_z(\varepsilon; f) \leq \frac{35}{a_1} R_z(\varepsilon; f). \quad \square \quad (\text{A.1.55})$$

#### A.1.6. Proof of Theorem 2.3.2

We will start by showing that the coverage is guaranteed. Recalling that we introduced the notation  $j^w$  to denote the step that the localization procedure chooses an interval relatively far away from the right one:

$$j^w = \min\{j : |\hat{i}_j - i_j^*| \geq 5\}. \quad (\text{A.1.56})$$

Then we know that  $|\hat{i}_{j^w-1} - i_{j^w-1}^*| \leq 4$ , so we have that  $|\hat{i}_{j^w+k} - i_{j^w+k}^*| \leq 6 * 2^{k+1} - 2$  for

all  $k \geq -1$ . Now we introduce the following lemma bounding the probability of stopping at least  $K + 1$  steps after  $j^w$ .

**Lemma A.1.8.** *For  $j^w$  defined in Equation (A.1.56), and for  $K \geq 0$ , we have*

$$P(\hat{j} \geq j^w + K + 1) \leq \Phi(-2)^K.$$

In particular, for  $K_\alpha = \lceil \frac{\log \alpha}{\log \Phi(-2)} \rceil$ ,  $P(\hat{j} \geq j^w + K_\alpha + 1) \leq \alpha$ .

Note that when  $\hat{j} \leq j^w + K_\alpha$ , we have  $|\hat{i}_{\hat{j}} - i_j^*| \leq 12 \cdot 2^{K_\alpha} - 2$ , implying that  $Z(f) \in [L, U]$ .

Therefore, we have

$$P(Z(f) \in CI_{z,\alpha}) \geq P(\hat{j} \leq j^w + K_\alpha) = 1 - P(\hat{j} \geq j^w + K_\alpha + 1) \geq 1 - \alpha. \quad \square$$

*Proof of Lemma A.1.8.* Now we will compute the probability that the stopping rule does not stop  $K$  steps after  $j^w$ . When  $j^w = \infty$ ,  $\hat{j}$  can never be larger than  $j^w$ , hence we can only consider the event  $\{j^w < \infty\}$ .

$$\begin{aligned} & \mathbb{E}_{l,s} \left( \mathbb{1}\{\hat{j} \geq j^w + K + 1\} \mathbb{1}\{j^w < \infty\} \right) \\ &= \mathbb{E}_{l,s} \left( \sum_{j_1=3}^{\infty} \mathbb{1}\{\hat{j} \geq j_1 + K + 1\} \mathbb{1}\{j^w = j_1\} \right) \\ &= \mathbb{E}_l \left( \sum_{j_1=3}^{\infty} \mathbb{E}_s(\mathbb{1}\{\hat{j} \geq j_1 + K + 1\} | Y_l) \mathbb{1}\{j^w = j_1\} \right) \tag{A.1.57} \\ &\leq \mathbb{E}_l \left( \sum_{j_1=3}^{\infty} \Phi(-2)^K \mathbb{1}\{j^w = j_1\} \right) \\ &\leq \Phi(-2)^K. \end{aligned}$$

The rationale for the first inequality in Equation (A.1.57) is as follows. Define the set of all possible sequences (i.e.  $(i_0, \dots, i_{j_1+1+K})$ ) starting from stage 0 to the stage  $j_1 + K + 1$  that satisfies  $|i_j - i_j^*| \leq 4$  for  $j \leq j_1$ , and  $|i_j - i_j^*| \geq 5$  for  $j = j_1$  as  $Se(j_1, K + 1)$ .  $\forall s \in Se(j_1, K + 1)$ , denote  $(i_l, \dots, i_h)$  in  $s$  as  $s(l, h)$ , and denote the sequence  $(\hat{i}_l, \dots, \hat{i}_h)$

produced by the localization procedure as  $\hat{s}(l, h)$ . If  $l = h$ , we will abbreviate  $s(l, l)$  into  $s(l)$  and  $\hat{s}(l, l)$  into  $\hat{s}(l)$ . Then we know that for  $s \in Se(j_1, K+1)$  with  $s(j_1) \leq i_{j_1}^* - 5$ , we have  $s(j) + 6 < i_j^*$  for  $j = j_1 + 1, \dots, K+1$ , therefore,  $\mu_{j,s(j)+6} - \mu_{j,s(j)+5} \leq 0$ . On the other hand, for  $s \in Se(j_1, K+1)$  with  $s(j_1) \geq i_{j_1}^* + 5$ , we have  $\mu_{j,s(j)-6} - \mu_{j,s(j)-5} \leq 0$ . Now we define a sign function indicating which side  $s(j)$  is on to  $i_j^*$ ,

$$Sg(s, j) = \text{sign}\{i_j^* - s(j)\}.$$

And for ease of expression, denote  $\tau_{j,i} = W_2(t_{j,i}) - W_2(t_{j,i-1})$ . Now we go back to the analysis of the first inequality in Equation (A.1.57):

$$\begin{aligned} & \mathbb{E}_s(\mathbb{1}\{\hat{j} \geq j_1 + K + 1\} | Y_l) \mathbb{1}\{j^w = j_1\} \\ &= \mathbb{E}_s\left(\sum_{s \in Se(j_1, K+1)} \mathbb{1}\{\hat{j} \geq j_1 + K + 1, \hat{s}(0, j_1 + 1 + K) = s\} | Y_l\right) \mathbb{1}\{j^w = j_1\} \\ &\leq \mathbb{E}_s\left(\sum_{s \in Se(j_1, K+1)} \mathbb{1}\{\min\{\tilde{X}_{j,s(j)+6} - \tilde{X}_{j,s(j)+5}, \tilde{X}_{j,s(j)-6} - \tilde{X}_{j,s(j)-5}\} \geq 2\sqrt{2}c_s\varepsilon\sqrt{m_j},\right. \\ &\quad \left.\forall j = j_1 + 1, \dots, j_1 + K\} \mathbb{1}\{\hat{s}(0, j_1 + 1 + K) = s\} | Y_l\right) \mathbb{1}\{j^w = j_1\} \\ &\leq \sum_{s \in Se(j_1, K+1)} \mathbb{E}_s\left(\mathbb{1}\{\tilde{X}_{j,s(j)+6}Sg(s, j) - \tilde{X}_{j,s(j)+5}Sg(s, j) \geq 2\sqrt{2}c_s\varepsilon\sqrt{m_j},\right. \\ &\quad \left.\forall j = j_1 + 1, \dots, j_1 + K\} \mathbb{1}\{\hat{s}(0, j_1 + 1 + K) = s\} | Y_l\right) \\ &\leq \sum_{s \in Se(j_1, K+1)} \mathbb{E}_s\left(\mathbb{1}\{m_j \cdot \mu_{j,s(j)+6}Sg(s, j) - m_j \cdot \mu_{j,s(j)+5}Sg(s, j) + \tau_{j,s(j)+6}Sg(s, j)\right. \\ &\quad \left.- \tau_{j,s(j)+5}Sg(s, j) \geq 2\sqrt{2}c_s\varepsilon\sqrt{m_j}, \forall j = j_1 + 1, \dots, j_1 + K\} \mathbb{1}\{\hat{s}(0, j_1 + 1 + K) = s\} | Y_l\right) \\ &\leq \sum_{s \in Se(j_1, K+1)} \mathbb{E}_s\left(\mathbb{1}\{\tau_{j,s(j)+6}Sg(s, j) - \tau_{j,s(j)+5}Sg(s, j) \geq 2\sqrt{2}c_s\varepsilon\sqrt{m_j},\right. \\ &\quad \left.\forall j = j_1 + 1, \dots, j_1 + K\} \mathbb{1}\{\hat{s}(0, j_1 + 1 + K) = s\} | Y_l\right) \\ &= \sum_{s \in Se(j_1, K+1)} \Phi(-2)^K \mathbb{E}_s\left(\mathbb{1}\{\hat{s}(0, j_1 + 1 + K) = s\} | Y_l\right) \\ &= \Phi(-2)^K \mathbb{1}\{j^w = j_1\}. \end{aligned}$$

(A.1.58)

□

Next we turn to the expected length, for which we introduce the following lemma for the length of the confidence interval for the minimizer.

**Lemma A.1.9** (Length of Confidence Interval for the Minimizer). *For  $0 < \alpha < 0.3$ , the expected length of the confidence interval given in (2.3.6) satisfies*

$$\mathbb{E}(CI_{z,\alpha}(Y)) \leq (24 \times 2^{K_\alpha} - 3) \times 17.5 \times \rho_z(\varepsilon; f) \leq C_{z,\alpha} L_{z,\alpha}(\varepsilon; f).$$

*Proof of Lemma A.1.9.* Recall that we have the following notation for indicating the stage where the localization procedure starts choosing the interval not close enough to the targeting interval:

$$\tilde{j} = \min\{j : |\hat{i}_j - i_j^*| \geq 2\}.$$

Now we have

$$\begin{aligned} \mathbb{E}(m_{\hat{j}}) &= \mathbb{E}(m_{\hat{j}} \mathbb{1}\{\hat{j} \geq j^* - 3\}) + \mathbb{E}(m_{\hat{j}} \mathbb{1}\{\hat{j} \leq j^* - 4\}) \\ &\leq 8m_{j^*} + \mathbb{E}(m_{\hat{j}} \mathbb{1}\{\hat{j} \geq \tilde{j}, \hat{j} \leq j^* - 4\}) + \mathbb{E}(m_{\hat{j}} \mathbb{1}\{\hat{j} \leq \tilde{j} - 1, \hat{j} \leq j^* - 4\}) \\ &\leq 2\rho_z(\varepsilon; f) + \mathbb{E}(m_{\hat{j}} \mathbb{1}\{\tilde{j} \leq \hat{j} \leq j^* - 4\}) + \sum_{j=1}^{j^*-4} m_j \mathbb{E}(\mathbb{1}\{\hat{j} = j, \tilde{j} \geq j + 1\}). \end{aligned} \tag{A.1.59}$$

We will bound the second and the third term in Equation (A.1.59) as follows:

$$\begin{aligned}
\mathbb{E}(m_{\tilde{j}} \mathbb{1}\{\tilde{j} \leq \hat{j} \leq j^* - 4\}) &\leq \mathbb{E}(m_{\tilde{j}} \mathbb{1}\{\tilde{j} \leq j^* - 4\}) \leq \sum_{j=1}^{j^*-4} m_j \mathbb{E}(\mathbb{1}\{\tilde{j} = j\}) \\
&\leq \sum_{j=1}^{j^*-4} m_j \mathbb{E}(\mathbb{1}\{X_{j,i_j^*+3} \leq X_{j,i_j^*+1}, t_{j,i_j^*+3} \leq 1\} + \mathbb{1}\{X_{j,i_j^*+2} \leq X_{j,i_j^*+1}, t_{j,i_j^*+1} \leq 1\} \\
&\quad + \mathbb{1}\{X_{j,i_j^*-3} \leq X_{j,i_j^*-1}, t_{j,i_j^*-3} \geq m_j\} + \mathbb{1}\{X_{j,i_j^*-2} \leq X_{j,i_j^*-1}, t_{j,i_j^*-2} \geq m_j\}) \\
&\quad + \mathbb{1}\{X_{j,i_j^*-4} \leq X_{j,i_j^*-1}, t_{j,i_j^*-4} \geq m_j\} + \mathbb{1}\{X_{j,i_j^*+4} \leq X_{j,i_j^*+1}, t_{j,i_j^*+4} \leq 1\}) \\
&\leq \sum_{j=1}^{j^*-4} 2^{j^*-j} m_{j^*} \times 2(\Phi(-\frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} m_j \sqrt{m_j} \frac{1}{c_l \sqrt{2\varepsilon}}) + \Phi(-\frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} 2m_j \sqrt{m_j} \frac{1}{c_l \sqrt{2\varepsilon}}) \\
&\quad + \Phi(-\frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} 3m_j \sqrt{m_j} \frac{1}{c_l \sqrt{2\varepsilon}})) \\
&\leq \sum_{j=1}^{j^*-4} 2^{j^*-j} \frac{\rho_z(\varepsilon; f)}{2} \left( \Phi(-2^{\frac{3}{2}(j^*-j)} (\frac{1}{8})^{\frac{3}{2}} \frac{1}{2c_l}) + \Phi(-2^{\frac{3}{2}(j^*-j)} \times 2 \times (\frac{1}{8})^{\frac{3}{2}} \frac{1}{2c_l}) \right. \\
&\quad \left. + \Phi(-2^{\frac{3}{2}(j^*-j)} \times 3 \times (\frac{1}{8})^{\frac{3}{2}} \frac{1}{2c_l}) \right) \\
&= 4 \sum_{j=1}^{j^*-4} 2^{j^*-j-3} \rho_z(\varepsilon; f) \left( \Phi(-2^{\frac{3}{2}(j^*-j-4)} \sqrt{2/3}) + \Phi(-\frac{1}{\sqrt{2}} 2^{\frac{3}{2}(j^*-j-3)} \sqrt{2/3}) \right. \\
&\quad \left. + \Phi(-3 \times 2^{\frac{3}{2}(j^*-j-3)} \sqrt{2/3}) \right) \\
&\leq 4\rho_z(\varepsilon; f) \times \sum_{j=1}^{\infty} (2^j \Phi(-\frac{1}{2\sqrt{3}} 2^{\frac{3}{2}j}) + 2^j \Phi(-\frac{1}{\sqrt{3}} 2^{\frac{3}{2}j}) + 2^j \Phi(-\frac{3}{2\sqrt{3}} 2^{\frac{3}{2}j})) \\
&\leq 8\rho_z(\varepsilon; f) \times (\Phi(-\sqrt{2/3}) + \Phi(-2\sqrt{2/3}) + \Phi(-\sqrt{6}) + \\
&\quad [2\Phi(-4/\sqrt{3}) + 2\Phi(-8/\sqrt{3}) + 2\Phi(-12/\sqrt{3})] \frac{1}{1 - 2^{\frac{\Phi(-8\sqrt{2/3})}{\Phi(-4/\sqrt{3})}}}).
\end{aligned} \tag{A.1.60}$$

The last inequality is due to  $\frac{\Phi(-2\sqrt{2}x)}{\Phi(-x)}$  decreases with  $x > 0$  increases. Now we turn to the



third term in Equation (A.1.59).

$$\begin{aligned}
& \sum_{j=1}^{j^*-4} m_j \mathbb{E}(\mathbb{1}\{\hat{j} = j, \tilde{j} \geq j+1\}) \\
&= \sum_{j=1}^{j^*-4} m_j \mathbb{E}_l(\mathbb{E}_s(\hat{j} = j | Y_l) \mathbb{1}\{\tilde{j} \geq j+1\}) \\
&\leq \sum_{j=1}^{j^*-4} m_j \mathbb{E}_l(\mathbb{E}_s(\tilde{X}_{j, \hat{i}_j+6} - \tilde{X}_{j, \hat{i}_j+5} \leq 2\sqrt{2}c_s \varepsilon \sqrt{m_j} | Y_l) \mathbb{1}\{\tilde{j} \geq j+1\}) + \\
&\quad \mathbb{E}_s(\tilde{X}_{j, \hat{i}_j-6} - \tilde{X}_{j, \hat{i}_j-5} \leq 2\sqrt{2}c_s \varepsilon \sqrt{m_j} | Y_l) \mathbb{1}\{\tilde{j} \geq j+1\}) \\
&\leq \sum_{j=1}^{j^*-4} m_j \mathbb{E}_l\left(2\Phi\left(2 - \frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} m_j \frac{\sqrt{m_j}}{c_s \sqrt{2\varepsilon}}\right) \mathbb{1}\{\tilde{j} \geq j+1\}\right) \\
&\leq \sum_{j=1}^{j^*-4} 2^{j^*-j} \frac{\rho_z(\varepsilon; f)}{4} \times 2\Phi\left(2 - \frac{2^{\frac{3}{2}(j^*-j-3)}}{2c_s}\right) \mathbb{E}_l(\mathbb{1}\{\tilde{j} \geq j+1\}) \\
&\leq \sum_{j=1}^{j^*-4} 2^{j^*-j} \frac{\rho_z(\varepsilon; f)}{4} \times 2\Phi\left(2 - 2^{\frac{3}{2}(j^*-j-4)} \sqrt{2/3}\right) \\
&\leq 8\rho_z(\varepsilon; f) \left(\Phi(2 - \sqrt{2/3}) + 2\Phi(2 - 4/\sqrt{3}) + 4\Phi(2 - 8 \times \sqrt{2/3}) \frac{1}{1 - 0.008}\right).
\end{aligned} \tag{A.1.61}$$

Combining them together, now we can turn to the original Equation (A.1.59), for which we have

$$\mathbb{E}(m_{\hat{j}}) < 17.5\rho_z(\varepsilon; f). \tag{A.1.62}$$

Therefore,

$$\mathbb{E}(CI_{z, \alpha}) \leq (24 \times 2^{K_\alpha} - 3) \times \mathbb{E}(m_{\hat{j}}) \leq (24 \times 2^{K_\alpha} - 3) \times 17.5\rho_z(\varepsilon; f). \tag{A.1.63}$$

Since  $L_{z, \alpha}(\varepsilon; f) \geq b_\alpha \omega_z(\varepsilon/3; f) \geq b_\alpha \rho_z(\varepsilon; f)/3$  when  $0 < \alpha < 0.3$ , we have the statement.

□

### A.1.7. Proof of Theorem 2.3.3

In addition to the notation introduced in the proof of Theorem 2.3.1, we define the following bias and variance terms.

$$\hat{f} = \frac{1}{m_{\hat{j}}} \int_{t_{\hat{i}_{\hat{j}}+\Delta-1}^{t_{\hat{i}_{\hat{j}}+\Delta}} f(t) dt, \quad (\text{A.1.64})$$

$$\hat{\mathfrak{Z}} = \frac{1}{m_{\hat{j}}} (W_3(t_{\hat{i}_{\hat{j}}+\Delta}) - W_3(t_{\hat{i}_{\hat{j}}+\Delta-1})), \quad (\text{A.1.65})$$

where

$$\Delta = 2 \left( \mathbb{1}\{\tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}+6} - \tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}+5} \leq 2\sigma_j\} - \mathbb{1}\{\tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}-6} - \tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}-5} \leq 2\sigma_j\} \right).$$

Therefore, we have:

$$\begin{aligned} \mathbb{E}_{l,s,e}((\hat{M} - M(f))^2) &= \mathbb{E}_{l,s,e}((\hat{f} - M(f))^2 + \hat{\mathfrak{Z}}^2 + 2\hat{\mathfrak{Z}}(\hat{f} - M(f))) \\ &= \mathbb{E}_{l,s}((\hat{f} - M(f))^2 + \frac{3\varepsilon^2}{m_{\hat{j}}}) \leq \mathbb{E}_{l,s}((\hat{f} - M(f))^2) + \frac{24\varepsilon^2}{\rho_z(\varepsilon; f)} \mathbb{E}(2^{\hat{j}-j^*}). \end{aligned} \quad (\text{A.1.66})$$

The second equation is because  $Y_l$ ,  $Y_s$  and  $Y_e$  are mutually independent, and taking the conditional expectation leads to the equation.

For the second term of the right hand side of Inequality (A.1.66), we have the following lemma that we will prove later:

**Lemma A.1.10.**

$$\mathbb{E}(2^{\hat{j}-j^*}) < \frac{35}{4} \frac{\rho_z(\varepsilon; f) \rho_m(\varepsilon; f)^2}{\varepsilon^2}. \quad (\text{A.1.67})$$

For the first term of the right hand side of Inequality (A.1.66), we have

$$\begin{aligned}
& \mathbb{E}_{l,s}((\hat{f} - M(f))^2) \\
&= \mathbb{E}_{l,s}((\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}) + \mathbb{E}_{l,s}((\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} > \hat{j}\}).
\end{aligned} \tag{A.1.68}$$

And for the first term in Equation A.1.68, we have

$$\begin{aligned}
& \mathbb{E}_{l,s}((\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}) \\
& \leq \mathbb{E}_{l,s}\left(\left((\hat{f} - \mu_{\hat{j}, \hat{i}_{\tilde{j}}})_+ + (\mu_{\hat{j}, \hat{i}_{\tilde{j}}} - M(f))\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) \\
& \leq 2\mathbb{E}_{l,s}\left(\left((\hat{f} - \mu_{\hat{j}, \hat{i}_{\tilde{j}}})_+\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) + 2\mathbb{E}_{l,s}\left(\left(\mu_{\hat{j}, \hat{i}_{\tilde{j}}} - M(f)\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) \\
& \leq 2\mathbb{E}_{l,s}\left(\left((\hat{f} - \mu_{\hat{j}, \hat{i}_{\tilde{j}}})_+\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) + 2\mathbb{E}_{l,s}\left(\left(\left(\mu_{\hat{j}, \hat{i}_{\tilde{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}}^+\right) + (\mu_{\tilde{j}, \hat{i}_{\tilde{j}}} - M(f))\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) \\
& \leq 2\mathbb{E}_{l,s}\left(\left((\hat{f} - \mu_{\hat{j}, \hat{i}_{\tilde{j}}})_+\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) + 2\mathbb{E}_{l,s}\left(\left(\frac{4}{3}\left((\mu_{\hat{j}, \hat{i}_{\tilde{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}}^+)\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right)\right. \\
& \quad \left. + 2\mathbb{E}_{l,s}\left(4(\mu_{\tilde{j}, \hat{i}_{\tilde{j}}} - M(f))^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right)\right).
\end{aligned} \tag{A.1.69}$$

Therefore, going back to Inequality A.1.68, we have

$$\begin{aligned}
& \mathbb{E}_{l,s}((\hat{f} - M(f))^2) \\
& \leq 2\mathbb{E}_{l,s}\left(\left((\hat{f} - \mu_{\hat{j}, \hat{i}_{\tilde{j}}})_+\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) + \mathbb{E}_{l,s}((\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} > \hat{j}\}) \\
& \quad + 2\mathbb{E}_{l,s}\left(\left(\frac{4}{3}\left((\mu_{\hat{j}, \hat{i}_{\tilde{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}}^+)\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) + 2\mathbb{E}_{l,s}\left(4(\mu_{\tilde{j}, \hat{i}_{\tilde{j}}} - M(f))^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right)\right).
\end{aligned} \tag{A.1.70}$$

To bound each term in Inequality (A.1.70), we introduce and prove the following proposition.

**Proposition A.1.3.**

$$\mathbb{E}_{l,s}((\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} > \hat{j}\}) < 12003\rho_m(\varepsilon; f)^2, \tag{A.1.71}$$

$$\mathbb{E}_{l,s}\left(\left((\hat{f} - \mu_{\hat{j}, \hat{i}_{\tilde{j}}})_+\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) < 13064\rho_m(\varepsilon; f)^2, \tag{A.1.72}$$

$$\mathbb{E}_{l,s}\left(\left(\mu_{\tilde{j}, \hat{i}_{\tilde{j}}} - M(f)\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) < 3104\rho_m(\varepsilon; f)^2, \tag{A.1.73}$$

$$\mathbb{E}_{l,s}\left(\left(\left(\mu_{\hat{j}, \hat{i}_{\tilde{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}}^+\right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}\right) < 50857\rho_m(\varepsilon; f)^2. \tag{A.1.74}$$

With proposition applied to Inequality (A.1.70) and Lemma A.1.10, and going back to Inequality (A.1.66), we have the statement of the theorem.

Now we are left with proving Proposition A.1.3 and Lemma A.1.10. Before we proceed, we introduce and prove a lemma

**Lemma A.1.11.**  $P(\hat{j} < \infty) = 1$ .

*Proof.* To prove this, we only need to prove  $\lim_{j \rightarrow \infty} P(\hat{j} > j) = 0$ . Suppose  $j \geq j^* + 3$ . For  $j_1 \geq j^* + 2$ ,

$$\min\{\mu_{j_1, \hat{i}_{j_1}+6} - \mu_{j_1, \hat{i}_{j_1}+5}, \mu_{j_1, \hat{i}_{j_1}-6} - \mu_{j_1, \hat{i}_{j_1}-5}\} < 13.5m_{j_1} \frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)}.$$

Since

$$3\varepsilon^2 \geq \rho_z(\varepsilon; f)\rho_m(\varepsilon; f)^2 \geq \frac{1}{2}\varepsilon^2, \quad (\text{A.1.75})$$

we have

$$\begin{aligned} & \min\{\mu_{j_1, \hat{i}_{j_1}+6} - \mu_{j_1, \hat{i}_{j_1}+5}, \mu_{j_1, \hat{i}_{j_1}-6} - \mu_{j_1, \hat{i}_{j_1}-5}\}m_{j_1}/(c_s\sqrt{2m_{j_1}\varepsilon^2}) \\ & \leq \frac{1}{c_s\sqrt{2}\varepsilon} 2^{\frac{-3(j_1-j^*)}{2}} m_{j^*}^{\frac{3}{2}} \frac{13.5\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} \\ & \leq \frac{\sqrt{3}}{c_s\sqrt{2}\rho_m(\varepsilon; f)\sqrt{\rho_z(\varepsilon; f)}} \rho_z(\varepsilon; f)^{\frac{3}{2}} \frac{13.5\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} 2^{\frac{-3(j_1-j^*+2)}{2}} \\ & \leq 13.5 \cdot 2^{\frac{-3(j_1-j^*+2)}{2}}. \end{aligned} \quad (\text{A.1.76})$$

Therefore,

$$\begin{aligned} & P(\hat{j} > j) \\ & = \mathbb{E}_l \left( \mathbb{E}_s(\Pi_{j_1=j^*+2}^{j-1} \mathbb{1}\{\min\{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}, \tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}\} > 2\sigma_{j_1}\} | Y_l) \right) \\ & \leq \mathbb{E}_l \left( \Pi_{j_1=j^*+2}^{j-1} \Phi(-2 + 13.5 \cdot 2^{\frac{-3(j_1-j^*+2)}{2}}) \right) < \Phi(-1.85)^{j-j^*-2}. \end{aligned} \quad (\text{A.1.77})$$

Therefore,  $\lim_{j \rightarrow \infty} P(\hat{j} > j) \leq \lim_{j \rightarrow \infty} \Phi(-1.85)^{j-j^*-2} = 0$ .

□

Continuing with the proof of the Proposition A.1.3, we have the following lemmas that we will prove in the Section A.2 (page 212, 217, 219 and 221).

**Lemma A.1.12.**

$$\begin{aligned} & \mathbb{E}_{l,s}((\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} > \hat{j}\}) \\ & \leq (5760V + 2)\rho_m(\varepsilon; f)^2 + 78V\rho_m(\varepsilon; f)^2 + \frac{1}{16}\rho_m(\varepsilon; f)^2, \end{aligned} \quad (\text{A.1.78})$$

where  $V = \sup_{x \geq 0} x^2 \Phi(2 - x)$ .

**Lemma A.1.13.**

$$\mathbb{E}_{l,s} \left( ((\hat{f} - \mu_{\hat{j}, \hat{i}_{\tilde{j}}})_+)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\} \right) \leq 6355.2V\rho_m(\varepsilon; f)^2, \quad (\text{A.1.79})$$

where  $V = \sup_{x \geq 0} x^2 \Phi(2 - x)$ .

**Lemma A.1.14.**

$$\mathbb{E}_{l,s} \left( (\mu_{\hat{j}, \hat{i}_{\tilde{j}}} - M(f))^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\} \right) < 3 \times (2^8 + 2^8 \frac{\Phi(-1.85)}{(1 - \Phi(-2 + \frac{1}{12}))^2}) \rho_m(\varepsilon; f)^2 (23\frac{1}{8})Q, \quad (\text{A.1.80})$$

where  $Q = \sup_{x \geq 0} x^2 \Phi(-x)$ .

**Lemma A.1.15.**

$$\mathbb{E}_{l,s} \left( \left( (\mu_{\hat{j}, \hat{i}_{\tilde{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}})_+ \right)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\} \right) \leq Q \times 277075 \times \rho_m(\varepsilon; f)^2 + Q \times 23850.1 \times \rho_m(\varepsilon; f)^2, \quad (\text{A.1.81})$$

where  $Q = \sup_{x \geq 0} x^2 \Phi(-x)$ .

These four lemmas combined with Lemma A.1.3 give the statement of Proposition A.1.3.

Finally we will prove Lemma A.1.10.

*Proof of Lemma A.1.10.* Note that this lemma is used to bound the term  $\frac{8\varepsilon^2 c_e^2}{\rho_z(\varepsilon; f)} \mathbb{E}(2^{\hat{j}-j^*})$ ,

so in the proof we will start with bounding this term. We have

$$\begin{aligned}
& \frac{8\varepsilon^2 c_e^2}{\rho_z(\varepsilon; f)} \mathbb{E}(2^{\hat{j}-j^*}) \\
& \leq 16c_e^2 \rho_m(\varepsilon; f)^2 \mathbb{E}(2^{\hat{j}-j^*}) \\
& = 16c_e^2 \rho_m^2 [\mathbb{E}(2^{\hat{j}-j^*} \mathbb{1}\{\hat{j} \leq j^* + 2\}) + \mathbb{E}(\{2^{\hat{j}-j^*} \mathbb{1}\{\hat{j} \geq j^* + 3\})] \\
& \leq 64c_e^2 \rho_m(\varepsilon; f)^2 + 16c_e^2 \rho_m(\varepsilon; f)^2 (2^{\hat{j}-j^*} \mathbb{1}\{\hat{j} \geq j^* + 3\}) \\
& = 64c_e^2 \rho_m(\varepsilon; f)^2 \\
& \quad + 16c_e^2 \rho_m(\varepsilon; f)^2 \mathbb{E}_{l,s} \left( \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{1}\{\hat{j} = j_1, t_{\hat{j}, \hat{i}_{\hat{j}}} \leq \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right) \\
& \quad + \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{1}\{\hat{j} = j_1, t_{\hat{j}} > \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\}.
\end{aligned} \tag{A.1.82}$$

Now we will bound the second term and the third term in the Inequality (A.1.82). Without loss of generality, we can assume

$$\sup\{t > Z(f) : f(t) \leq \rho_m(\varepsilon; f) + M(f)\} = \rho_z(\varepsilon; f) + Z(f),$$

because otherwise the following would hold

$$\min\{t < Z(f) : f(t) \leq \rho_m(\varepsilon; f) + M(f)\} = Z(f) - \rho_z(\varepsilon; f).$$

and one only need to flip everything around with  $Z(f)$  being the center. Then for the

second term we have

$$\begin{aligned}
& 16c_e^2 \rho_m(\varepsilon; f)^2 \mathbb{E}_{l,s} \left( \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{1}\{\hat{j} = j_1, t_{\hat{j}, \hat{i}_{\hat{j}}} \leq \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right) \\
&= 16c_e^2 \rho_m(\varepsilon; f)^2 \mathbb{E}_{l,s} \left[ \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{1}\{\hat{j} = j_1, t_{j_1, \hat{i}_{j_1}} \leq \frac{7\rho_m(\varepsilon; f)}{16} + Z(f), \right. \\
&\quad \left. \forall j^*+2 \leq j \leq j_1-1, \mathcal{E}_{j, \hat{i}_j+6} \frac{1}{\sqrt{2}c_s \varepsilon} \geq 2 - \frac{\sqrt{m_j}}{\sqrt{2}c_s \varepsilon} (\mu_{j, \hat{i}_j+6} - \mu_{j, \hat{i}_j+5}) \} \right] \\
&\leq 16c_e^2 \rho_m(\varepsilon; f)^2 \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_l \left[ \mathbb{E}_s \left( \mathbb{1}\{\hat{j} = j_1, t_{j_1, \hat{i}_{j_1}} \leq \frac{7\rho_m(\varepsilon; f)}{16} + Z(f), \right. \right. \\
&\quad \left. \left. \forall j^*+2 \leq j \leq j_1-1, \mathcal{E}_{j, \hat{i}_j+6} \frac{1}{\sqrt{2}c_s} \geq 2 - \frac{\sqrt{m_j}}{\sqrt{2}c_s \varepsilon} (\mu_{j, \hat{i}_j+6} - \mu_{j, \hat{i}_j+5}) \} | Y_l \right) \right] \\
&\leq 16c_e^2 \rho_m(\varepsilon; f)^2 \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_l \left[ \mathbb{1}\{t_{j_1, \hat{i}_{j_1}} \leq \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right. \\
&\quad \left. \mathbb{E}_s \left( \mathbb{1}\{\forall j^*+2 \leq j \leq j_1-1, \mathcal{E}_{j, \hat{i}_j+6} \geq 2 - \frac{\sqrt{m_j}}{\sqrt{2}c_s \varepsilon} \frac{\rho_m(\varepsilon; f)(\frac{7}{16}\rho_z(\varepsilon; f) + 6m_j)}{\rho_z(\varepsilon; f)}\} | Y_l \right) \right] \\
&\leq 16c_e^2 \rho_m(\varepsilon; f)^2 \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_l \left[ \mathbb{1}\{t_{j_1, \hat{i}_{j_1}} \leq \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right. \\
&\quad \left. \Pi_{j=j^*+2}^{j_1-1} \Phi \left( -2 + \frac{\rho_m(\varepsilon; f) \sqrt{\rho_z(\varepsilon; f)}}{\varepsilon} \frac{2^{\frac{j^*-j-2}{2}}}{\sqrt{2}c_s} \left( \frac{7}{16} + 6 * 2^{j^*-j-2} \right) \right) \right] \\
&\leq 16c_e^2 \rho_m(\varepsilon; f)^2 \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_l \left[ \mathbb{1}\{t_{j_1, \hat{i}_{j_1}} \leq \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right. \\
&\quad \left. \Pi_{j=j^*+2}^{j_1-1} \Phi \left( -2 + \frac{2^{\frac{j^*-j-2}{2}}}{\sqrt{2}} \left( \frac{7}{16} + 6 * 2^{j^*-j-2} \right) \right) \right] \\
&\leq 16c_e^2 \rho_m(\varepsilon; f)^2 \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_l \left[ \mathbb{1}\{t_{j_1, \hat{i}_{j_1}} \leq \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right] \Phi(-1.8)^{j_1-j^*-2}.
\end{aligned}
\tag{A.1.83}$$

Now we go to the third term in the Inequality (A.1.82).

$$\begin{aligned}
& 16c_e^2 \rho_m(\varepsilon; f)^2 \mathbb{E}_{l,s} \left( \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{1}\{\hat{j} = j_1, t_{\hat{j}} > \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right) \\
&= 16c_e^2 \rho_m(\varepsilon; f)^2 \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_{l,s} \left( \mathbb{1}\{\hat{j} = j_1, t_{j_1} > \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right) \\
&= 16c_e^2 \rho_m(\varepsilon; f)^2 \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_{l,s} \left( \mathbb{1}\{\hat{j} = j_1, t_{j_1} > \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\}, \right. \\
&\quad \left. \forall j^*+2 \leq j \leq j_1-1, -\mathcal{E}_{j, \hat{i}_j-5} \geq 2 - \frac{\sqrt{m_j}}{c_s} (\mu_{j, \hat{i}_j-6} - \mu_{j, \hat{i}_j-5}) \right) \tag{A.1.84} \\
&\leq 16c_e^2 \rho_m(\varepsilon; f)^2 \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_l \left[ \mathbb{1}\{t_{j_1} > \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right. \\
&\quad \left. \mathbb{E}_s \left( \mathbb{1}\{\forall j^*+2 \leq j \leq j_1-1, -\mathcal{E}_{j, \hat{i}_j-5} \frac{1}{\sqrt{m_j} c_s \varepsilon} \geq 2\} | Y_l \right) \right] \\
&\leq 16c_e^2 \rho_m(\varepsilon; f)^2 \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_l \left[ \mathbb{1}\{t_{\hat{j}} > \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right] \Phi(-2)^{j_1-j^*-2}.
\end{aligned}$$

Combining Inequality (A.1.83) and Inequality (A.1.84), back to the original Inequality (A.1.82)

$$\begin{aligned}
& \frac{8\varepsilon^2 c_e^2}{\rho_z(\varepsilon; f)} \mathbb{E}(2^{\hat{j}-j^*}) \\
&\leq 64c_e^2 \rho_m(\varepsilon; f)^2 + 16c_e^2 \rho_m(\varepsilon; f)^2 \left( \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_l \left[ \mathbb{1}\{t_{j_1, \hat{i}_{j_1}} \leq \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right] \Phi(-1.8)^{j_1-j^*-2} \right. \\
&\quad \left. + \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \mathbb{E}_l \left[ \mathbb{1}\{t_{\hat{j}} > \frac{7\rho_m(\varepsilon; f)}{16} + Z(f)\} \right] \Phi(-2)^{j_1-j^*-2} \right) \tag{A.1.85} \\
&\leq 64c_e^2 \rho_m(\varepsilon; f)^2 + 16c_e^2 \rho_m(\varepsilon; f)^2 \sum_{j_1 \geq j^*+3} 2^{j_1-j^*} \Phi(-1.8)^{j_1-j^*-2} \\
&= 64c_e^2 \rho_m(\varepsilon; f)^2 + 16c_e^2 \rho_m(\varepsilon; f)^2 * 8\Phi(-1.8) * \frac{1}{1-2\Phi(-1.8)} \\
&< 70c_e^2 \rho_m(\varepsilon; f)^2 = 210\rho_m(\varepsilon; f)^2.
\end{aligned}$$



Therefore, we have

$$\mathbb{E}(2^{\hat{j}-j^*}) \leq \frac{35}{4} \frac{\rho_m(\varepsilon; f)^2 \rho_z(\varepsilon; f)}{\varepsilon}.$$

□

#### A.1.8. Proof of Theorem 2.3.4

We will prove the following two lemmas separately, which give rise to the theorem.

**Lemma A.1.16** (Coverage of the Confidence Interval for the Minimum). *For any  $0 < \alpha < 1$ , the confidence interval  $CI_{m,\alpha}$  given in (2.3.10) is a  $1 - \alpha$  confidence interval.*

**Lemma A.1.17** (Length of the Confidence Interval for the Minimum). *For  $0 < \alpha < 1$ , the expected length of the confidence interval given in (2.3.10) satisfies*

$$\mathbb{E}(|f_{hi} - f_{lo}|) \leq c_{m,\alpha} \rho_m(\varepsilon; f), \text{ for all } f \in \mathcal{F},$$

where  $c_{m,\alpha}$  is a constant depending only on  $\alpha$ .

Further, when  $0 < \alpha < 0.3$ , we have

$$\mathbb{E}(|f_{hi} - f_{lo}|) \leq c_{m,\alpha} \rho_m(\varepsilon; f) \leq C_{m,\alpha} L_{m,\alpha}(\varepsilon; f), \text{ for all } f \in \mathcal{F},$$

where  $C_{m,\alpha}$  is an absolute constant depending only on  $\alpha$ .

*Proof of Lemma A.1.16.* Define five events:

$$\begin{aligned} E &= \{Z(f) \notin [t_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+, \hat{i}_{\hat{j}-K_{\frac{\alpha}{4}}-1}-5}, t_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+, \hat{i}_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+}+4}]\} \\ E_1 &= \{\hat{j} \geq j^w + K_{\frac{\alpha}{4}} + 1\} \\ F &= \{\hat{j} \leq j^* - 2 - \tilde{K}_{\frac{\alpha}{4}}\} \\ G &= \{f_{hi} < M(f)\} \\ H &= \{f_{lo} > M(f)\}. \end{aligned} \tag{A.1.86}$$

By definition  $\{M(f) \in [f_{lo}, f_{hi}]\} = G^c \cap H^c$ . We will bound the probabilities of the above events.

Recalling  $K_\alpha = \lceil \frac{\log \alpha}{\log \Phi(-2)} \rceil$ , then with Lemma A.1.8 we have

$$P(\hat{j} \geq j^w + K_\alpha + 1) \leq \alpha,$$

so  $P(E_1) \leq \frac{\alpha}{4}$ .

When the event  $E_1^c = \{\hat{j} \leq j^w + K_\alpha\}$  occurs, we have

$$Z(f) \in [t_{(\hat{j}-K_\alpha-1)_+, \hat{i}_{(\hat{j}-K_\alpha-1)_+}} - 5, t_{(\hat{j}-K_\alpha-1)_+, \hat{i}_{(\hat{j}-K_\alpha-1)_+} + 4}],$$

so  $P(E) \leq \frac{\alpha}{4}$ .

To bound  $P(F)$ , we introduce the following lemma (proved in Section A.2 page 226) showing the procedure can not stop too early.

**Lemma A.1.18.** *When  $\tilde{K} \geq 4$ , we have*

$$P(\hat{j} \leq j^* - 2 - \tilde{K}) \leq \Phi(-2^{\frac{3}{2}(\tilde{K}-2)-\frac{1}{2}} + 2) \frac{2}{1 - \exp(-40)}.$$

Now with this lemma, and take  $\tilde{K}_\alpha = \max\{4, 2 + \lceil \log_2(2 - \Phi^{-1}(\frac{\alpha}{3})) \rceil\} > \max\{4, 2 + \lceil \frac{2}{3} \log_2 \max\{2 - \Phi^{-1}((1 - e^{-40})\frac{\alpha}{2}), 1\} + \frac{1}{3} \rceil\}$ , we know that

$$P(\hat{j} \leq j^* - 2 - \tilde{K}_\alpha) \leq \alpha.$$

This gives  $P(F) \leq \frac{\alpha}{4}$ .

Now we will introduce two more lemmas that build up the remaining foundation of the proof, which are proved in Section A.2 (page 226 and 227).

**Lemma A.1.19.**

$$P(G \mid Z(f) \in [t_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+, \hat{i}_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+} - 5}, t_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+, \hat{i}_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+} + 4}]) \leq \frac{\alpha}{4}. \quad (\text{A.1.87})$$

**Lemma A.1.20.**

$$P(H|E^c \cap F^c) \leq \frac{\alpha}{4}. \quad (\text{A.1.88})$$

With these additional lemmas, we have

$$\begin{aligned} & P(M(f) \in CI_{m,\alpha}(Y)) \\ & \geq P(E^c \cap F^c \cap G^c \cap H^c) \\ & \geq P(E_1^c \cap F^c \cap G^c \cap H^c) \\ & \geq (1 - P(H|E_1^c \cap F^c) - P(G|E_1^c \cap F^c))P(E_1^c \cap F^c) \\ & \geq -P(H|E_1^c \cap F^c)P(E_1^c \cap F^c) + P(G^c \cap E_1^c \cap F^c) \\ & \geq -\frac{\alpha}{4}P(E_1^c \cap F^c) + P(G^c \cap E_1^c) - P((G^c \cap E_1^c) \cap F) \\ & \geq -\frac{\alpha}{4} + P(E_1^c) - P(E_1^c \cap G) - P(F) \\ & \geq -\frac{\alpha}{4} + 1 - P(E_1) - P(G|E_1^c) - P(F) \\ & \geq -\frac{\alpha}{4} + 1 - \frac{\alpha}{4} - \frac{\alpha}{4} - \frac{\alpha}{4} = 1 - \alpha. \end{aligned} \quad (\text{A.1.89})$$

□

*Proof of Lemma A.1.17.*

$$\begin{aligned} & \mathbb{E}(|f_{hi} - f_{lo}|) \\ & = \mathbb{E}((S_{i_R - i_L, \frac{\alpha}{4}} c_e + z_{\frac{\alpha}{4}} c_e + \sqrt{3}) \frac{\varepsilon}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}}) \\ & < (S_{i_R - i_L, \frac{\alpha}{4}} + z_{\frac{\alpha}{4}} + \sqrt{3}) \frac{2^{\frac{\tilde{K}_{\frac{\alpha}{4}}}{2}} c_e \varepsilon}{\sqrt{m_{j^*}}} \mathbb{E}(2^{\frac{1}{2}(\hat{j} - j^*)}) \\ & \leq (S_{i_R - i_L, \frac{\alpha}{4}} + z_{\frac{\alpha}{4}} + \sqrt{3}) 2^{\frac{\tilde{K}_{\frac{\alpha}{4}}}{2}} c_e \cdot 4\rho_m(\varepsilon; f) \mathbb{E}(2^{\frac{1}{2}(\hat{j} - j^*)}). \end{aligned} \quad (\text{A.1.90})$$

Similarly to the way we bound variance in Theorem 2.3.3, we have

$$\begin{aligned}
& \mathbb{E}(2^{\frac{1}{2}(\hat{j}-j^*)}) \\
& \leq 2\mathbb{E}(\mathbb{1}\{\hat{j} \leq j^* + 2\}) + \mathbb{E}(2^{\frac{1}{2}(\hat{j}-j^*)}\mathbb{1}\{\hat{j} \geq j^* + 3\}) \\
& \leq 2 + 2\sqrt{2}\Phi(-1.85)\frac{1}{1 - 2\Phi(-1.85)} \\
& < 2.16.
\end{aligned} \tag{A.1.91}$$

According to the definition of  $S_{i_R-i_L, \frac{\alpha}{4}}$ ,  $S_{i_R-i_L, \frac{\alpha}{4}}$  is decided by the following inequality

$$(1 - \Phi(-S_{i_R-i_L, \frac{\alpha}{4}}))^{i_R-i_L} \geq 1 - \frac{\alpha}{4}. \tag{A.1.92}$$

Therefore,

$$S_{i_R-i_L, \frac{\alpha}{4}} = -\Phi^{-1}(1 - (1 - \frac{\alpha}{4})^{\frac{1}{i_R-i_L}}). \tag{A.1.93}$$

Furthermore, we have

$$\begin{aligned}
& i_R - i_L \\
& = 9 \times 2 \times 2^{\tilde{K}_{\frac{\alpha}{4}}} \times 2^{K_{\frac{\alpha}{4}}},
\end{aligned} \tag{A.1.94}$$

so we know that  $(S_{i_R-i_L, \frac{\alpha}{4}} + z_{\frac{\alpha}{4}} + \sqrt{3})2^{\frac{\tilde{K}_{\frac{\alpha}{4}}}{2}}c_e$  only depend on  $\alpha$ . Therefore,

$$\mathbb{E}(|f_{hi} - f_{lo}|) \leq c_{m,\alpha}\rho_m(\varepsilon; f). \tag{A.1.95}$$

Since for  $0 < \alpha < 0.3$ , we have

$$\rho_m(\varepsilon; f) \leq 3\rho_m(\varepsilon/3; f) \leq \frac{3}{b_\alpha}L_{m,\alpha}(\varepsilon; f),$$

we get our statement.

□

### A.1.9. Analysis of Lower Bounds of the Benchmarks in Regression Setting

To establish the optimality of the procedures, we need to analyze the lower bounds of the benchmarks. Compared with the white noise model, we will incur an additional discretization error.

Define the discretization errors for  $Z(f)$  and  $M(f)$  as

$$\begin{aligned}\mathfrak{D}_z(n, f) &= \max\{Z(g) : g \in \mathcal{F}, g(x_i) = f(x_i), i = 0, \dots, n\} \\ &\quad - \min\{Z(g) : g \in \mathcal{F}, g(x_i) = f(x_i), i = 0, \dots, n\}\end{aligned}\tag{A.1.96}$$

$$\begin{aligned}\mathfrak{D}_m(n, f) &= \max\{M(g) : g \in \mathcal{F}, g(x_i) = f(x_i), i = 0, \dots, n\} \\ &\quad - \min\{M(g) : g \in \mathcal{F}, g(x_i) = f(x_i), i = 0, \dots, n\}\end{aligned}\tag{A.1.97}$$

It is easy to see that  $0 \leq \mathfrak{D}_z(n, f) < \frac{2}{n}$  and any value in  $[0, \frac{2}{n})$  can be taken by  $\mathfrak{D}_z(n, f)$  for some  $f \in \mathcal{F}$ .

The lower bounds for the benchmarks are given as follows.

**Proposition A.1.4.** *Let  $\tilde{R}_{z,n}(\sigma; f)$ ,  $\tilde{R}_{m,n}(\sigma; f)$ ,  $\tilde{L}_{z,\alpha,n}(\sigma; f)$ ,  $\tilde{L}_{m,\alpha,n}(\sigma; f)$  be defined as in (2.4.2). Suppose  $0 < \alpha < 0.3$ . There exist constants  $\tilde{C}_z, \tilde{C}_m, \tilde{C}_{z,\alpha}, \tilde{C}_{m,\alpha} > 0$  such that for all  $f \in \mathcal{F}$ ,*

$$\begin{aligned}\tilde{R}_{z,n}(\sigma; f) &\geq \tilde{C}_z \sup_{g \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right) \left(1 \wedge \sqrt{n\rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right)}\right) \vee \frac{1}{4}\mathfrak{D}_z(n, f), \\ \tilde{R}_{m,n}(\sigma; f) &\geq \tilde{C}_m \sup_{g \in \mathcal{G}_n(f)} \rho_m\left(\frac{\sigma}{\sqrt{n}}; g\right) \left(1 \wedge \sqrt{n\rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right)}\right) \vee \frac{1}{4}\mathfrak{D}_m(n, f), \\ \tilde{L}_{z,\alpha,n}(\sigma; f) &\geq \tilde{C}_{z,\alpha} \sup_{g \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right) \left(1 \wedge \sqrt{n\rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right)}\right) \vee \frac{(1-2\alpha)}{2}\mathfrak{D}_z(n, f), \\ \tilde{L}_{m,\alpha,n}(\sigma; f) &\geq \tilde{C}_{m,\alpha} \sup_{g \in \mathcal{G}_n(f)} \rho_m\left(\frac{\sigma}{\sqrt{n}}; g\right) \left(1 \wedge \sqrt{n\rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right)}\right) \vee \frac{(1-2\alpha)}{2}\mathfrak{D}_m(n, f),\end{aligned}\tag{A.1.98}$$

where

$$\mathcal{G}_n(f) = \{g \in \mathcal{F} : g(x_i) = f(x_i), \text{ for all } 0 \leq i \leq n\}. \quad (\text{A.1.99})$$

Compared with the lower bounds in the white noise model, the lower bounds in the regression model contain additional discretization errors, which are in general non-vanishing for fixed  $n$  as the noise level  $\sigma \rightarrow 0$ .

*Proof of Proposition A.1.4.* Similar to white noise model. The probability density under truth  $f$  is:

$$p(y_0, \dots, y_n | f) = \prod_{i=0}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i))^2}{2\sigma^2}\right).$$

Hence

$$\frac{p(y_0, \dots, y_n | f)}{p(y_0, \dots, y_n | g)} = \exp\left(\frac{\sum_{i=0}^n (f(x_i) - g(x_i))(2y_i - f(x_i) - g(x_i))}{2\sigma^2}\right).$$

Let  $\theta_1 = 1$ , denoting the truth is  $f$ ,  $\theta_2 = -1$ , denoting the truth is  $g$ . And suppose  $\hat{\theta}$  is an estimator of  $\theta$ . Then we know that  $\frac{\sum_{i=0}^n (f(x_i) - g(x_i))(y_i - \frac{1}{2}f(x_i) - \frac{1}{2}g(x_i))}{\sigma^2}$  is sufficient statistic for  $\theta$ , we further standardize this statistic by  $l_n(f, g) = \sqrt{\sum_{i=0}^n \frac{1}{n} (f(x_i) - g(x_i))^2}$  and  $\sigma$ ,

$$\check{W} = \frac{\sum_{i=0}^n (f(x_i) - g(x_i))(y_i - \frac{1}{2}f(x_i) - \frac{1}{2}g(x_i))}{l_n(f, g)\sqrt{n}\sigma} \sim N\left(\theta \frac{l_n(f, g)}{2\frac{\sigma}{\sqrt{n}}}, 1\right).$$

Then let  $\hat{\theta} = \frac{2\hat{Z} - (Z(f) + Z(g))}{Z(f) - Z(g)}$ . We know that

$$\begin{aligned} \mathbb{E}_f(|\hat{Z} - Z(f)|) &= |Z(f) - Z(g)| \mathbb{E}_{\theta=1}\left(\frac{1}{2}|\hat{\theta} - \theta|\right), \\ \mathbb{E}_g(|\hat{Z} - Z(g)|) &= |Z(f) - Z(g)| \mathbb{E}_{\theta=-1}\left(\frac{1}{2}|\hat{\theta} - \theta|\right). \end{aligned}$$

Therefore, similar to white noise model, we have

$$\tilde{R}_{z,n}(\sigma; f) \geq \sup\{|Z(g) - Z(f)| : l_n(f, g) \leq \sigma/\sqrt{n}\} \Phi(-0.5). \quad (\text{A.1.100})$$

For minimum, similar procedure shows that

$$\tilde{R}_{m,n}(\sigma; f) \geq \sup\{|M(g) - M(f)| : l_n(f, g) \leq \sigma/n\} \Phi(-0.5). \quad (\text{A.1.101})$$

For confidence interval, for  $0 < \alpha < 0.3$ , similar to the white noise model, we have, for  $CI \in \mathcal{I}_{z,\alpha,n}(\{f, g\})$ ,

$$\begin{aligned} \mathbb{E}_f L(CI) &\geq |Z(f) - Z(g)|(1 - 2\alpha - TV(P_{f,n}, P_{g,n})) \\ &\geq |Z(f) - Z(g)|(1 - 2\alpha - \sqrt{\chi^2(P_{f,n}, P_{g,n})}). \end{aligned}$$

where  $P_{f,n}$  is the distribution of the regression model with  $n+1$  observations corresponding to  $f$ .

Further, we have

$$\begin{aligned} \chi^2(P_{f,n}, P_{g,n}) &= \\ &\int \exp\left(\frac{\sum_{i=0}^n (f(x_i) - g(x_i))(2y_i - f(x_i) - g(x_i))}{\sigma^2}\right) p(y_0, \dots, y_n | g) dy_0 dy_1 \dots dy_n - 1 \\ &= \exp\left(\frac{l_n(f, g)^2}{\sigma^2/n}\right) - 1. \end{aligned} \quad (\text{A.1.102})$$

Picking  $g \in \mathcal{F}$  such that  $l_n(f, g) \leq \frac{1}{3} \frac{\sigma}{\sqrt{n}}$ , then we have  $\mathbb{E}_f L(CI) \geq (0.6 - 2\alpha)|Z(f) - Z(g)|$ .

Hence

$$\tilde{L}_{z,\alpha,n}(\sigma; f) \geq (0.6 - 2\alpha) \sup\{|Z(g) - Z(f)| : l_n(f, g) \leq \frac{1}{3} \sigma/\sqrt{n}\}. \quad (\text{A.1.103})$$

Similarly, we have

$$\tilde{L}_{m,\alpha,n}(\sigma; f) \geq (0.6 - 2\alpha) \sup\{|M(g) - M(f)| : l_n(f, g) \leq \frac{1}{3} \sigma/\sqrt{n}\}. \quad (\text{A.1.104})$$

Further, we have the following lemma, which we prove in Section A.2 (page 228). Recall that the function class of convex functions having the same values with  $f$  on  $x_0, x_1, \dots, x_n$  is  $\mathcal{G}_n(f) = \{g \in \mathcal{F} : g(x_i) = f(x_i), i = 0, 1, \dots, n\}$ , defined in Equation (A.1.99). Then we know that  $P_{g,n} = P_{f,n}$  for all  $g \in \mathcal{G}_n(f)$ .

**Lemma A.1.21.** *For any  $h \in \mathcal{G}_n(f)$ . When  $\rho_z(\frac{\sigma}{\sqrt{6n}}; h) \geq 1/2n$ , let*

$$g_{n,\sigma,h}(t) = \max\{h(t), M(h) + \rho_m(\frac{\sigma}{\sqrt{6n}}; h)\}.$$

*Then we have*

$$l_n(f, g_{n,\sigma,h}) \leq \sigma^2/n.$$

*When  $\rho_z(\frac{\sigma}{\sqrt{6n}}; h) < 1/2n$ , let*

$$g_{n,\sigma,h}(t) = \max\{h(t), M(h) + \rho_m(\frac{\sigma}{\sqrt{6n}}; h)\sqrt{2n\rho_z(\frac{\sigma}{\sqrt{6n}}; h)}\},$$

*then we will have*

$$l_n(f, g_{n,\sigma,h}) \leq \sigma^2/n.$$

Let  $t_l(h) = \inf\{g_{n,\sigma,h}(t) > h(t)\}$ ,  $t_r(h) = \sup\{g_{n,\sigma,h}(t) > h(t)\}$ , similar to the white noise model, we know that for any  $\delta > 0$ , exists  $g_{n,\sigma,h,\delta,l}$ ,  $g_{n,\sigma,h,\delta,r} \in \mathcal{F}$ , such that

$$l_n(f, g_{n,\sigma,h,\delta,l}) \leq \sigma^2/n, l_n(f, g_{n,\sigma,h,\delta,r}) \leq \sigma^2/n,$$

$$Z(g_{n,\sigma,h,\delta,l}) \leq t_l + \delta, Z(g_{n,\sigma,h,\delta,r}) \geq t_r - \delta,$$

and

$$M(g_{n,\sigma,h,\delta,r}) = M(g_{n,\sigma,h,\delta,l}) = \min\{\rho_m(\frac{\sigma}{\sqrt{6n}}; h), \rho_m(\frac{\sigma}{\sqrt{6n}}; h)\sqrt{2n\rho_z(\frac{\sigma}{\sqrt{6n}}; h)}\}.$$



Therefore,

$$\begin{aligned}
& \sup\{|Z(g) - Z(f)| : l_n(f, g) \leq \sigma/\sqrt{n}, g \in \mathcal{F}\} \\
& \geq \sup_{h \in \mathcal{G}_n(f)} \frac{1}{2} \lim_{\delta \rightarrow 0^+} (Z(g_{n,\sigma,h,\delta,r}) - Z(g_{n,\sigma,h,\delta,l})) \\
& = \sup_{h \in \mathcal{G}_n(f)} \frac{1}{2} (t_r - t_l) \\
& \geq \frac{1}{2} \sup_{h \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{\sqrt{6n}}; h\right) \{1 \wedge \sqrt{2n\rho_z\left(\frac{\sigma}{\sqrt{6n}}; h\right)}\} \\
& \geq \frac{1}{2} \sup_{h \in \mathcal{G}_n(f)} 54^{-\frac{1}{4}} \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right) \left(1 \wedge \sqrt{n\rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)}\right),
\end{aligned} \tag{A.1.105}$$

$$\begin{aligned}
& \sup\{|Z(g) - Z(f)| : l_n(f, g) \leq \sigma/3\sqrt{n}, g \in \mathcal{F}\} \\
& \geq \sup_{h \in \mathcal{G}_n(f)} \frac{1}{2} \lim_{\delta \rightarrow 0^+} (Z(g_{n,\sigma/3,h,\delta,r}) - Z(g_{n,\sigma/3,h,\delta,l})) \\
& = \sup_{h \in \mathcal{G}_n(f)} \frac{1}{2} (t_r - t_l) \geq \\
& \geq \frac{1}{2} \sup_{h \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{3\sqrt{6n}}; h\right) \{1 \wedge \sqrt{2n\rho_z\left(\frac{\sigma}{3\sqrt{6n}}; h\right)}\} \\
& \geq \frac{1}{2} \sup_{h \in \mathcal{G}_n(f)} \frac{1}{9} 6^{-\frac{1}{4}} \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right) \left(1 \wedge \sqrt{n\rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)}\right),
\end{aligned} \tag{A.1.106}$$

$$\begin{aligned}
& \sup\{|M(g) - M(f)| : l_n(f, g) \leq \sigma/\sqrt{n}, g \in \mathcal{F}\} \\
& \geq \min\{\rho_m\left(\frac{\sigma}{\sqrt{6n}}; h\right), \rho_m\left(\frac{\sigma}{\sqrt{6n}}; h\right) \sqrt{2n\rho_z\left(\frac{\sigma}{\sqrt{6n}}; h\right)}\} \\
& \geq 54^{-\frac{1}{4}} \min\{\rho_m\left(\frac{\sigma}{\sqrt{n}}; h\right), \rho_m\left(\frac{\sigma}{\sqrt{n}}; h\right) \sqrt{n\rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)}\},
\end{aligned} \tag{A.1.107}$$

and

$$\begin{aligned}
& \sup\{|M(g) - M(f)| : l_n(f, g) \leq \sigma/3\sqrt{n}, g \in \mathcal{F}\} \\
& \geq \min\{\rho_m(\frac{\sigma}{3\sqrt{6n}}; h), \rho_m(\frac{\sigma}{3\sqrt{6n}}; h) \sqrt{2n\rho_z(\frac{\sigma}{3\sqrt{6n}}; h)}\} \\
& \geq \frac{1}{9}6^{-\frac{1}{4}} \min\{\rho_m(\frac{\sigma}{\sqrt{n}}; h), \rho_m(\frac{\sigma}{\sqrt{n}}; h) \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; h)}\}.
\end{aligned} \tag{A.1.108}$$

Getting back to Inequalities (A.1.100), (A.1.101), (A.1.103), (A.1.104), we have

$$\tilde{R}_{z,n}(\sigma; f) \geq \frac{1}{2}\Phi(-0.5)54^{-\frac{1}{4}} \sup_{h \in \mathcal{G}_n(f)} \rho_z(\frac{\sigma}{\sqrt{n}}; h) \left(1 \wedge \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; h)}\right), \tag{A.1.109}$$

$$\tilde{R}_{m,n}(\sigma; f) \geq \Phi(-0.5)54^{-\frac{1}{4}} \sup_{h \in \mathcal{G}_n(f)} \rho_z(\frac{\sigma}{\sqrt{n}}; h) \left(1 \wedge \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; h)}\right), \tag{A.1.110}$$

$$\tilde{L}_{z,\alpha,n}(\sigma; f) \geq \frac{1}{2}(0.6 - 2\alpha)\frac{1}{9}6^{-\frac{1}{4}} \sup_{h \in \mathcal{G}_n(f)} \rho_z(\frac{\sigma}{\sqrt{n}}; h) \left(1 \wedge \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; h)}\right), \tag{A.1.111}$$

$$\tilde{L}_{m,\alpha,n}(\sigma; f) \geq (0.6 - 2\alpha)\frac{1}{9}6^{-\frac{1}{4}} \sup_{h \in \mathcal{G}_n(f)} \rho_m(\frac{\sigma}{\sqrt{n}}; h) \left(1 \wedge \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; h)}\right). \tag{A.1.112}$$

Now we turn to the discretization error.

Since for any  $g \in \mathcal{G}_n(f)$ , we have  $\frac{dP_{f,n}}{dP_{g,n}}(y_0, y_1, \dots, y_n) = 1$  for all  $(y_0, y_1, \dots, y_n) \in \mathcal{R}^n$ .

Therefore, for any estimator  $\hat{Z}$ , we have

$$\begin{aligned}
\mathbb{E}_g|\hat{Z} - Z(g)| + \mathbb{E}_f|\hat{Z} - Z(f)| &= \mathbb{E}_f \left( |\hat{Z} - Z(g)| + |\hat{Z} - Z(f)| \right) \\
&\geq \mathbb{E}_f |Z(f) - Z(g)| = |Z(f) - Z(g)|.
\end{aligned}$$

Hence we have

$$\tilde{R}_{z,n}(\sigma; f) \geq \frac{1}{2} \sup_{g \in \mathcal{G}_n(f)} |Z(f) - Z(g)| \geq \frac{1}{4}\mathfrak{D}_z(n, f).$$

Similarly, we have  $\tilde{R}_{m,n}(\sigma; f) \geq \frac{1}{4}\mathfrak{D}_m(n, f)$ . For the confidence interval, we have for any

$g \in \mathcal{G}_n(f)$ , and for any  $CI \in I_{z,\alpha,n}(\{f, g\})$ ,

$$\begin{aligned}\mathbb{E}_f L(CI) &\geq (1 - P_f(Z(f) \notin CI) - P_f(Z(g) \notin CI))_+ |Z(f) - Z(g)| \\ &\geq (1 - 2\alpha) |Z(f) - Z(g)|.\end{aligned}$$

Hence we have

$$\tilde{L}_{z,\alpha,n}(\sigma; f) \geq (1 - 2\alpha) \cdot \frac{1}{2} \mathfrak{D}_z(n, f).$$

Similarly, we have  $\tilde{L}_{m,\alpha,n}(\sigma; f) \geq (1 - 2\alpha) \cdot \frac{1}{2} \mathfrak{D}_m(n, f)$ .

□

#### A.1.10. Proof of Theorem 2.4.1

With Proposition A.1.4, to prove the theorem, we only need to prove the following two propositions:

**Proposition A.1.5.** *For  $\hat{Z}$  defined in (2.4.5), we have*

$$\mathbb{E}(|\hat{Z} - Z(f)|) \leq \check{C}_1 \rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right) + \frac{2}{n}. \quad (\text{A.1.113})$$

**Proposition A.1.6.** *For  $\hat{Z}$  defined in (2.4.5), if  $\sup_{h \in \mathcal{G}_n(f)} \{\rho_z(\frac{\sigma}{\sqrt{n}}; h)\} < \frac{1}{2n}$ , we have*

$$\mathbb{E}(|\hat{Z} - Z(f)|) \leq \check{C}_2 \sup_{h \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)} + \mathfrak{D}_z(n, f). \quad (\text{A.1.114})$$

Let  $C_1 = \frac{\sqrt{2}\check{C}_1 + 4 + \check{C}_2}{\check{C}_z} + 4$ , where  $\check{C}_z$  is defined in (A.1.98), gives the statement of the theorem.

We first give the main part of the proof of the two propositions and then give the proofs of the lemmas in there.

*Proof of Proposition A.1.5.*

$$\begin{aligned}\mathbb{E}(|\hat{Z} - Z(f)|) &= \mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} < \tilde{\mathbf{j}}\}|\hat{Z} - Z(f)|) + \mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} \geq \tilde{\mathbf{j}}\}|\hat{Z} - Z(f)|) \\ &\leq \mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} < \tilde{\mathbf{j}}\}1.5m_{\mathbf{j}}) + \mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} \geq \tilde{\mathbf{j}}\}|\hat{Z} - Z(f)|)\end{aligned}\tag{A.1.115}$$

To bound the two terms, we give two lemmas below, the proofs of the lemmas are in Section A.2 (page 228, 230).

**Lemma A.1.22.**

$$\mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} < \tilde{\mathbf{j}}\}1.5m_{\mathbf{j}}) \leq c_{z1}\rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right) + \frac{1.5}{n}\mathbb{1}\{J \leq \mathbf{j}^* - 3\}.\tag{A.1.116}$$

**Lemma A.1.23.**

$$\mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} \geq \tilde{\mathbf{j}}\}|\hat{Z} - Z(f)|) \leq c_{z2}\rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right).\tag{A.1.117}$$

Therefore,

$$\mathbb{E}(|\hat{Z} - Z(f)|) \leq (c_{z1} + c_{z2})\rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right) + \frac{1.5}{n}\mathbb{1}\{J \leq \mathbf{j}^* - 3\} \leq \check{C}_1\rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right) + \frac{1.5}{n}.\tag{A.1.118}$$

□

*Proof of Proposition A.1.6.* since  $\sup_{h \in \mathcal{G}_n(f)} \{\rho_z(\frac{\sigma}{\sqrt{n}}; h)\} < \frac{1}{2n}$ , we know that  $|\{i : f(x_i) = \min\{f(x_k) : 0 \leq k \leq n\}\}| = 1$ . Suppose  $i_{min} \in \{i : f(x_i) = \min\{f(x_k) : 0 \leq k \leq n\}\}$ . Let  $\tilde{h}$  be the piece wise linear function such that  $\tilde{h}(x_i) = f(x_i)$  for all  $0 \leq i \leq n$ , and  $\tilde{h}$  is linear on all the sub-intervals  $[k/n, (k+1)/n]$ , for  $0 \leq k \leq n-1$ . It is clear that  $Z(\tilde{h}) = x_{i_{min}}$ .

Then we have

$$\begin{aligned}
& \mathbb{E}(|\hat{Z} - Z(f)|) \\
& \leq \mathbb{E}(|\hat{Z} - Z(\tilde{h})|) + |Z(\tilde{h}) - Z(f)| \\
& \leq \mathbb{E}(|\hat{Z} - Z(\tilde{h})|) + \mathfrak{D}_z(n, f) \\
& = \mathbb{E}(\mathbb{1}\{\check{j} < \infty\}|\hat{Z} - Z(\tilde{h})|) + \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}|\hat{Z} - Z(\tilde{h})|) + \mathfrak{D}_z(n, f).
\end{aligned} \tag{A.1.119}$$

Also, we can further split the first and second terms by the  $\{\hat{j} < \tilde{j}\}$  and  $\{\hat{j} \geq \tilde{j}\}$  to have

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\check{j} < \infty\}|\hat{Z} - Z(\tilde{h})|) \\
& = \mathbb{E}(\mathbb{1}\{\check{j} < \infty\}\mathbb{1}\{\check{j} < \tilde{j}\}|\hat{Z} - Z(\tilde{h})|) + \mathbb{E}(\mathbb{1}\{\check{j} < \infty\}\mathbb{1}\{\check{j} \geq \tilde{j}\}|\hat{Z} - Z(\tilde{h})|),
\end{aligned} \tag{A.1.120}$$

and

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}|\hat{Z} - Z(\tilde{h})|) \\
& = \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}\mathbb{1}\{\hat{j} < \tilde{j}\}|\hat{Z} - Z(\tilde{h})|) + \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}\mathbb{1}\{\hat{j} \geq \tilde{j}\}|\hat{Z} - Z(\tilde{h})|) \\
& = \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}\mathbb{1}\{\hat{j} \geq \tilde{j}\}|\hat{Z} - Z(\tilde{h})|).
\end{aligned} \tag{A.1.121}$$

Therefore,

$$\begin{aligned}
& \mathbb{E}(|\hat{Z} - Z(f)|) \\
& \leq \mathbb{E}(\mathbb{1}\{\check{j} < \infty\}\mathbb{1}\{\check{j} < \tilde{j}\}|\hat{Z} - Z(\tilde{h})|) + \mathbb{E}(\mathbb{1}\{\check{j} < \infty\}\mathbb{1}\{\check{j} \geq \tilde{j}\}|\hat{Z} - Z(\tilde{h})|) \\
& \quad + \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}\mathbb{1}\{\hat{j} \geq \tilde{j}\}|\hat{Z} - Z(\tilde{h})|) + \mathfrak{D}_z(n, f) \\
& = \mathbb{E}(\mathbb{1}\{\check{j} < \infty\}\mathbb{1}\{\check{j} < \tilde{j}\}|\hat{Z} - Z(\tilde{h})|) + \mathbb{E}(\mathbb{1}\{\check{j} \geq \tilde{j}\}|\hat{Z} - Z(\tilde{h})|) + \mathfrak{D}_z(n, f)
\end{aligned} \tag{A.1.122}$$

Finally, with the help of the following lemmas (proved in Section A.2, page 230, 231), we prove the proposition.

**Lemma A.1.24.**

$$\mathbb{E}(\mathbb{1}\{\check{j} < \infty\} \mathbb{1}\{\check{j} < \tilde{j}\} |\hat{Z} - Z(\tilde{h})|) \leq \check{c}_{z1} \rho_z\left(\frac{\sigma}{\sqrt{n}}; \tilde{h}\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; \tilde{h}\right)} \quad (\text{A.1.123})$$

**Lemma A.1.25.**

$$\mathbb{E}(\mathbb{1}\{\hat{j} \geq \tilde{j}\} |\hat{Z} - Z(\tilde{h})|) \leq \check{c}_{z2} \rho_z\left(\frac{\sigma}{\sqrt{n}}; \tilde{h}\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; \tilde{h}\right)} \quad (\text{A.1.124})$$

□

#### A.1.11. Proof of Theorem 2.4.2

With Proposition A.1.4, we prove the theorem by proving the following three lemmas:

**Lemma A.1.26** (length of the confidence interval for minimizer).

$$\mathbb{E}_f L(\text{CI}_{z,\alpha}(Y)) < \tilde{C}_{2,\alpha} (C_0 \rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right) + \frac{1}{n}).$$

**Lemma A.1.27.** *When  $\sup_{h \in \mathcal{G}_n(f)} \{\rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)\} < \frac{1}{2n}$ , we have*

$$\mathbb{E}_f L(\text{CI}_{z,\alpha}(Y)) < \tilde{C}_{2,\alpha} \sup_{h \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)} + 2\mathfrak{D}_z(n, f)$$

**Lemma A.1.28** (coverage of the confidence interval for minimizer).

$$P(Z(f) \in \text{CI}_{z,\alpha}(Y)) \geq 1 - \alpha.$$

Let  $C_{2,\alpha} = \max\left\{\frac{\tilde{C}_{2,\alpha}(C_0+2)\sqrt{2}}{\tilde{C}_{z,\alpha}}, \frac{\tilde{C}_{2,\alpha}}{\tilde{C}_{z,\alpha}} + \frac{4}{1-2\alpha}\right\}$ , then we have the statement of the theorem.

*Proof of Lemma A.1.26.*

$$\begin{aligned}
& \mathbb{E}_f L(\mathbf{CI}_{z,\alpha}(Y)) \\
&= \mathbb{E}((U - L) \frac{2^{J-\tilde{J}}}{n}) \leq (24 \times 2^{K_{\frac{\alpha}{2}}} - 3) \cdot \mathbb{E}(\frac{2^{J-\tilde{J}}}{n}) \\
&= (24 \times 2^{K_{\frac{\alpha}{2}}} - 3) \frac{2^J}{n} \mathbb{E}(\sum_{j=1}^{j^*-1} 2^{-j} \mathbb{1}\{\tilde{J} = j\} + \sum_{j=j^*}^{\infty} 2^{-j} \mathbb{1}\{\tilde{J} = j\}) \\
&\leq (24 \times 2^{K_{\frac{\alpha}{2}}} - 3) \frac{2^J}{n} (\sum_{j=1}^{j^*-1} 2^{-j} \mathbb{E}(\mathbb{1}\{\tilde{J} = j, \tilde{J} > j\} + \mathbb{1}\{\tilde{J} = j, \tilde{J} \leq j\}) + 2^{-j^*})
\end{aligned} \tag{A.1.125}$$

To bound the first two terms, we will introduce two lemmas. The proofs of the lemmas are given at Section A.2 (page 231 and 232).

**Lemma A.1.29.**

$$\sum_{j=1}^{j^*-1} \mathbb{E}(2^{-j} \mathbb{1}\{\tilde{J} = j, \tilde{J} > j\}) \leq 2^{-j^*} c_{z3} + 2^{-J} \mathbb{1}\{J \leq j^* - 1\}. \tag{A.1.126}$$

**Lemma A.1.30.**

$$\sum_{j=1}^{j^*-1} \mathbb{E}(2^{-j} \mathbb{1}\{\tilde{J} = j, \tilde{J} \leq j\}) \leq 2^{-j^*} c_{z4}. \tag{A.1.127}$$

With these lemmas, we have

$$\begin{aligned}
& \mathbb{E}_f L(\mathbf{CI}_{z,\alpha}(Y)) \\
&\leq (24 \times 2^{K_{\frac{\alpha}{2}}} - 3) \left( \frac{2^{J-j^*}}{n} (c_{z4} + c_{z3} + 1) + \frac{1}{n} \mathbb{1}\{J \leq j^* - 1\} \right) \\
&\leq \tilde{C}_{2,\alpha} (C_0 \rho_z(\frac{\sigma}{\sqrt{n}}; f) + \frac{1}{n} \mathbb{1}\{J \leq j^* - 1\}),
\end{aligned} \tag{A.1.128}$$

where  $\tilde{C}_{2,\alpha} = (24 \times 2^{K_{\alpha}} - 3)$ ,  $C_0 = \frac{c_{z3} + c_{z4} + 1}{4}$ . □

*Proof of Lemma A.1.27.* To prove the lemma, we introduce the following lemmas while postponing their proofs.

**Lemma A.1.31.** When  $\sup_{h \in \mathcal{G}_n(f)} \{\rho_z(\frac{\sigma}{\sqrt{n}}; h)\} < \frac{1}{2n}$ ,

$$\mathbb{E}(\mathbb{1}\{\check{j} < \infty\} L(\mathbf{CI}_{z,\alpha}(Y))) \leq \check{c}_{1,\alpha} \sup_{h \in \mathcal{G}_n(f)} \rho_z(\frac{\sigma}{\sqrt{n}}; h) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; h)}. \quad (\text{A.1.129})$$

**Lemma A.1.32.** When  $\sup_{h \in \mathcal{G}_n(f)} \{\rho_z(\frac{\sigma}{\sqrt{n}}; h)\} < \frac{1}{2n}$ ,

$$\mathbb{E}(\mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{t_{hi} - t_{lo} \geq \frac{3}{n}\} L(\mathbf{CI}_{z,\alpha}(Y))) \leq \check{c}_{2,\alpha} \sup_{h \in \mathcal{G}_n(f)} \rho_z(\frac{\sigma}{\sqrt{n}}; h) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; h)}. \quad (\text{A.1.130})$$

**Lemma A.1.33.** When  $\sup_{h \in \mathcal{G}_n(f)} \{\rho_z(\frac{\sigma}{\sqrt{n}}; h)\} < \frac{1}{2n}$ ,

$$\begin{aligned} & \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{t_{hi} - t_{lo} < \frac{3}{n}\} L(\mathbf{CI}_{z,\alpha}(Y))) \\ & \leq \check{c}_{3,\alpha} \sup_{h \in \mathcal{G}_n(f)} \rho_z(\frac{\sigma}{\sqrt{n}}; h) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; h)} + 2\mathfrak{D}_z(n, f). \end{aligned} \quad (\text{A.1.131})$$

With these lemmas, we have the statement of Lemma A.1.27.

For the proofs of the lemmas, the main parts are in Section A.2 (page 233, 234 and 236), but here we mention the common thing that will be used in all of them.

When  $\sup_{h \in \mathcal{G}_n(f)} \{\rho_z(\frac{\sigma}{\sqrt{n}}; h)\} < \frac{1}{2n}$ , we know that  $|\{k : f(x_k) = \min\{f(x_i) : 0 \leq i \leq n\}\}| = 1$ , we denote this unique element to be  $i_m$ .

Let  $\tilde{h}$  be the piece wise linear function such that  $\tilde{h}(x_i) = f(x_i)$  for all  $0 \leq i \leq n$ , and  $\tilde{h}$  is linear on all the sub-intervals  $[k/n, k+1/n]$ , for  $0 \leq k \leq n-1$ . Then we know that  $\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) < \frac{1}{2n}$

Suppose  $Y_{e,1} = \{y_{e,i} + \sqrt{3}\sigma z_{3,i} : (L-1) \vee 0 \leq i \leq (U+1) \wedge n\}$ ,  $Y_{e,2} = \{y_{e,i} - \sqrt{3}\sigma z_{3,i} : (L-1) \vee 0 \leq i \leq (U+1) \wedge n\}$ . Then we know that  $\mathbf{Y}_l, \mathbf{Y}_s, Y_{e,1}, Y_{e,2}$  are independent.

□

*Proof of Lemma A.1.28.* In this proof, to make the main idea more clear, the proofs of the



lemmas used are postponed to Section A.2.

With a bit abuse of notation, define the following events:

$$\begin{aligned}
E &= \left\{ Z(f) \in \left[ (\hat{\mathbf{i}}_{\hat{\mathbf{j}}} - (6 \cdot 2^{K_{\alpha/2}+1} - 2) - 1) \frac{2^{J-\hat{\mathbf{j}}}}{n} - \frac{1}{2n}, \right. \right. \\
&\quad \left. \left. (\hat{\mathbf{i}}_{\hat{\mathbf{j}}} + (6 \cdot 2^{K_{\alpha/2}+1} - 2)) \frac{2^{J-\hat{\mathbf{j}}}}{n} - \frac{1}{2n} \right] \cap [0, 1] \right\} \\
F_1 &= \left\{ i_l \leq \min \left\{ i : f(x_i) = \min \{ f(x_k) : 0 \leq k \leq n \} \right\} \right\} \\
F_2 &= \left\{ i_r + 1 \geq \max \left\{ i : f(x_i) = \min \{ f(x_k) : 0 \leq k \leq n \} \right\} \right\}.
\end{aligned} \tag{A.1.132}$$

For  $\mathbf{j}^{\mathbf{w}}$  defined in (A.1.5), we have the following lemma (proved in Section A.2, on page 246).

**Lemma A.1.34.** *For  $K \geq 1$ ,*

$$\Phi(\hat{\mathbf{j}} \geq \mathbf{j}^{\mathbf{w}} + K + 1) \leq \Phi(-2)^K. \tag{A.1.133}$$

Therefore, with this lemma, we have

$$P(E^c) \leq P(|\mathbf{i}_{\hat{\mathbf{j}}}^* - \hat{\mathbf{i}}_{\hat{\mathbf{j}}}| > 6 \cdot 2^{K_{\frac{\alpha}{2}}+1} - 2) \leq P(\mathbb{1}\{\hat{\mathbf{j}} > \mathbf{j}^{\mathbf{w}} + K_{\alpha/2}\}) \leq \frac{\alpha}{2}. \tag{A.1.134}$$

Therefore,

$$\begin{aligned}
& P(Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)) \\
&= \mathbb{E}(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\}) + \mathbb{E}(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E^c\}) \\
&\leq \mathbb{E}(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} < \infty\}) + \\
&\quad \mathbb{E}(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\}) + \frac{\alpha}{2} \\
&= \mathbb{E}(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\}) + \frac{\alpha}{2} \\
&\leq \mathbb{E}\left(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} (\mathbb{1}\{F_1 \cap F_2\} + \mathbb{1}\{F_1^c\} + \mathbb{1}\{F_2^c\})\right) + \frac{\alpha}{2} \\
&\leq \mathbb{E}\left(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\}\right) \\
&\quad + \mathbb{E}(\mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} (\mathbb{1}\{F_1^c\} + \mathbb{1}\{F_2^c\})) + \frac{\alpha}{2}.
\end{aligned} \tag{A.1.135}$$

We introduce the following lemma, which is proved in Section A.2 on page 246.

**Lemma A.1.35.**

$$\mathbb{E}(\mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1^c\}) \leq \alpha_1, \mathbb{E}(\mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_2^c\}) \leq \alpha_1. \tag{A.1.136}$$

Therefore

$$P(Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)) \leq \mathbb{E}\left(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\}\right) + \frac{\alpha}{2} + 2\alpha_1. \tag{A.1.137}$$

Now we turn to the only term left

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \right) \\
&= \mathbb{E} \left( \mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) = 0\} \right) \\
&\quad + \mathbb{E} \left( \mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
&\quad \left. \mathbb{1}\{i_{hi} - i_{lo} \geq 3 \text{ or } (i_{hi} - n)i_{lo} = 0\} \right) \\
&\quad + \mathbb{E} \left( \mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
&\quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \right) \\
&= \mathbb{E} \left( \mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) = 0\} \right) \\
&\quad + \mathbb{E} \left( \mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
&\quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \right). \tag{A.1.138}
\end{aligned}$$

The second equation is because when under the event  $E \cap F_1 \cap F_2 \cap \{\check{j} = \infty\} \cap \{(i_l - U)(i_r - L + 1) \neq 0\} \cap \{i_{hi} - i_{lo} \geq 3 \text{ or } (i_{hi} - n)i_{lo} = 0\}$ ,  $Z(f) \in \mathbf{CI}_{z,\alpha}(Y)$ .

We have the following lemmas, which are proved in Section A.2 on page 247 and 248.

**Lemma A.1.36.**

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) = 0\} \right) \\
&\leq 3\alpha_2 \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) = 0\} \right). \tag{A.1.139}
\end{aligned}$$

**Lemma A.1.37.**

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
&\quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \right) \\
&\leq 6\alpha_2 P \left( \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
&\quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \right). \tag{A.1.140}
\end{aligned}$$

With these two lemmas, we finally have

$$P(Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)) \leq 6\alpha_2 + \frac{\alpha}{2} + 2\alpha_1 \leq \alpha \quad (\text{A.1.141})$$

□

### A.1.12. Proof of Theorem 2.4.3

With Proposition A.1.4, to prove this theorem, it's sufficient to prove the following two propositions:

#### Proposition A.1.7.

$$\mathbb{E}(|\hat{M} - M(f)|) \leq \check{C}_{3,0}\rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right) + \sqrt{2}(\min\{f(x_i) : 0 \leq i \leq n\} - M(f)). \quad (\text{A.1.142})$$

**Proposition A.1.8.** *When  $\sup_{h \in \mathcal{G}_n(f)} \{\rho_z(\frac{\sigma}{\sqrt{n}}; h)\} < \frac{1}{2n}$ , we have*

$$\begin{aligned} \mathbb{E}(|\hat{M} - M(f)|) &\leq \check{C}_3 \sup_{h \in \mathcal{G}_n(f)} \rho_m\left(\frac{\sigma}{\sqrt{n}}; h\right) \sqrt{n\rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)} \\ &\quad + \sqrt{2}(\min\{f(x_i) : 0 \leq i \leq n\} - M(f)). \end{aligned} \quad (\text{A.1.143})$$

Let  $C_3 = \frac{\sqrt{2}\check{C}_{3,0} + \check{C}_3}{\check{C}_m} + 4\sqrt{2}$  gives the statement of Theorem 2.4.3.

*Proof of Proposition A.1.7.* Then we will have

$$\mathbb{E}((\hat{M} - M(f))^2) = \mathbb{E}((\hat{M} - M(f))^2 \mathbb{1}_{\{\check{j} < \infty\}}) + \mathbb{E}((\hat{M} - M(f))^2 \mathbb{1}_{\{\check{j} = \infty\}}). \quad (\text{A.1.144})$$

For the first term we have

$$\begin{aligned}
\mathbb{E}((\hat{M} - M(f))^2 \mathbb{1}\{\check{j} < \infty\}) &= \mathbb{E}\left(\left((\hat{\mathbf{f}} - M(f)) + \mathfrak{E}_{\check{j}, \hat{\mathbf{i}}_{\check{j}}, e} \frac{1}{2^{J-\check{j}}}\right)^2 \mathbb{1}\{\check{j} < \infty\}\right) \\
&= \mathbb{E}\left((\hat{\mathbf{f}} - M(f))^2 + 2(\hat{\mathbf{f}} - M(f))\mathfrak{E}_{\check{j}, \hat{\mathbf{i}}_{\check{j}}, e} \frac{1}{2^{J-\check{j}}} + (\mathfrak{E}_{\check{j}, \hat{\mathbf{i}}_{\check{j}}, e} \frac{1}{2^{J-\check{j}}})^2 \mathbb{1}\{\check{j} < \infty\}\right) \quad (\text{A.1.145}) \\
&= \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\check{j} < \infty\}) + \mathbb{E}\left((\mathfrak{E}_{\check{j}, \hat{\mathbf{i}}_{\check{j}}, e} \frac{1}{2^{J-\check{j}}})^2 \mathbb{1}\{\check{j} < \infty\}\right).
\end{aligned}$$

We introduce following two lemmas (proved in Section A.2 on page 250 and 251) to bound the two terms.

**Lemma A.1.38.**

$$\mathbb{E}((\mathfrak{E}_{\check{j}, \hat{\mathbf{i}}_{\check{j}}, e} \frac{1}{2^{J-\check{j}}})^2 \mathbb{1}\{\check{j} < \infty\}) \leq c_{m1} \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2. \quad (\text{A.1.146})$$

**Lemma A.1.39.**

$$\mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\check{j} < \infty\}) \leq c_{m2} \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2. \quad (\text{A.1.147})$$

For the second term in Equation (A.1.144), let

$$\begin{aligned}
\mathbf{i} &= \arg \min_{\hat{\mathbf{i}}_J - 2 \leq i \leq \hat{\mathbf{i}}_J + 2} f(x_{i-1}), \\
f_i &= f(x_{i-1}), \\
\delta_i &= y_{e, i-1} - f(x_{i-1}), \\
\eta &= \min\{\delta_i : \hat{\mathbf{i}}_J - 2 \leq i \leq \hat{\mathbf{i}}_J + 2\},
\end{aligned} \quad (\text{A.1.148})$$

then we know  $\mathbb{E}(\eta|\hat{\mathbf{i}}_J) \leq 0$ , and we have

$$\begin{aligned}
& \mathbb{E}((\hat{M} - M(f))^2 \mathbb{1}\{\check{j} = \infty\}) \\
& \leq \mathbb{E}((f_{\mathbf{i}} - M(f) + \delta_{\mathbf{i}})^2 \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{\hat{M} > M(f)\}) \\
& \quad + \mathbb{E}((f_{\mathbf{i}} - M(f) + \eta)^2 \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{\hat{M} < M(f)\}) \\
& \leq 2\mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\check{j} = \infty\}) + 2\gamma_e^2 \sigma^2 \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}) \\
& \quad + \mathbb{E}(\mathbb{E}(\eta^2 \mathbb{1}\{\eta < 0\} | \mathbf{Y}_l, \mathbf{Y}_s) \mathbb{1}\{\check{j} = \infty\}) \\
& \leq 2\mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\check{j} = \infty\}) + 2\gamma_e^2 \sigma^2 \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}) + \sigma^2 \gamma_e^2 Q_2 \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}),
\end{aligned} \tag{A.1.149}$$

where  $Q_2 = \int_0^\infty x^2 5\Phi(x)^4 \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx \leq \frac{5}{2}$ .

To bound it we have the following lemmas, which are proved in Section A.2 on page 263 and 264.

**Lemma A.1.40.**

$$\begin{aligned}
& \mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\check{j} = \infty\}) \\
& \leq c_{m6} \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2 + (\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2
\end{aligned} \tag{A.1.150}$$

**Lemma A.1.41.**

$$\sigma^2 \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}) \leq 32 \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2 \tag{A.1.151}$$

Combining them together, we have

$$\begin{aligned}
& \mathbb{E}((\hat{M} - M(f))^2) \\
& \leq (c_{m1} + c_{m2} + 144\gamma_e^2 + 2m_6) \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2 + 2(\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2 \\
& \leq C_{3,0} \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2 + 2(\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2.
\end{aligned} \tag{A.1.152}$$

Therefore,

$$\begin{aligned}\mathbb{E}(|\hat{M} - M(f)|) &\leq \sqrt{\mathbb{E}((\hat{M} - M(f))^2)} \\ &\leq \check{C}_{3,0}\rho_m(\frac{\sigma}{\sqrt{n}}; f) + \sqrt{2}(\min\{f(x_i) : 0 \leq i \leq n\} - M(f)).\end{aligned}\tag{A.1.153}$$

□

*Proof of Proposition A.1.8.* Since we have

$$\sup_{h \in \mathcal{G}_n(f)} \rho_m(\frac{\sigma}{\sqrt{n}}; h) \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; h)} \geq \sqrt{n} \frac{1}{\sqrt{2}} \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{2}},\tag{A.1.154}$$

we only need to prove that

$$\mathbb{E}(|\hat{M} - M(f)|) \leq \check{c}_{m1}\sigma + \sqrt{2}(\min\{f(x_i) : 0 \leq i \leq n\} - M(f)).\tag{A.1.155}$$

We recycle all the notation in the proof of Proposition A.1.7, especially in Equation (A.1.148) and (A.2.105).

Similar to the proof of Proposition of A.1.7, we have

$$\begin{aligned}\mathbb{E}((\hat{M} - M(f))^2) &= \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\check{j} < \infty\}) + \mathbb{E}((\mathfrak{E}_{\check{j}, \check{\mathbf{i}}_{\check{j}}, e} \frac{1}{2^{J-\check{j}}})^2 \mathbb{1}\{\check{j} < \infty\}) + \\ &\quad 2\mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\check{j} = \infty\}) + 2\gamma_e^2 \sigma^2 \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}) + \sigma^2 \gamma_e^2 Q_2 \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}),\end{aligned}\tag{A.1.156}$$

where  $Q_2 = \int_0^\infty x^2 \cdot 5\Phi(x)^4 \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx \leq \frac{5}{2}$ .

Since we have

$$\mathbb{E}((\mathfrak{E}_{\check{j}, \check{\mathbf{i}}_{\check{j}}, e} \frac{1}{2^{J-\check{j}}})^2 \mathbb{1}\{\check{j} < \infty\}) = \mathbb{E}(\frac{\sigma^2}{2^{J-\check{j}}} \mathbb{1}\{\check{j} < \infty\}) \leq \sigma^2,\tag{A.1.157}$$

we are only left with bounding the terms:  $\mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\check{j} < \infty\})$ ,  $\mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\check{j} = \infty\})$ .

We have the following two lemmas, which are proved in Section A.2 on page 264 and 269.

**Lemma A.1.42.**

$$\mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\check{j} < \infty\}) \leq \check{c}_{m2}^2 \sigma^2. \quad (\text{A.1.158})$$

**Lemma A.1.43.**

$$\mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\check{j} = \infty\}) \leq \check{c}_{m3}^2 \sigma^2 + (\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2. \quad (\text{A.1.159})$$

With these lemmas, we know that

$$\begin{aligned} & \mathbb{E}((\hat{M} - M(f))^2) \\ & \leq (\check{c}_{m2}^2 + 1 + 2\check{c}_{m3}^2 + 2\gamma_e^2 + \gamma_e^2 Q_2) \sigma^2 + 2(\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2. \end{aligned} \quad (\text{A.1.160})$$

Therefore, we have

$$\begin{aligned} & \mathbb{E}(|\hat{M} - M(f)|) \\ & \leq \sqrt{\check{c}_{m2}^2 + 1 + 2\check{c}_{m3}^2 + 2\gamma_e^2 + \gamma_e^2 Q_2} \sigma + \sqrt{2}(\min\{f(x_i) : 0 \leq i \leq n\} - M(f)) \\ & = \check{C}_3 \frac{\sigma}{\sqrt{2}} + \sqrt{2}(\min\{f(x_i) : 0 \leq i \leq n\} - M(f)). \end{aligned} \quad (\text{A.1.161})$$

□

#### A.1.13. Proof of Theorem 2.4.4

With Proposition A.1.4, we prove the theorem by proving the following lemmas.



**Lemma A.1.44** (length of the confidence interval for minimum 0).

$$\mathbb{E}_f L(\text{CI}_{m,\alpha}(Y)) \leq \check{C}_{4,\alpha} \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right) + \sqrt{2} \left(\min\{f(x_i) : i = 0, 1, \dots, n\} - \check{h}\right),$$

where  $\check{h} = \inf\{M(g) : g \in \mathcal{F}, \text{ and } g(x_i) = f(x_i), i = 0, 1, \dots, n\}$ .

**Lemma A.1.45** (length of the confidence interval for minimum 1). *When*

$$\sup_{h \in \mathcal{G}_n(f)} \{\rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)\} < \frac{1}{2n},$$

*we have*

$$\mathbb{E}_f L(\text{CI}_{m,\alpha}(Y)) \leq \check{C}_{5,\alpha} \sigma + \sqrt{2} \left(\min\{f(x_i) : i = 0, 1, \dots, n\} - \check{h}\right),$$

where  $\check{h} = \min\{M(g) : g \in \mathcal{F}, \text{ and } g(x_i) = f(x_i), i = 0, 1, \dots, n\}$ .

Note that when  $\sup_{h \in \mathcal{G}_n(f)} \{\rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)\} < \frac{1}{2n}$ ,

$$\sup_{h \in \mathcal{G}_n(f)} \rho_m\left(\frac{\sigma}{\sqrt{n}}; h\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)} \geq \frac{\sigma}{\sqrt{2}},$$

hence with these two lemmas we know that

$$\begin{aligned} & \mathbb{E}_f L(\text{CI}_{m,\alpha}(Y)) \\ & \leq (\sqrt{2}\check{C}_{4,\alpha} + \sqrt{2}\check{C}_{5,\alpha}) \sup_{h \in \mathcal{G}_n(f)} \rho_m\left(\frac{\sigma}{\sqrt{n}}; h\right) \left(1 \wedge \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)}\right) + \sqrt{2}\mathfrak{D}_m(n, f). \end{aligned} \quad (\text{A.1.162})$$

When  $0 < \alpha < 0.3$ , let

$$C_{4,\alpha} = \frac{\sqrt{2}\check{C}_{4,\alpha} + \sqrt{2}\check{C}_{5,\alpha}}{\check{C}_{m,\alpha}} + \frac{2\sqrt{2}}{1 - 2\alpha} \quad (\text{A.1.163})$$

gives the statement with respect to the length.

**Lemma A.1.46** (coverage of the confidence interval for minimum).

$$P(M(f) \in \text{CI}_{m,\alpha}(Y)) \geq 1 - \alpha.$$

*Proof of Lemma A.1.44.*

$$I_{hi} - I_{lo} + 1 \leq 2 + 9 \cdot 2^{j_l - j_s} \leq 2 + 9 \cdot 2^{K_{\frac{\alpha}{4}} + \tilde{K}_{\frac{\alpha}{4}} + 1}. \quad (\text{A.1.164})$$

Therefore,

$$S_{I_{hi} - I_{lo} + 1, \frac{\alpha}{4}} \leq -\Phi^{-1}\left(\frac{\alpha}{4(2 + 9 \cdot 2^{K_{\frac{\alpha}{4}} + \tilde{K}_{\frac{\alpha}{4}} + 1})}\right). \quad (\text{A.1.165})$$

$$\begin{aligned} & \mathbb{E}_f L(\text{CI}_{m,\alpha}(Y)) \\ & \leq (S_{I_{hi} - I_{lo} + 1, \frac{\alpha}{4}} - \Phi^{-1}\left(\frac{\alpha}{4}\right) + \sqrt{3}) \gamma_e \mathbb{E}\left(\frac{\sigma}{\sqrt{2^{J-j_l}}}\right) \\ & \quad + \mathbb{E}\left((\hat{\mathbf{f}}_1 - z_{\alpha/4} \frac{\sqrt{3}\sigma}{\sqrt{2^{J-j_l}}} - \frac{\sqrt{3}\sigma}{\sqrt{2^{J-j_l}}} - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\}\right). \end{aligned} \quad (\text{A.1.166})$$

We first bound the first term.

$$\begin{aligned} & \mathbb{E}\left(\frac{\sigma}{\sqrt{2^{J-j_l}}}\right) \\ & \leq \mathbb{E}\left(\frac{\sigma}{\sqrt{2^{J-\hat{\mathbf{j}}-\tilde{K}_{\frac{\alpha}{4}}}}}\mathbb{1}\{\hat{\mathbf{j}} \leq J - \tilde{K}_{\frac{\alpha}{4}}\}\right) + \sigma \mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} > J - \tilde{K}_{\frac{\alpha}{4}}\}) \\ & \leq \mathbb{1}\{\mathbf{j}^* + 3 \leq J\} \left(\mathbb{E}\left(\frac{\sigma}{\sqrt{2^{J-\hat{\mathbf{j}}-\tilde{K}_{\frac{\alpha}{4}}}}}\mathbb{1}\{\hat{\mathbf{j}} \leq J - \tilde{K}_{\frac{\alpha}{4}}\}\right) + \sigma \mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} > J - \tilde{K}_{\frac{\alpha}{4}}\})\right) \\ & \quad + \mathbb{1}\{\mathbf{j}^* + 2 \geq J\} \sigma \\ & \leq \mathbb{1}\{\mathbf{j}^* + 3 \leq J\} \left(\frac{\sigma}{\sqrt{2^{J-\mathbf{j}^*-\tilde{K}_{\frac{\alpha}{4}}-3}}} + \sum_{j=\mathbf{j}^*+3}^J \frac{\sigma}{\sqrt{2^{J-j-1-\tilde{K}_{\frac{\alpha}{4}}}}} \Phi\left(-2 + \frac{1}{6}\right)^{j-\mathbf{j}^*-2} \right. \\ & \quad \left. + \frac{\sigma}{\sqrt{2^{J-\mathbf{j}^*}}} \sqrt{2^{J-\mathbf{j}^*}} \Phi\left(-2 + \frac{1}{6}\right)^{(J-1-\tilde{K}_{\frac{\alpha}{4}}-\mathbf{j}^*)_+}\right) \\ & \quad + \mathbb{1}\left\{\frac{1}{n} > \frac{\rho_z(\frac{\sigma}{\sqrt{n}}; f)}{32}\right\} \sqrt{n} \sqrt{2\rho_z(\frac{\sigma}{\sqrt{n}}; f)} \rho_m(\frac{\sigma}{\sqrt{n}}; f) \\ & = 2^{1+\frac{\tilde{K}_{\frac{\alpha}{4}}}{2}} \tilde{C}_4 \rho_m(\frac{\sigma}{\sqrt{n}}; f) + 8\rho_m(\frac{\sigma}{\sqrt{n}}; f) \bar{C}_{1,\alpha}. \end{aligned} \quad (\text{A.1.167})$$

Now we turn to the second term,

$$\begin{aligned}
& \mathbb{E} \left( (\hat{\mathbf{f}}_1 - z_{\alpha/4} \frac{\sqrt{3}\sigma}{\sqrt{2^{J-j_i}}} - \frac{\sqrt{3}\sigma}{\sqrt{2^{J-j_i}}} - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\
& \leq \mathbb{E} \left( (\hat{\mathbf{f}}_1 - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\
& \leq \mathbb{E} \left( (\hat{\mathbf{f}}_1 - M(f))_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) + \mathbb{E} \left( (M(f) - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\
& \leq \mathbb{E} \left( (\hat{M} - M(f))_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) + \mathbb{E} \left( (M(f) - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right),
\end{aligned} \tag{A.1.168}$$

where  $\hat{M}$  is defined in (2.4.9).

Then according to Proposition A.1.7, we have

$$\mathbb{E} \left( (\hat{M} - M(f))_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \leq \check{C}_{3,0} \rho_m \left( \frac{\sigma}{\sqrt{n}}; f \right) + \sqrt{2} (\min\{f(x_i) : 0 \leq i \leq n\} - M(f)). \tag{A.1.169}$$

Now we turn to the term  $\mathbb{E} \left( (M(f) - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right)$ .

For  $1 \leq k \leq n-2$ , define linear functions  $\tilde{v}_{l,k}(t) = \frac{f(x_k) - f(x_{k-1})}{1/n}(t - x_k) + f(x_k)$ ,  $\tilde{v}_{r,k}(t) = \frac{f(x_{k+2}) - f(x_{k+1})}{1/n}(t - x_{k+1}) + f(x_{k+1})$ . Then we know that when  $g$  is in  $\mathcal{F}$  such that  $g(x_i) = f(x_i)$ , for all  $0 \leq i \leq n$ ,  $\min_{t \in [x_k, x_{k+1}]} g(t)$  can and can only take value in

$$\left[ \min_{t \in [x_i, x_{i+1}]} \max\{\tilde{v}_{l,k}(t), \tilde{v}_{r,k}(t)\}, \min\{f(x_i), f(x_{i+1})\} \right].$$

Denote  $\tilde{h}(k) = \min_{t \in [x_i, x_{i+1}]} \max\{\tilde{v}_{l,k}(t), \tilde{v}_{r,k}(t)\}$ , for  $1 \leq k \leq n-2$ .

For  $k = 0$ , define linear function  $\tilde{v}_{r,0}(t) = \frac{f(x_2) - f(x_1)}{1/n}(t - x_1) + f(x_1)$ , and let  $\tilde{h}(0) = \min_{t \in [x_0, x_1]} \tilde{v}_{r,0}(t)$ , then similarly, we know that when  $g$  is in  $\mathcal{F}$  such that  $g(x_i) = f(x_i)$ , for all  $0 \leq i \leq n$ ,  $\min_{t \in [x_k, x_{k+1}]} g(t)$  can and can only take value in  $[\tilde{h}(0), \min\{f(x_0), f(x_1)\}]$ .

For  $k = n-1$ , define linear function  $\tilde{v}_{l,n-1}(t) = \frac{f(x_{n-1}) - f(x_{n-2})}{1/n}(t - x_{n-1}) + f(x_{n-1})$ , and let  $\tilde{h}(n-1) = \min_{t \in [x_{n-1/n}, 1]} \tilde{v}_{l,n-1}(t)$ , then similarly, we know that when  $g$  is in  $\mathcal{F}$  such that  $g(x_i) = f(x_i)$ , for all  $0 \leq i \leq n$ ,  $\min_{t \in [x_{n-1/n}, 1]} g(t)$  can and can only take value in

$$[\tilde{h}(n-1), \min\{f(x_{n-1}), f(x_n)\}].$$

Now we know that

$$\begin{aligned} \max\{M(g) : g \in \mathcal{F}, \text{ and } g(x_i) = f(x_i), \text{ for all } 0 \leq i \leq n\} &= \min\{f(x_i) : 0 \leq i \leq n\}, \\ \min\{M(g) : g \in \mathcal{F}, \text{ and } g(x_i) = f(x_i), \text{ for all } 0 \leq i \leq n\} &= \min\{\tilde{h}(i) : 0 \leq i \leq n-1\}. \end{aligned} \quad (\text{A.1.170})$$

Denote  $\check{h} = \min\{\tilde{h}(i) : 0 \leq i \leq n-1\}$ , and then we have

$$\begin{aligned} &\mathbb{E} \left( (M(f) - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\ &\leq (M(f) - \check{h}) + \mathbb{E} \left( (\check{h} - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\ &\leq (M(f) - \check{h}) + \sum_{i=I_{lo}-1}^{I_{hi}-2} \mathbb{E} \left( (\tilde{h}(i) - h(i))_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right). \end{aligned} \quad (\text{A.1.171})$$

We take the definition of  $\delta_i$  in Equation (A.1.148):  $\delta_i = y_{e,i-1} - f(x_{i-1})$ .

For  $(I_{lo} - 1) \vee 1 \leq k \leq (I_{hi} - 2) \wedge (n - 2)$ , we have

$$\begin{aligned} &\mathbb{E} \left( (\tilde{h}(i) - h(i))_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\ &\leq \mathbb{E} \left( \left( \min_{t \in [x_i, x_{i+1}]} \max\{\tilde{v}_{l,i}(t), \tilde{v}_{r,i}(t)\} - \right. \right. \\ &\quad \min_{t \in [x_i, x_{i+1}]} \max\{\tilde{v}_{l,i}(t) + (\delta_{i+1} - \delta_i - 2H)n(t - x_i) + \delta_{i+1} - H, \\ &\quad \left. \tilde{v}_{r,i}(t) + (\delta_{i+2} - \delta_{i+3} - 2H)n(x_{i+1} - t) + \delta_{i+2} - H\} \right)_+ \\ &\quad \left. \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\ &\leq P(\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J) (\mathbb{E}(2|\delta_{i+1}| + |\delta_i| + 2|\delta_{i+2}| + |\delta_{i+3}|) + 3H) \\ &\leq \left( 6 \cdot \gamma_e \sigma \sqrt{\frac{2}{\pi}} + 3\gamma_e S_{I_{hi}-I_{lo}+3, \frac{1}{8}} \sigma \right) P(\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J) \\ &\leq \bar{C}_{2,\alpha} \rho_m \left( \frac{\sigma}{\sqrt{n}}; f \right). \end{aligned} \quad (\text{A.1.172})$$

The last inequality is due to  $\sigma \mathbb{E}(\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J) < \mathbb{E}(\frac{\sigma}{\sqrt{2^{J-\hat{\mathbf{j}}}}})$ , and (A.1.167).

When  $I_{lo} = 1$ ,

$$\begin{aligned}
& \mathbb{E} \left( \left( \tilde{h}(0) - h(0) \right)_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\
& \leq \mathbb{E} \left( \left( \min_{t \in [0, 1/n]} \tilde{v}_{r,0}(t) - \min_{t \in [0, 1/n]} (\tilde{v}_{r,0}(t) + (\delta_3 - \delta_2 + 2H)n(t - x_1) + \delta_2 - H) \right)_+ \right. \\
& \quad \left. \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \tag{A.1.173} \\
& \leq P(\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J)(3H + 3\gamma_e \sigma \sqrt{\frac{2}{\pi}}) \\
& < \bar{C}_{2,\alpha} \rho_m(\frac{\sigma}{\sqrt{n}}; f).
\end{aligned}$$

When  $I_{hi} - 2 = n - 1$ ,

$$\begin{aligned}
& \mathbb{E} \left( \left( \tilde{h}(n-1) - h(n-1) \right)_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\
& \leq \mathbb{E} \left( \left( \min_{t \in [\frac{n-1}{n}, 1]} \tilde{v}_{l,n-1}(t) - \min_{t \in [\frac{n-1}{n}, 1]} (\tilde{v}_{l,n-1}(t) + (\delta_n - \delta_{n-1} - 2H)n(t - x_{n-1}) + \delta_n - H) \right)_+ \right. \\
& \quad \left. \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\
& \leq P(\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J)(3H + 3\gamma_e \sigma \sqrt{\frac{2}{\pi}}) \\
& < \bar{C}_{2,\alpha} \rho_m(\frac{\sigma}{\sqrt{n}}; f). \tag{A.1.174}
\end{aligned}$$

Going back to Inequality (A.1.171), we have

$$\mathbb{E} \left( (M(f) - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \leq (I_{hi} - I_{lo}) \bar{C}_{2,\alpha} \rho_m(\frac{\sigma}{\sqrt{n}}; f) + (M(f) - \check{h}). \tag{A.1.175}$$

Combing all the terms together, we have

$$\mathbb{E}_f L(CI_{m,\alpha}(Y)) \leq \check{C}_{4,\alpha} \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right) + \sqrt{2} \left(\min\{f(x_i) : 0 \leq i \leq n\} - \check{h}\right). \quad (\text{A.1.176})$$

□

*Proof of Lemma A.1.45.* The proof of this lemma is very similar to that of lemma A.1.44. For simplicity, we will omit the parts that are the same and only point out the places that are different.

Similar to Inequality (A.1.166), we have

$$\begin{aligned} & \mathbb{E}_f L(\text{CI}_{m,\alpha}(Y)) \\ & \leq (S_{I_{hi}-I_{lo}+1, \frac{\alpha}{4}} - \Phi^{-1}\left(\frac{\alpha}{4}\right) + \sqrt{3}) \gamma_e \mathbb{E}\left(\frac{\sigma}{\sqrt{2^{J-j_i}}}\right) \\ & \quad + \mathbb{E}\left(\left(\hat{\mathbf{f}}_1 - z_{\alpha/4} \frac{\sqrt{3}\sigma}{\sqrt{2^{J-j_i}}} - \frac{\sqrt{3}\sigma}{\sqrt{2^{J-j_i}}} - \mathbf{f}_{lo}\right)_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\}\right) \\ & \leq (S_{I_{hi}-I_{lo}+1, \frac{\alpha}{4}} - \Phi^{-1}\left(\frac{\alpha}{4}\right) + \sqrt{3}) \gamma_e \sigma + \\ & \quad \mathbb{E}\left(\left(\hat{\mathbf{f}}_1 - M(f)\right)_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\}\right) + \mathbb{E}\left(\left(M(f) - \mathbf{f}_{lo}\right)_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\}\right). \end{aligned} \quad (\text{A.1.177})$$

For the second term, according to the definition of  $\hat{\mathbf{f}}_1$  and Proposition A.1.8, we have

$$\begin{aligned} & \mathbb{E}\left(\left(\hat{\mathbf{f}}_1 - M(f)\right)_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\}\right) \\ & \leq \mathbb{E}\left(\left(\hat{M} - M(f)\right)_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\}\right) \\ & \leq \check{C}_3 \sup_{h \in \mathcal{G}_n(f)} \rho_m\left(\frac{\sigma}{\sqrt{n}}; h\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)} + \sqrt{2} \left(\min\{f(x_i) : 0 \leq i \leq n\} - M(f)\right), \end{aligned} \quad (\text{A.1.178})$$

where  $\hat{M}$  is defined in (2.4.9).

For  $\mathbb{E}\left(\left(M(f) - \mathbf{f}_{lo}\right)_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\}\right)$ , according to the arguments in the proof of Lemma

A.1.44, we have

$$\begin{aligned} & \mathbb{E} \left( (M(f) - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\ & \leq (M(f) - \check{h}) + \sum_{i=I_{lo}-1}^{I_{hi}-2} \mathbb{E} \left( (\tilde{h}(i) - h(i))_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right). \end{aligned} \quad (\text{A.1.179})$$

For  $(I_{lo} - 1) \vee 1 \leq k \leq (I_{hi} - 2) \wedge (n - 2)$ , we have

$$\begin{aligned} & \mathbb{E} \left( (\tilde{h}(i) - h(i))_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\ & \leq \left( 6 \cdot \gamma_e \sqrt{\frac{2}{\pi}} + 3\gamma_e S_{I_{hi}-I_{lo}+3, \frac{1}{8}} \right) P(\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J) \sigma. \end{aligned} \quad (\text{A.1.180})$$

When  $I_{lo} = 1$ ,

$$\begin{aligned} & \mathbb{E} \left( \left( \tilde{h}(0) - h(0) \right)_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\ & \leq P(\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J) (3H + 3\gamma_e \sigma \sqrt{\frac{2}{\pi}}). \end{aligned} \quad (\text{A.1.181})$$

When  $I_{hi} - 2 = n - 1$ ,

$$\begin{aligned} & \mathbb{E} \left( \left( \tilde{h}(n-1) - h(n-1) \right)_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\ & \leq P(\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J) (3H + 3\gamma_e \sigma \sqrt{\frac{2}{\pi}}). \end{aligned} \quad (\text{A.1.182})$$

Therefore,

$$\begin{aligned} & \mathbb{E} \left( (M(f) - \mathbf{f}_{lo})_+ \mathbb{1}\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J\} \right) \\ & \leq (I_{hi} - I_{lo}) \left( 6 \cdot \gamma_e \sqrt{\frac{2}{\pi}} + 3\gamma_e S_{I_{hi}-I_{lo}+3, \frac{1}{8}} \right) P(\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} > J) \sigma \\ & \quad + (M(f) - \check{h}). \end{aligned} \quad (\text{A.1.183})$$

Hence

$$\mathbb{E}_f L(\mathbf{CI}_{m,\alpha}(Y)) \leq \check{C}_{5,\alpha} \sigma + \sqrt{2} (\min\{f(x_i) : i = 0, 1, \dots, n\} - \check{h}). \quad (\text{A.1.184})$$

□

*Proof of Lemma A.1.46.* Similar to the proof of lemma A.1.16, define the following events:

$$\begin{aligned} \mathbf{E} &= \{Z(f) \notin [\frac{2^{J-j_l}(I_{lo} - 1)}{n}, \frac{2^{J-j_l}I_{hi} - 1}{n}] \cap [0, 1]\} \\ \mathbf{E}_1 &= \{\check{j} \geq \mathbf{j}^{\mathbf{w}} + K_{\frac{\alpha}{4}} + 1, \text{ and } \mathbf{j}^{\mathbf{w}} + K_{\frac{\alpha}{4}} + 1 \leq J\} \\ \mathbf{F} &= \{\check{j} \leq \mathbf{j}^* - 2 - \tilde{K}_{\frac{\alpha}{4}}\} \\ \mathbf{G} &= \{\mathbf{f}_{hi} < M(f)\} \\ \mathbf{H} &= \{\mathbf{f}_{lo} > M(f)\}. \end{aligned} \quad (\text{A.1.185})$$

Then we know that

$$\mathbf{E}_1^c \subset \mathbf{E}^c. \quad (\text{A.1.186})$$

So we have

$$\{M(f) \in \mathbf{CI}_{m,\alpha}(Y)\} \supset \mathbf{E}^c \cap \mathbf{F}^c \cap \mathbf{G}^c \cap \mathbf{H}^c \supset \mathbf{E}_1^c \cap \mathbf{F}^c \cap \mathbf{G}^c \cap \mathbf{H}^c. \quad (\text{A.1.187})$$



Then we have

$$\begin{aligned}
& P(M(f) \in \mathbf{CI}_{m,\alpha}(Y)) \\
& \geq P(\mathbf{E}_1^c \cap \mathbf{F}^c \cap \mathbf{G}^c \cap \mathbf{H}^c) \\
& = P(\mathbf{G}^c \cap \mathbf{H}^c | \mathbf{E}_1^c \cap \mathbf{F}^c) (1 - P(\mathbf{E}_1) - P(\mathbf{F}) + P(\mathbf{F} \cap \mathbf{E}_1)) \\
& = (1 - P(\mathbf{G} | \mathbf{E}_1^c \cap \mathbf{F}^c) - P(\mathbf{H} | \mathbf{E}_1^c \cap \mathbf{F}^c)) \\
& \quad + P(\mathbf{G} \cap \mathbf{H} | \mathbf{E}_1^c \cap \mathbf{F}^c) (1 - P(\mathbf{E}_1) - P(\mathbf{F}) + P(\mathbf{F} \cap \mathbf{E}_1)) \\
& \geq 1 - P(\mathbf{G} | \mathbf{E}_1^c \cap \mathbf{F}^c) - P(\mathbf{H} | \mathbf{E}_1^c \cap \mathbf{F}^c) - P(\mathbf{E}_1) - P(\mathbf{F}).
\end{aligned} \tag{A.1.188}$$

According to Lemma A.1.34, we have

$$\begin{aligned}
P(\mathbf{E}_1) & = P(\hat{\mathbf{j}} \geq \mathbf{j}^{\mathbf{w}} + K_{\frac{\alpha}{4}} + 1, \mathbf{j}^{\mathbf{w}} + K_{\frac{\alpha}{4}} + 1 \leq J) \\
& \leq P(\check{\mathbf{j}} \geq \mathbf{j}^{\mathbf{w}} + K_{\frac{\alpha}{4}} + 1, \mathbf{j}^{\mathbf{w}} + K_{\frac{\alpha}{4}} + 1) \\
& \leq \Phi(-2)^{K_{\frac{\alpha}{4}}} \leq \frac{\alpha}{4}.
\end{aligned} \tag{A.1.189}$$

Similar to the proof of Lemma A.1.16, especially the proof of Lemma A.1.18, which consists the proof of Lemma A.1.16, we have

$$P(\mathbf{F}) \leq P(\check{\mathbf{j}} \leq \mathbf{j}^* - 2 - \tilde{K}_{\frac{\alpha}{4}}) \leq \frac{\alpha}{4}. \tag{A.1.190}$$

For the remaining terms in Inequality (A.1.188), we claim

**Lemma A.1.47.**

$$P(\mathbf{H} | \mathbf{E}_1^c \cap \mathbf{F}^c) \leq \frac{\alpha}{4}. \tag{A.1.191}$$

*Proof.* With a little abuse of notation, let  $\mathbf{A}$  denote the event  $\{\hat{\mathbf{j}} + \tilde{K}_{\alpha/4} \leq J\}$  in the proof

of this lemma. Then

$$\begin{aligned}
P(H|E_1^c \cap F^c) = \\
P(H|E_1^c \cap F^c \cap A)P(A|E_1^c \cap F^c) + P(H|E_1^c \cap F^c \cap A^c)(1 - P(A|E_1^c \cap F^c)).
\end{aligned} \tag{A.1.192}$$

We start with the second term, for which we introduce another lemma.

**Lemma A.1.48.** *For  $h(i)$  defined in Algorithm 2,*

$$P(h(i) \leq \min_{t \in [x_i, x_{i+1}]} f(t) \text{ for all } I_{lo} - 1 \leq i \leq I_{hi} - 2 | Y_l, Y_s) \geq 1 - \alpha/4.$$

*Proof.* We take the definition of  $\delta_i$  in Equation (A.1.148):  $\delta_i = y_{e,i-1} - f(x_{i-1})$ . Since

$$\begin{aligned}
&P(\max\{|\delta_i| : (I_{lo} - 1) \vee 1 \leq i \leq (I_{hi} + 1) \wedge (n + 1)\} > H | Y_l, Y_s) \\
&\leq P(\max\{\delta_i : (I_{lo} - 1) \vee 1 \leq i \leq (I_{hi} + 1) \wedge (n + 1)\} > H | Y_l, Y_s) \\
&\quad + P(-\min\{\delta_i : I_{lo} \leq i \leq I_{hi}\} > H | Y_l, Y_s) \leq \alpha/4,
\end{aligned} \tag{A.1.193}$$

we have, condition on  $Y_l, Y_s$ , with probability at least  $1 - \alpha/4$ ,  $y_{e,i} - H \leq f(x_i)$  and  $y_{e,i} + H \geq f(x_i)$  for all  $(I_{lo} - 2)_+ \leq i \leq I_{hi} \vee n$ . With a bit of abuse of notation, let B denote the event that  $y_{e,i} - H \leq f(x_i)$  and  $y_{e,i} + H \geq f(x_i)$  for all  $(I_{lo} - 2)_+ \leq i \leq I_{hi} \vee n$ . On event B, for  $(I_{lo} - 1) \vee 1 \leq i \leq (I_{hi} - 2) \wedge (n - 2)$ , consider two linear functions  $\tilde{v}_{l,i}(t) = \frac{f(x_i) - f(x_{i-1})}{1/n}(t - x_i) + f(x_i)$ ,  $\tilde{v}_{r,i}(t) = \frac{f(x_{i+2}) - f(x_{i+1})}{1/n}(t - x_{i+1}) + f(x_{i+1})$ , then for  $t \in [x_i, x_{i+1}]$ ,  $f(t) \geq \max\{\tilde{v}_{l,i}(t), \tilde{v}_{r,i}(t)\} \geq \max\{v_{l,i}(t), v_{r,i}(t)\}$ , hence  $h(i) \leq \inf_{t \in [x_i, x_{i+1}]} f(t)$ .

Also, on event B, if  $I_{lo} - 1 = 0$ , then consider the linear function  $\tilde{v}_{r,0}(t) = \frac{f(x_2) - f(x_1)}{1/n}(t - x_1) + f(x_1)$ , for  $t \in [0, 1/n]$ ,  $f(t) \geq \tilde{v}_{r,0}(t) \geq v_{r,0}(t)$ , hence  $h(0) \leq \min_{t \in [0, 1/n]} f(t)$ .

Similarly, on event B, if  $I_{hi} - 2 = n - 1$ , we have  $h(n - 1) \leq \min_{t \in [n-1/n, 1]} f(t)$ .

Therefore, on event B,  $\min\{h(i) : I_{lo} - 1 \leq i \leq I_{hi} - 2\} \leq \inf_{t \in [x_{I_{lo}-1}, x_{I_{hi}-1}]} f(t)$ . Therefore,

$$P(h(i) \leq \min_{t \in [x_i, x_{i+1}]} f(t) \text{ for all } I_{lo} - 1 \leq i \leq I_{hi} - 2 | \mathbf{Y}_l, \mathbf{Y}_s) \geq P(B | \mathbf{Y}_l, \mathbf{Y}_s) \geq 1 - \alpha/4. \quad (\text{A.1.194})$$

□

Recalling that on event  $\mathbf{E}_1^c$ , we have  $Z(f) \in [x_{I_{lo}-1}, x_{I_{hi}-1}]$ , together with the lemma, we have

$$\begin{aligned} & P(H | \mathbf{E}_1^c \cap \mathbf{F}^c \cap A^c) \\ & \leq P(\min\{h(i) : I_{lo} - 1 \leq i \leq I_{hi} - 2\} > M(f) | \mathbf{E}_1^c \cap \mathbf{F}^c \cap A^c) \\ & = P(\min\{h(i) : I_{lo} - 1 \leq i \leq I_{hi} - 2\} > \min_{t \in [x_{I_{lo}-1}, x_{I_{hi}-1}]} f(t) | \mathbf{E}_1^c \cap \mathbf{F}^c \cap A^c) \leq \alpha/4. \end{aligned} \quad (\text{A.1.195})$$

Now we turn to the first term in Inequality (A.1.192).

$$\begin{aligned} & \left\{ \min_{I_{lo} \leq i \leq I_{hi}} \text{ave}_f(j_l, i) \leq M(f) + \frac{\sqrt{3}\sigma}{\sqrt{2^{J-j_l}}} \right\} \cap \mathbf{E}_1^c \cap A \\ & \supset \left\{ \min_{I_{lo} \leq i \leq I_{hi}} \text{ave}_f(j_l, i) \leq M(f) + \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right) \right\} \cap \{j_l > j^* - 2\} \cap \mathbf{E}_1^c \cap A \\ & \supset F^c \cap \{j_l > j^* - 2\} \cap \mathbf{E}_1^c \cap \{\hat{j} + \tilde{K}_{\alpha/4} \leq J\} \\ & \supset \mathbf{E}_1^c \cap A \cap \mathbf{F}^c. \end{aligned} \quad (\text{A.1.196})$$

Denote  $i_{min} = \arg \min_{I_{lo} \leq i \leq I_{hi}} \text{ave}_f(j_l, i)$ . When there is more than one qualifying for  $i_{min}$ , take anyone.

Therefore,

$$P(\mathbf{H} | \mathbf{E}_1^c \cap \mathbf{F}^c \cap A) \leq P(\mathfrak{E}_{j_l, i_{min}, e} \geq -\Phi^{-1}\left(\frac{\alpha}{4}\right)\sigma\gamma_e) \leq \frac{\alpha}{4}. \quad (\text{A.1.197})$$

Therefore,

$$P(\mathbf{H}|\mathbf{E}_1^c \cap \mathbf{F}^c) \leq \frac{\alpha}{4}. \quad (\text{A.1.198})$$

□

Similar to the arguments in proof of Lemma A.1.19, we have

$$P(\mathbf{G}|\mathbf{E}_1^c \cap \mathbf{F}^c) \leq \frac{\alpha}{4}. \quad (\text{A.1.199})$$

Returning to the main theorem, we have,

$$P(M(f) \in \mathbf{CI}_{m,\alpha}(Y)) \geq 1 - \alpha. \quad (\text{A.1.200})$$

□

## A.2. Proofs of Supporting Technical Lemmas for Chapter 2

We prove all the technical lemmas supporting Section A.1 in this section.

*Proof of Lemma A.1.1.* The inequalities are due to

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_3) - f(x_2)}{x_3 - x_2} \leq \frac{(x_3 - x_1)(f(x_2) - \frac{f(x_1)(x_3 - x_2) + f(x_3)(x_2 - x_1)}{x_3 - x_1})}{(x_2 - x_1)(x_3 - x_2)} \leq 0,$$

and

$$\frac{f(x_3) - f(x_1)}{x_3 - x_1} = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \cdot \frac{x_2 - x_1}{x_3 - x_1} + \frac{f(x_3) - f(x_2)}{x_3 - x_2} \cdot \frac{x_3 - x_2}{x_3 - x_1}.$$

□

*Proof of Lemma A.1.2.* Let  $t = x^{\frac{3}{2}}\sqrt{2/3} - 2$ , then we have

$$\begin{aligned} \frac{2x\Phi(2 - (2x)^{\frac{3}{2}}\sqrt{2/3})}{x\Phi(2 - \sqrt{2/3}x^{3/2})} &\leq 2 \frac{\int_{-\infty}^{-2\sqrt{2}t - (4\sqrt{2}-2)} \exp(-\frac{u^2}{2}) du}{\int_{-\infty}^{-t} \exp(-\frac{u^2}{2}) du} \\ &\leq 4\sqrt{2} \frac{\int_{-\infty}^{-2\sqrt{2}t} \exp(-\frac{u^2}{2}) du \exp(-\frac{(4\sqrt{2}-2)^2}{2})}{\int_{-\infty}^{-2\sqrt{2}t} \exp(-\frac{u^2}{16}) du} < 0.008. \end{aligned} \quad (\text{A.2.1})$$

□

*Proof of Lemma A.1.3.* Let

$$q(x) = x^2\Phi(-x)$$

Then

$$q(x)' = x(2\Phi(-x) - \frac{x}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})).$$

Taking further derivative, we know that  $\text{sign}((2\Phi(-x) - \frac{x}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}))') = \text{sign}(x^2 - 3)$ . Hence  $q(x)'/x$  goes down and then goes up, its first root is the place that  $q(x)$  takes maximum. Since  $q(1.19)' > 0$ ,  $q(1.2)' < 0$ , we have  $\sup_{x>0} q(x) \leq 1.2^2\Phi(-1.19) < 0.168514 < 0.169$ . Therefore  $Q \leq 1.2^2\Phi(-1.19) < 0.169$ . Only in this proof, let  $u(x) = x^2\Phi(2 - x)$ . We have  $u(x)' = x(2\Phi(2 - x) - x\frac{1}{\sqrt{2\pi}} \exp(-\frac{(2-x)^2}{2}))$ . Since  $\text{sign}((2\Phi(2 - x) - x\frac{1}{\sqrt{2\pi}} \exp(-\frac{(2-x)^2}{2}))') = \text{sign}(x(x - 2) - 3)$ , and  $\min_{x>0} u(x)' < 0 < u(1)'$ , we know  $u(x)'$  has at least 1 root. And its first root (when the root is unique, its first root is its unique root) is where  $u(x)$  takes maximum, since  $u(2.18)' > 0$ ,  $u(2.19)' < 0$ , we have  $u(x) \leq 2.19^2\Phi(2 - 2.18) < 2.0555$ . Hence  $V < 2.0555$ . □

*Proof of Lemma A.1.4.* Since we have for  $t > 0$ ,

$$\Phi(-t) \geq \frac{1}{\sqrt{2\pi}} \frac{t}{t^2 + 1} \exp(-t^2/2), \quad (\text{A.2.2})$$

we set  $t(\alpha) = \sqrt{2 \log(1/\alpha)} - \sqrt{\log(2 \log(1/\alpha))}$ . So we get, for  $\alpha < 0.03$ ,

$$\begin{aligned} \Phi(-t(\alpha)) &\geq \frac{1}{\sqrt{2\pi}} \alpha \exp(\log(2 \log(1/\alpha))) \cdot \left(\sqrt{\frac{\exp(1)}{1}} - \frac{1}{2}\right) \frac{t(\alpha)}{t(\alpha)^2 + 1} \\ &\geq \alpha \cdot (2 \log(1/\alpha))^{1.14} \frac{1}{\sqrt{2\pi}} \frac{t(\alpha)}{t(\alpha)^2 + 1}. \end{aligned} \quad (\text{A.2.3})$$

Further, denote  $x = 2 \log(1/\alpha)$ , we have

$$\frac{t(\alpha)}{t(\alpha)^2 + 1} x = \frac{t(\alpha)^2}{t(\alpha)^2 + 1} \frac{x}{t(\alpha)} \geq \frac{t(\alpha)^2}{t(\alpha)^2 + 1} \sqrt{x} > 0.6\sqrt{x} > 1.58. \quad (\text{A.2.4})$$

The inequalities are because of  $t(\alpha) = \sqrt{x} - \sqrt{\log x}$ ,  $t$  increases with  $x$  when  $x > 2$ , and  $x > 7$  when  $\alpha < 0.03$ .

Therefore, for  $\alpha < 0.03$

$$\Phi(-t(\alpha)) \geq 0.82\alpha. \quad (\text{A.2.5})$$

Therefore, for  $\alpha \leq 0.005$ ,  $z_{3\alpha} \geq t(\frac{3}{0.82}\alpha)$ ,  $z_{2.06\alpha} \geq t(\frac{2.06}{0.82}\alpha)$ .

Note that for  $\alpha < 0.02$ ,  $t(\alpha) \geq \sqrt{\log(1/\alpha)} \times 0.689$ .

Hence for  $\alpha \leq 0.005$ ,

$$\begin{aligned} z_{3\alpha} &\geq t(\frac{3}{0.82}\alpha) \geq 0.689 \times \sqrt{\log(0.82/3\alpha)} \geq 0.599\sqrt{\log(1/\alpha)}, \\ z_{2.06\alpha} &\geq t(\frac{2.06}{0.82}\alpha) \geq 0.689 \times \sqrt{\log(0.82/2.06\alpha)} \geq 0.627\sqrt{\log(1/\alpha)}. \end{aligned} \quad (\text{A.2.6})$$

We are now left with bounding

$$\inf_{\alpha \in (0.005, 0.08]} \frac{z_{2.06\alpha}}{\sqrt{\log 1/\alpha}}.$$

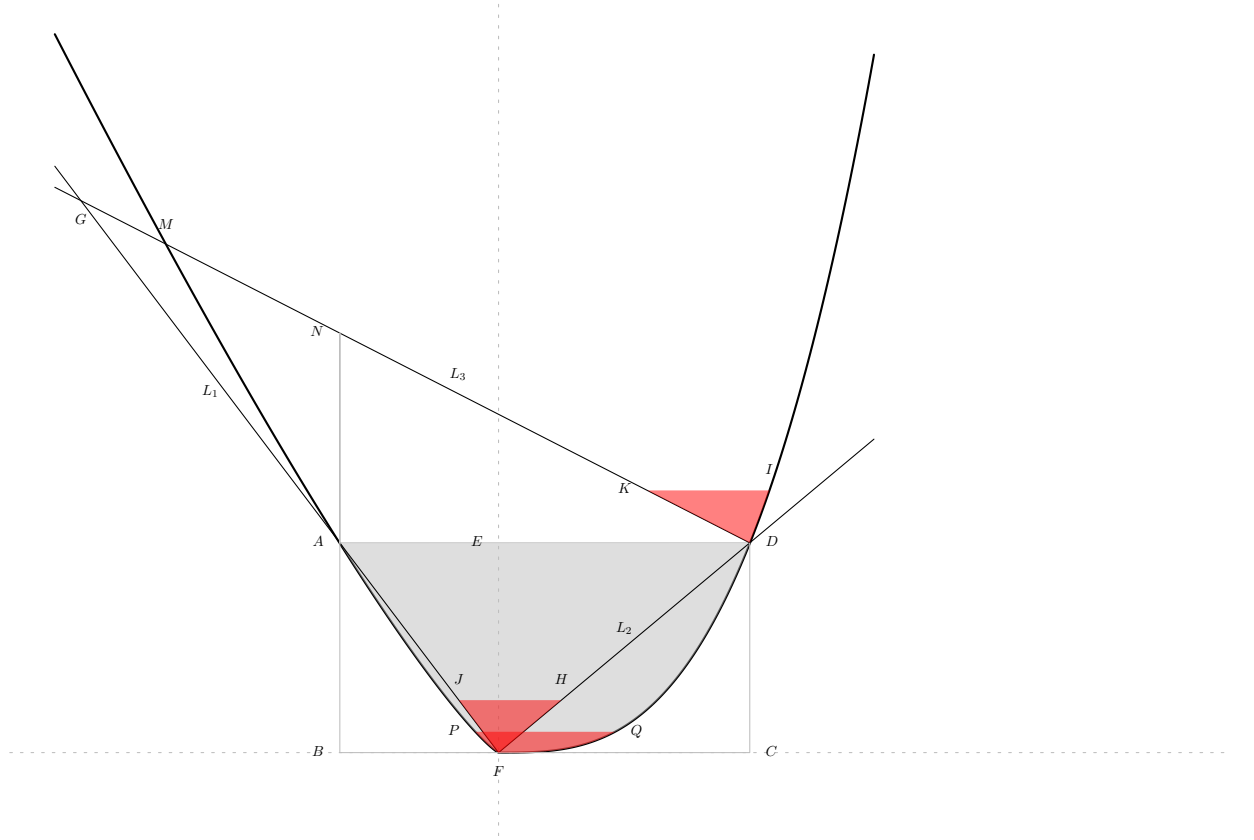
Note that both  $z_{2.06\alpha}$  and  $\sqrt{\log 1/\alpha}$  increases with  $\alpha$  decreasing. Therefore,

$$\inf_{\alpha \in (0.005, 0.08]} \frac{z_{2.06\alpha}}{\sqrt{\log 1/\alpha}} \geq \min_{5 \leq k \leq 79} \frac{z_{2.06 \frac{k+1}{1000}}}{\sqrt{\log 1000/k}} \geq 0.61.$$

Therefore, for  $\alpha < 0.08$ ,  $\frac{z_{2.06\alpha}}{\sqrt{\log 1/\alpha}} \geq 0.61$ .

□

*Proof of Lemma A.1.5.*



For  $\mu$  that will be designated later, define  $x_l = \arg \min\{t : f(t) \leq M(f) + \mu\}$ ,  $x_r = \arg \min\{t : f(t) \geq M(f) + \mu\}$ . Without loss of generality, we can assume  $x_r + x_l \geq 2Z(f)$ .

As shown in the figure, the function in bold is  $f$ , and the following points have the following

coordinates:

$$F : (Z(f), M(f)) \quad A : (x_l, M(f) + \mu) \quad D : (x_r, M(f) + \mu) \quad N : (x_l, M(f) + 2\mu) \quad (\text{A.2.7})$$

Define four linear functions:

$$\begin{aligned} L_0(t) &= M(f) + \mu & (AD), \\ L_1(t) &= M(f) + (t - Z(f)) \frac{\mu}{x_l - Z(f)} & (AF), \\ L_2(t) &= M(f) + (t - Z(f)) \frac{\mu}{x_r - Z(f)} & (FD), \\ L_3(t) &= M(f) + \mu + (t - x_r) \frac{\mu}{x_l - x_r} & (ND). \end{aligned} \quad (\text{A.2.8})$$

Define the following functions:

$$g_1 = \max(f, L_0), \quad g_2 = \max(f, L_3), \quad g_3 = \max(L_1, L_2, L_0), \quad g_4 = \max(L_1, L_2). \quad (\text{A.2.9})$$

Therefore, with  $\mu$  increasing from  $0^+$  to  $\infty$ ,  $\|g_1 - f\|$  and  $\|g_2 - f\|$  increase from  $0^+$  to  $\infty$ .

Then we know that for any given  $\sigma > 0$ , either  $\exists \mu > 0$ , s.t.  $\|g_2 - f\| = \sigma$ , or  $\exists \mu$  such that the following three things hold.

Property 1.  $x_l(\mu) + x_r(\mu) = 2Z(f)$ .

Property 2. Suppose  $g_{2,l}$  and  $g_{2,r}$  are constructed essentially in the same way as  $g_2$  but one on the left side ( $g_{2,l}$ ) and one on the right side ( $g_{2,r}$ ). Then  $(\|g_{2,l} - f\| - \sigma) \cdot (\|g_{2,r} - f\| - \sigma) < 0$ .

Property 3. And further, for the side ( $h \in \{l, r\}$ ) that  $\|g_{2,h} - f\| - \sigma < 0$ ,  $\exists \mu > \tau_h > 0$  such that for any  $\tau \in (0, \tau_h)$ ,

$$|x_h(\mu - \tau) - Z(f)| \geq \frac{|x_l(\mu - \tau) - Z(f)| + |x_r(\mu - \tau) - Z(f)|}{2}.$$



And for the other side  $\tilde{h} \in \{l, r\}/\{h\}$ ,  $\|g_{2,\tilde{h}} - f\| - \sigma > 0$ ,  $\exists \mu > \tau_{\tilde{h}} > 0$  such that for any  $\tau \in (0, \tau_{\tilde{h}})$ ,

$$|x_{\tilde{h}}(\mu + \tau) - Z(f)| \geq \frac{|x_l(\mu + \tau) - Z(f)| + |x_r(\mu + \tau) - Z(f)|}{2}.$$

To show the main idea more clearly, we assume for the moment that for the  $\sigma$  that we will designated later, there exists a  $\mu$  such that on at least one side, we have  $\|g_2 - f\| = \sigma$  and use  $\sigma$  to denote  $\|g_2 - f\|$ . For the  $\sigma$  that does not have a corresponding  $\mu$ , we will discuss it later.

To lower bound  $\|g_1 - f\|$  by a quantity related to  $\|g_2 - f\|$ , we have

$$\begin{aligned} \|g_2 - f\|^2 &\leq \frac{1}{3}\mu^3 \frac{1}{\frac{\mu}{Z(f)-x_l} - \frac{\mu}{x_r-x_l}} + 2 \times \frac{1}{3}(x_r - x_l) \times \mu^2 + 2 \times \|g_1 - f\|^2 \\ &= \frac{1}{3}\mu^2 \frac{(Z(f) - x_l)(x_r - x_l)}{x_r - Z(f)} + \frac{2}{3}\mu^2(x_r - x_l) + 2\|g_1 - f\|^2 \\ &\leq \mu^2(x_r - x_l) + 2 \times \|g_1 - f\|^2 \leq 5\|g_1 - f\|^2. \end{aligned} \tag{A.2.10}$$

To lower bound  $\|g_3 - g_4\|$  with  $\|g_1 - f\|$  or  $\|g_2 - f\|$ , we have

$$\|g_3 - g_4\|^2 \geq \frac{1}{3}\mu^2(x_r - x_l) \geq \frac{1}{3}\|g_1 - f\|^2, \tag{A.2.11}$$

and

$$\begin{aligned} &\|g_2 - f\|^2 \\ &\leq \frac{1}{3}\mu^3 \frac{1}{\frac{\mu}{Z(f)-x_l} - \frac{\mu}{x_r-x_l}} + \frac{1}{3}(x_r - x_l) \times \mu^2 + \|g_1 - f\|^2 + 2 \times \mu^2 \times \frac{1}{2}(x_r - x_l) \\ &= \frac{1}{3}\mu^2 \frac{4x_r - x_l - 3Z(f)}{x_r - Z(f)}(x_r - x_l) + \|g_1 - f\|^2 \\ &\leq \frac{5}{3}\mu^2(x_r - x_l) + \|g_1 - f\|^2 \leq 8\|g_3 - g_4\|^2. \end{aligned} \tag{A.2.12}$$

Define linear function  $g_5 = \max\{L_3, L_2\}$ , then we know that  $\rho_z(\gamma; g_2) \leq \rho_z(\gamma; g_5), \forall \gamma > 0$ .

Since we also have

$$\begin{aligned}\gamma^2 &= \frac{1}{3}\rho_m(\gamma; g_5)^3 \times \left(\frac{x_r - x_l}{\mu} + \frac{x_r - Z(f)}{\mu}\right), \\ \gamma^2 &= \frac{1}{3}\rho_m(\gamma; g_4)^3 \times \frac{x_r - x_l}{\mu} = \frac{1}{3}\rho_z(\gamma; g_4)^3 \left(\frac{\mu}{x_r - Z(f)}\right)^3 \frac{x_r - x_l}{\mu},\end{aligned}\tag{A.2.13}$$

we have

$$\begin{aligned}\rho_z(\gamma; g_2)^3 &\leq \rho_z(\gamma; g_5)^3 = \left(\rho_m(\gamma; g_5) \frac{x_r - x_l}{\mu}\right)^3 \\ &= \frac{\left(\frac{x_r - x_l}{\mu}\right)^3 3\gamma^2}{\frac{x_r - x_l}{\mu} + \frac{x_r - Z(f)}{\mu}} \\ &= \frac{\left(\frac{x_r - x_l}{\mu}\right)^3 \rho_z(\gamma; g_4)^3 \left(\frac{\mu}{x_r - Z(f)}\right)^3 \frac{x_r - x_l}{\mu}}{\frac{x_r - x_l}{\mu} + \frac{x_r - Z(f)}{\mu}} \\ &= \rho_z(\gamma; g_4)^3 \frac{(x_r - x_l)^4}{(x_r - Z(f))^3 (2x_r - x_l - Z(f))} \leq \frac{16}{3}\rho_z(\gamma; g_4)^3.\end{aligned}\tag{A.2.14}$$

Also, we have

$$\begin{aligned}\rho_z(\gamma; g_4) &= \left(\frac{\gamma}{\|g_3 - g_4\|}\right)^{\frac{2}{3}} (x_r - Z(f)) \leq \left(\frac{\sqrt{8}\gamma}{\|g_2 - f\|}\right)^{\frac{2}{3}} (x_r - Z(f)) \\ &\leq \left(\frac{\sqrt{8}\gamma}{\|g_2 - f\|}\right)^{\frac{2}{3}} |Z(g_2) - Z(f)| = \left(\frac{\sqrt{8}\gamma}{\sigma}\right)^{\frac{2}{3}} |Z(g_2) - Z(f)|.\end{aligned}\tag{A.2.15}$$

Therefore, we have

$$\rho_z(\gamma; g_2) \leq \frac{2^{\frac{7}{3}}}{3^{\frac{1}{3}}} \left(\frac{\gamma}{\sigma}\right)^{\frac{2}{3}} |Z(g_2) - Z(f)|.\tag{A.2.16}$$

Further we have

$$|Z(g_2) - Z(f)| = \sup\{|t - Z(f)| : g_1(t) = M(g_1)\} \geq \rho_z\left(\frac{1}{\sqrt{5}}\sigma; f\right).\tag{A.2.17}$$

The  $\sigma$  we will specify later is no smaller than  $\sqrt{5}\varepsilon$ , and suppose  $\sigma \geq \sqrt{5}\varepsilon$  from now. This gives  $|Z(g_2) - Z(f)| \geq \rho_z(\frac{1}{\sqrt{5}}\sigma; f) \geq \rho_z(\varepsilon; f)$ .

As we know, for the problem of estimation  $Z(h)$  with  $h \in \{g_2, f\}$ , the following statistic is sufficient

$$WS = \frac{\int_0^1 (g_2(t) - f(t))dY(t) - \frac{1}{2} \int_0^1 (g_2(t)^2 - f(t)^2)dt}{\varepsilon \|g_2 - f\|}, \quad (\text{A.2.18})$$

and we have  $WS \sim N(\theta(h)\frac{\|g_2 - f\|}{2\varepsilon}, 1)$ , with  $\theta(g_2) = 1, \theta(f) = -1$ .

Define an event  $O = \{|\hat{Z} - Z(f)| > \frac{1}{2}|Z(g_2) - Z(f)|\}$ , then we have  $P_f(O) \leq 2c$ . This is because we have  $\mathbb{E}_f|\hat{Z} - Z(f)| \leq c\rho_z(\varepsilon; f)$ , and  $|Z(g_2) - Z(f)| \geq \rho_z(\varepsilon; f)$ . Since we further have  $|\hat{Z} - Z(g_2)| \geq |Z(g_2) - Z(f)| - |\hat{Z} - Z(f)|$ , the following inequalities hold

$$\begin{aligned} \mathbb{E}_{g_2}|\hat{Z} - Z(g_2)| &\geq \mathbb{E}_{g_2} \left( (|Z(g_2) - Z(f)| - |\hat{Z} - Z(f)|)_+ \right) \\ &\geq \mathbb{E}_{g_2} \left( \mathbb{1}\{O^c\} \left( |Z(g_2) - Z(f)| - \frac{1}{2}\rho_z(\varepsilon; f) \right) \right) \\ &\geq \Phi(\Phi^{-1}(1 - 2c) - \frac{\|g_2 - f\|}{\varepsilon}) \frac{1}{2}|Z(g_2) - Z(f)| \end{aligned} \quad (\text{A.2.19})$$

For  $c \leq 0.0063$ , let  $\sigma = \Phi^{-1}(1 - 2c)\varepsilon$ . Then  $\sigma > \sqrt{5}\varepsilon$ , thus  $|Z(g_2) - Z(f)| \geq \rho_z(\varepsilon; f)$ .

So we have

$$\begin{aligned} \mathbb{E}_{g_2}|\hat{Z} - Z(g_2)| &\geq \frac{1}{4}|Z(g_2) - Z(f)| \\ &\geq \frac{1}{4} \left( \frac{3}{27} \right)^{\frac{1}{3}} \Phi^{-1}(1 - 2c)^{\frac{2}{3}} \rho_z(\varepsilon; g_2). \end{aligned} \quad (\text{A.2.20})$$

Let  $f_1 = g_2$ , we have the result.

Now we consider the case when  $\sigma = \Phi^{-1}(1 - 2c)\varepsilon$  does not have a corresponding  $\mu$ . Then

$\exists \mu > 0$  such that Property 1., Property 2. and Property 3. hold.

Without loss of generality, we assume  $h$  defined in Property 3. is  $r$ . According to Property 1., and construction of  $g_{2,l}$  and  $g_{2,r}$  we know that  $\rho_z(\beta; g_{2,l}) = \rho_z(\beta; g_{2,r})$  for all  $\beta > 0$ . Besides  $g_{2,l}$  and  $g_{2,r}$ , we can construct  $g_{1,l}, g_{3,l}, g_{4,l}, g_{5,l}$  similarly to  $g_1, g_3, g_4, g_5$  on the left hand side, and also  $g_{1,r}, g_{3,r}, g_{4,r}, g_{5,r}$  on the right hand side. Then we know that  $g_{1,l} = g_{1,r}, g_{3,l} = g_{3,r}, g_{4,l} = g_{4,r}$ . According to Inequality (A.2.14) and Inequality (A.2.15), we have

$$\begin{aligned}
|Z(g_{2,r}) - Z(f)| &= |Z(g_{2,l}) - Z(f)| \geq \left( \frac{\|g_{2,l} - f\|}{\sqrt{8\varepsilon}} \right)^{\frac{3}{2}} \rho_z(\varepsilon; g_{4,l}) \\
&\geq \left( \frac{\|g_{2,l} - f\|}{\sqrt{8\varepsilon}} \right)^{\frac{2}{3}} \left( \frac{3}{16} \right)^{\frac{1}{3}} \rho_z(\varepsilon; g_{2,l}) \\
&= \left( \frac{\|g_{2,l} - f\|}{\sqrt{8\varepsilon}} \right)^{\frac{2}{3}} \left( \frac{3}{16} \right)^{\frac{1}{3}} \rho_z(\varepsilon; g_{2,r}) \\
&\geq \left( \frac{3}{2^7} \right)^{\frac{1}{3}} \Phi^{-1}(1 - 2c)^{\frac{2}{3}} \rho_z(\varepsilon; g_{2,r}).
\end{aligned} \tag{A.2.21}$$

The last inequality is because of  $\|g_{2,l} - f\| > \sigma = \Phi^{-1}(1 - 2c)\varepsilon$ , coming from Property 3. and Property 2.. Again, since  $\sigma \geq \sqrt{5}\varepsilon$ , we have  $|Z(g_{2,r}) - Z(f)| = |Z(g_{2,l}) - Z(f)| \geq \rho_z(\varepsilon; f)$ , which comes from (A.2.17).

Similar to the arguments in the case of  $g_2$ , we define event  $O = \{|\hat{Z} - Z(f)| > \frac{1}{2}|Z(g_{2,r}) - Z(f)|\}$ , then we have  $P_f(O) \leq 2c$ . And we have

$$\begin{aligned}
\mathbb{E}_{g_{2,r}}|\hat{Z} - Z(g_{2,r})| &\geq \mathbb{E}_{g_{2,r}} \left( (|Z(g_{2,r}) - Z(f)| - |\hat{Z} - Z(f)|)_+ \right) \\
&\geq \mathbb{E}_{g_{2,r}} \left( \mathbb{1}\{O^c\} \left( |Z(g_{2,r}) - Z(f)| - \frac{1}{2}\rho_z(\varepsilon; f) \right) \right) \\
&\geq \Phi(\Phi^{-1}(1 - 2c) - \frac{\|g_{2,r} - f\|}{\varepsilon}) \frac{1}{2} |Z(g_{2,r}) - Z(f)| \\
&\geq \frac{1}{4} \left( \frac{3}{2^7} \right)^{\frac{1}{3}} \Phi^{-1}(1 - 2c)^{\frac{2}{3}} \rho_z(\varepsilon; g_{2,r}).
\end{aligned} \tag{A.2.22}$$

We take  $f_1 = g_{2,l}$  and get the statement.

□

*Proof of Lemma A.1.6.* Without loss of generality, we can assume  $f(Z(f) + \rho_z(\varepsilon; f)) \leq M(f) + \rho_m(\varepsilon; f)$ . Denote  $x_l = \min\{t : f(t) \leq M(f) + \rho_m(\varepsilon; f)\}$ .

For  $0 < \delta < \frac{1}{2}\rho_z(\varepsilon; f)$ , denote

$$g_\delta(t) = \max \left\{ f(t), \right. \\ \left. M(f) + \rho_m(\varepsilon; f) + \frac{f(Z(f) + \rho_z(\varepsilon; f) - \delta) - M(f) - \rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f) + Z(f) - x_l - \delta}(t - x_l) \right\}. \quad (\text{A.2.23})$$

Then  $\|g_\delta - f\| \leq \varepsilon$ . And  $\rho_z(\varepsilon; g_\delta) \leq 3\rho_z(\varepsilon; f)$ .

Define event  $O$  to be  $O = \{|\hat{Z} - Z(f)| \geq \frac{1}{2}\rho_z(\varepsilon; f)\}$ . Then  $P_f(O) \leq 2c$ , thus  $P_{g_\delta}(O) \leq \Phi(1 + \Phi^{-1}(2c))$ .

Therefore,

$$\begin{aligned} & \mathbb{E}_{g_\delta} |\hat{Z} - Z(g_\delta)| \\ & \geq \mathbb{E}_{g_\delta} \left( \mathbb{1}\{O^c\} (|Z(f) - Z(g_\delta)| - \frac{1}{2}\rho_z(\varepsilon; f))_+ \right) \\ & \geq P_{g_\delta}(O^c) (\rho_z(\varepsilon; f) - \delta - \frac{1}{2}\rho_z(\varepsilon; f)) \\ & \geq (1 - \Phi(1 + \Phi^{-1}(2c))) (\rho_z(\varepsilon; f) - \delta - \frac{1}{2}\rho_z(\varepsilon; f)) \\ & \geq (1 - \Phi(1 + \Phi^{-1}(2c))) \left( \frac{1}{2} - \frac{\delta}{\rho_z(\varepsilon; f)} \right)_+ \rho_z(\varepsilon; f) \\ & \geq (1 - \Phi(1 + \Phi^{-1}(2c))) \left( \frac{1}{2} - \frac{\delta}{\rho_z(\varepsilon; f)} \right)_+ \frac{\rho_z(\varepsilon; g_\delta)}{3}. \end{aligned} \quad (\text{A.2.24})$$

Therefore,

$$\begin{aligned}
& \sup_{\frac{1}{2}\rho_z(\varepsilon;f) > \delta > 0} \frac{\mathbb{E}_{g_\delta} |\hat{Z} - Z(g_\delta)|}{\rho_z(\varepsilon; g_\delta)} \\
& \geq \limsup_{\delta \rightarrow 0^+} \frac{(1 - \Phi(1 + \Phi^{-1}(2c)))}{3} \left( \frac{1}{2} - \frac{\delta}{\rho_z(\varepsilon; f)} \right)_+ \\
& = \frac{1}{6} (1 - \Phi(1 + \Phi^{-1}(2c))) > 0.1666 (1 - \Phi(1 + \Phi^{-1}(2c))).
\end{aligned} \tag{A.2.25}$$

Note that the inequality is strict, so we have the statement.

□

*Proof of Lemma A.1.7.* Without loss of generality, we can assume

$$t_r = \max\{t : f(t) \leq M(f) + \rho_m(\gamma; f)\} = Z(f) + \rho_z(\gamma; f).$$

Denote

$$t_l = \min\{t : f(t) \leq M(f) + \rho_m(\gamma; f)\} = Z(f) + \rho_z(\gamma; f).$$

It's apparent that  $t_r$  and  $t_l$  depend on  $\gamma$ . For  $\frac{1}{4}\rho_z(\gamma; f) > \delta > 0$ , define

$$g_\delta(\gamma; f) = \max\{f, M(f) + \rho_m(\gamma; f) + \frac{f(t_r - \delta) - M(f) - \rho_m(\gamma; f)}{t_r - \delta - t_l}(t - t_l)\}.$$

Therefore, we know that  $\|g_\delta(\gamma; f) - f\| \leq \gamma$ . We will use  $g$  to refer to  $g_\delta(\gamma; f)$  when there is no ambiguity. According to the definition, we know that  $\limsup_{\delta \rightarrow 0^+} \rho_m(\gamma; g) \leq \rho_m(\gamma; f)$ . We will specify  $\gamma$  to be a quantity no smaller than  $\varepsilon$ , suppose  $\gamma \geq \varepsilon$  from now.

Denote  $O = \{|\hat{M} - M(f)| > \frac{1}{2}\rho_m(\varepsilon; f)\}$ . Since  $\mathbb{E}_f |\hat{M} - M(f)| \leq c\rho_m(\varepsilon; f)$ , we have  $P_f(O) \leq 2c$ , then we have

$$\begin{aligned}
\mathbb{E}_g |\hat{M} - M(g)| &\geq \mathbb{E}_g \left( \mathbb{1}\{O^c\} (|M(f) - M(g)| - |\hat{M} - M(f)|)_+ \right) \\
&\geq P_g(O^c) (|M(f) - M(g)| - \frac{1}{2}\rho_m(\varepsilon; f))_+ \\
&\geq \Phi(\Phi^{-1}(1 - 2c) - \frac{\gamma}{\varepsilon}) \left( |M(f) - M(g)| - \frac{1}{2}\rho_m(\varepsilon; f) \right)_+ \\
&= \Phi(\Phi^{-1}(1 - 2c) - \frac{\gamma}{\varepsilon}) \left( \rho_m(\gamma; f) - \frac{1}{2}\rho_m(\varepsilon; f) + f(t_r - \delta) - f(t_r) \right)_+.
\end{aligned} \tag{A.2.26}$$

For  $c \leq 0.103$ , let  $\gamma = \max\{\Phi^{-1}(1 - 2c)\varepsilon, \varepsilon\}$ . Then  $\gamma \geq \varepsilon$ .

Therefore, we have

$$\begin{aligned}
\sup_{0 < \delta < \frac{1}{4}\rho_z(\gamma; f)} \frac{\mathbb{E}_g |\hat{M} - M(g)|}{\rho_m(\varepsilon; g)} &\geq \limsup_{\delta \rightarrow 0^+} \frac{\mathbb{E}_g |\hat{M} - M(g)|}{\rho_m(\varepsilon; g)} \\
&\geq \limsup_{\delta \rightarrow 0^+} \frac{\Phi(z_{2c} - \max\{z_{2c}, 1\}) \left( \left( \frac{\gamma}{\varepsilon} \right)^{\frac{2}{3}} \rho_m(\varepsilon; f) - \frac{1}{2}\rho_m(\varepsilon; f) + f(t_r - \delta) - f(t_r) \right)_+}{\rho_m(\varepsilon; g)} \\
&\geq \frac{\Phi(z_{2c} - \max\{z_{2c}, 1\}) \left( \left( \frac{\gamma}{\varepsilon} \right)^{\frac{2}{3}} \rho_m(\varepsilon; f) - \frac{1}{2}\rho_m(\varepsilon; f) \right)}{\rho_m(\varepsilon; f)} \\
&= \Phi(z_{2c} - \max\{z_{2c}, 1\}) \left( \left( \frac{\gamma}{\varepsilon} \right)^{\frac{2}{3}} - \frac{1}{2} \right).
\end{aligned} \tag{A.2.27}$$

For  $0.103 \geq c \geq \frac{\Phi(-1)}{2}$ , we have

$$\sup_{g \in \mathcal{F}} \frac{\mathbb{E}_g |\hat{M} - M(g)|}{\rho_m(\varepsilon; g)} \geq \frac{\Phi(z_{2c} - 1)}{2} > 0.214362. \tag{A.2.28}$$

For  $c < \frac{\Phi(-1)}{2}$ , we have

$$\sup_{g \in \mathcal{F}} \frac{\mathbb{E}_g |\hat{M} - M(g)|}{\rho_m(\varepsilon; g)} \geq \frac{1}{2} \left( z_{2c}^{\frac{2}{3}} - \frac{1}{2} \right) > \frac{z_{2c}^{\frac{2}{3}}}{4}. \tag{A.2.29}$$

Note that for both cases, the inequality is strict, so we have the statement.

□

*Proof of Lemma A.1.12.* Without loss of generality, we assume

$$\sup\{t > Z(f) : f(t) \leq \rho_m(\varepsilon; f) + M(f)\} = \rho_z(\varepsilon; f) + Z(f).$$

$$\begin{aligned} & \mathbb{E}_{l,s}((\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} \geq \hat{j} + 1\}) \\ &= \mathbb{E}_{l,s}\left(\sum_{j_1=2}^{j^*+1} (\hat{f} - M(f))^2 \mathbb{1}\{\hat{j} = j_1, \tilde{j} \geq j_1 + 1\}\right) \\ & \quad + \mathbb{E}_{l,s}\left(\sum_{j_1=j^*+2}^{\infty} (\hat{f} - M(f))^2 \mathbb{1}\{\hat{j} = j_1, \tilde{j} \geq j_1 + 1\}\right) \end{aligned} \tag{A.2.30}$$



Now we will first bound the first term in Inequality (A.2.30)

$$\begin{aligned}
& \mathbb{E}_{l,s} \left( \sum_{j_1=2}^{j^*+1} (\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1\} \right) \\
& \leq \mathbb{E}_{l,s} \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}+2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} \right) \\
& \quad + \mathbb{E}_{l,s} \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} \right) \\
& \quad + \mathbb{E}_{l,s} \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \right. \\
& \quad \quad \left. \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} \right) \\
& \leq \mathbb{E}_{l,s} \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}+2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} \right) \\
& \quad + \mathbb{E}_{l,s} \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} \right) \\
& \quad + \mathbb{E}_{l,s} \left( \sum_{j_1=2}^{j^*+1} \frac{1}{2} [(\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 + (\mu_{j_1, \hat{i}_{j_1}+2} - M(f))^2] \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \right. \\
& \quad \quad \left. \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} \right) \\
& \leq \frac{3}{2} \mathbb{E}_{l,s} \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}+2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} \right) \\
& \quad + \frac{3}{2} \mathbb{E}_{l,s} \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} \right) \\
& = \frac{3}{2} \mathbb{E}_l \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}+2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \right) \\
& \quad \mathbb{E}_s \left( \mathbb{1}\{\mathcal{E}_{j_1, \hat{i}_{j_1}+6} \frac{1}{\sqrt{2}c_s\varepsilon} \leq 2 - \mu_{j_1, \hat{i}_{j_1}+6} \frac{\sqrt{m_{j_1}}}{\sqrt{2}c_s\varepsilon} + \mu_{j_1, \hat{i}_{j_1}+5} \frac{\sqrt{m_{j_1}}}{\sqrt{2}c_s\varepsilon} | Y_l \} \right) \\
& \quad + \frac{3}{2} \mathbb{E}_l \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \right) \\
& \quad \mathbb{E}_s \left( \mathbb{1}\{-\mathcal{E}_{j_1, \hat{i}_{j_1}+5} \frac{1}{\sqrt{2}c_s\varepsilon} \leq 2 - \mu_{j_1, \hat{i}_{j_1}-6} \frac{\sqrt{m_{j_1}}}{\sqrt{2}c_s\varepsilon} + \mu_{j_1, \hat{i}_{j_1}-5} \frac{\sqrt{m_{j_1}}}{\sqrt{2}c_s\varepsilon} | Y_l \} \right).
\end{aligned}$$

(A.2.31)

Now we will bound  $\mu_{j_1, \hat{i}_{j_1}-6} - \mu_{j_1, \hat{i}_{j_1}-5}$  by an expression of  $\mu_{j_1, \hat{i}_{j_1}-2} - M(f)$ . As we have  $|\hat{i}_{j_1} - i_{j_1}^*| \leq 1$ , we have  $i_{j_1}^* - 3 \leq \hat{i}_{j_1} - 2 \leq i_{j_1}^* - 1$ . We have

$$\begin{aligned} \mu_{j_1, \hat{i}_{j_1}-6} - \mu_{j_1, \hat{i}_{j_1}-5} &\geq m_{j_1} \frac{f(t_{j_1, \hat{i}_{j_1}-6}) - M(f)}{t_{j_1, \hat{i}_{j_1}-6} - Z(f)} \geq m_{j_1} \frac{f(t_{j_1, \hat{i}_{j_1}-3}) - M(f)}{t_{j_1, \hat{i}_{j_1}-3} - Z(f)} \\ &\geq m_{j_1} \frac{\mu_{j_1, \hat{i}_{j_1}-2} - M(f)}{4m_{j_1}} \geq \frac{1}{4}(\mu_{j_1, \hat{i}_{j_1}-2} - M(f)). \end{aligned} \quad (\text{A.2.32})$$

Similarly we have

$$\mu_{j_1, \hat{i}_{j_1}+6} - \mu_{j_1, \hat{i}_{j_1}+5} \geq \frac{1}{4}(\mu_{j_1, \hat{i}_{j_1}+2} - M(f)). \quad (\text{A.2.33})$$

In addition, for  $j_1 = j^*$  and  $j_1 = j^* + 1$  in the first term, we have

$$\begin{aligned} &\mathbb{E}_{l,s} \left( (\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2c_s} \sqrt{m_{j_1}} \varepsilon} \leq 2\} \right) \\ &\leq \rho_m(\varepsilon; f)^2 2^{2j^* - 2j_1}. \end{aligned}$$

Going back to Inequality (A.2.31),

$$\begin{aligned}
& \mathbb{E}_{l,s} \left( \sum_{j_1=2}^{j^*+1} (\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1\} \right) \\
& \leq \frac{1}{2} \left( 3 * \rho_m(\varepsilon; f)^2 + 3 * \frac{1}{4} \rho_m(\varepsilon; f)^2 \right) + \frac{3}{2} \mathbb{E}_l \left( \sum_{j_1=2}^{j^*-1} (\mu_{j_1, \hat{i}_{j_1}+2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \right. \\
& \quad \mathbb{E}_s \left( \mathbb{1}\{\mathcal{E}_{j_1, \hat{i}_{j_1}+6} \frac{1}{\sqrt{2}c_s\varepsilon} \leq 2 - \frac{1}{4}(\mu_{j_1, \hat{i}_{j_1}+2} - M(f)) \frac{\sqrt{m_{j_1}}}{\sqrt{2}c_s\varepsilon} | Y_l \} \right) \\
& \quad \left. + \frac{3}{2} \mathbb{E}_l \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \right. \right. \\
& \quad \mathbb{E}_s \left( \mathbb{1}\{-\mathcal{E}_{j_1, \hat{i}_{j_1}-5} \frac{1}{\sqrt{2}c_s\varepsilon} \leq 2 - \frac{1}{4}(\mu_{j_1, \hat{i}_{j_1}-2} - M(f)) \frac{\sqrt{m_{j_1}}}{\sqrt{2}c_s\varepsilon} | Y_l \} \right) \left. \right) \\
& = \frac{15}{8} \rho_m(\varepsilon; f)^2 + \frac{3}{2} \mathbb{E}_l \left( \sum_{j_1=2}^{j^*-1} (\mu_{j_1, \hat{i}_{j_1}+2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \right. \\
& \quad \Phi \left( 2 - (\mu_{j_1, \hat{i}_{j_1}+2} - M(f)) 2^{\frac{j^*-j_1-4}{2}} \frac{\sqrt{m_{j^*}}}{\sqrt{2}c_s\varepsilon} \right) \\
& \quad \left. + \frac{3}{2} \mathbb{E}_l \left( \sum_{j_1=2}^{j^*+1} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \Phi \left( 2 - (\mu_{j_1, \hat{i}_{j_1}-2} - M(f)) 2^{\frac{j^*-j_1-4}{2}} \frac{\sqrt{m_{j^*}}}{\sqrt{2}c_s\varepsilon} \right) \right) \right) \\
& \leq \frac{15}{8} \rho_m(\varepsilon; f)^2 + \frac{3}{2} \mathbb{E}_l \left( \sum_{j_1=2}^{j^*-1} 2^{4+j_1-j^*} \frac{2c_s^2\varepsilon^2}{m_{j^*}} \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \right. \\
& \quad \left[ 2^{\frac{j^*-j_1-4}{2}} \frac{\sqrt{m_{j^*}}}{\sqrt{2}c_s\varepsilon} (\mu_{j_1, \hat{i}_{j_1}+2} - M(f)) \right]^2 \Phi \left( 2 - (\mu_{j_1, \hat{i}_{j_1}+2} - M(f)) 2^{\frac{j^*-j_1-4}{2}} \frac{\sqrt{m_{j^*}}}{\sqrt{2}c_s\varepsilon} \right) \left. \right) \\
& \quad + \frac{3}{2} \mathbb{E}_l \left( \sum_{j_1=2}^{j^*+1} 2^{4+j_1-j^*} \frac{2c_s^2\varepsilon^2}{m_{j^*}} \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \right. \\
& \quad \left[ 2^{\frac{j^*-j_1-4}{2}} \frac{\sqrt{m_{j^*}}}{2\sqrt{2}c_s\varepsilon} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f)) \right]^2 \Phi \left( 2 - (\mu_{j_1, \hat{i}_{j_1}-2} - M(f)) 2^{\frac{j^*-j_1-4}{2}} \frac{\sqrt{m_{j^*}}}{2\sqrt{2}c_s\varepsilon} \right) \left. \right) \\
& \leq 3 \sum_{j_1=2}^{j^*-1} 2^{4+j_1-j^*} \frac{2c_s^2\varepsilon^2}{m_{j^*}} V + \frac{1}{2} (3 \times (2^5 + 2^6) \frac{c_s^2\varepsilon^2 V}{m_{j^*}}) + \frac{15}{8} \rho_m(\varepsilon; f)^2 \\
& \leq (96 \times 3 \times 8 \times V + 2^7 \times 3 \times 9 \times V + 2) \times \rho_m(\varepsilon; f)^2 \leq (5760V + 2) \rho_m(\varepsilon; f)^2,
\end{aligned} \tag{A.2.34}$$

where  $V = \sup_{x \geq 0} x^2 \Phi(2 - x)$ .

Now let's turn to the second term of Inequality (A.2.30).

$$\begin{aligned}
& \mathbb{E}_{l,s} \left( \sum_{j_1=j^*+2}^{\infty} (\hat{f} - M(f))^2 \mathbb{1}\{\hat{j} = j_1, \tilde{j} \geq j_1 + 1\} \right) \\
& \leq \mathbb{E}_{l,s} \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}+2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \right. \\
& \quad \left. \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} > 2\} \right) \\
& + \mathbb{E}_{l,s} \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \right. \\
& \quad \left. \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} > 2\} \right) \\
& + \mathbb{E}_{l,s} \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \right. \\
& \quad \left. \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} \right) \\
& \leq \frac{1}{16} \rho_m(\varepsilon; f)^2 + \mathbb{E}_{l,s} \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \right. \\
& \quad \left. \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2, \forall j^* + 1 \leq j \leq j_1, \frac{\tilde{X}_{j, \hat{i}_j+6} - \tilde{X}_{j, \hat{i}_j+5}}{\sqrt{2}c_s\sqrt{m_j}\varepsilon} > 2\} \right) \\
& + \mathbb{E}_{l,s} \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1, \hat{j} = j_1, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2, \right. \\
& \quad \left. \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2, \frac{\tilde{X}_{j, \hat{i}_j+6} - \tilde{X}_{j, \hat{i}_j+5}}{\sqrt{2}c_s\sqrt{m_j}\varepsilon} > 2, \forall j^* + 1 \leq j \leq j_1 - 1\} \right) \\
& \leq \mathbb{E}_l \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \mathbb{E}_s(\mathbb{1}\{\forall j^* + 1 \leq j \leq j_1, \right. \\
& \quad \left. \frac{\tilde{X}_{j, \hat{i}_j+6} - \tilde{X}_{j, \hat{i}_j+5}}{\sqrt{2}c_s\sqrt{m_j}\varepsilon} > 2, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} | Y_l) \right) + \frac{1}{16} \rho_m(\varepsilon; f)^2 \\
& + \mathbb{E}_l \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \mathbb{E}_s(\mathbb{1}\{\frac{\tilde{X}_{j_1, \hat{i}_{j_1}-6} - \tilde{X}_{j_1, \hat{i}_{j_1}-5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2, \right. \\
& \quad \left. \forall j^* + 1 \leq j \leq j_1 - 1, \frac{\tilde{X}_{j, \hat{i}_j+6} - \tilde{X}_{j, \hat{i}_j+5}}{\sqrt{2}c_s\sqrt{m_j}\varepsilon} > 2, \frac{\tilde{X}_{j_1, \hat{i}_{j_1}+6} - \tilde{X}_{j_1, \hat{i}_{j_1}+5}}{\sqrt{2}c_s\sqrt{m_{j_1}}\varepsilon} \leq 2\} | Y_l) \right)
\end{aligned} \tag{A.2.35}$$

$$\begin{aligned}
&\leq \frac{1}{16} \rho_m(\varepsilon; f)^2 + \mathbb{E}_l \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \right. \\
&\quad \left. [\Pi_{j=j^*+1}^{j_1} \Phi(-2 + \frac{(\mu_{j, \hat{i}_j+6} - \mu_{j, \hat{i}_j+5}) \sqrt{m_j}}{\sqrt{2} c_s \varepsilon})] \Phi(2 - \frac{\sqrt{m_{j_1}} (\mu_{j_1, \hat{i}_{j_1}-6} - \mu_{j_1, \hat{i}_{j_1}-5})}{\sqrt{2} c_s \varepsilon}) \right) \\
&\quad + \mathbb{E}_l \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \Phi(2 - \frac{\sqrt{m_{j_1}} (\mu_{j_1, \hat{i}_{j_1}-6} - \mu_{j_1, \hat{i}_{j_1}-5})}{\sqrt{2} c_s \varepsilon}) \right. \\
&\quad \left. \Phi(2 - \frac{\sqrt{m_{j_1}} (\mu_{j_1, \hat{i}_{j_1}+6} - \mu_{j_1, \hat{i}_{j_1}+5})}{\sqrt{2} c_s \varepsilon}) [\Pi_{j=j^*+1}^{j_1-1} \Phi(-2 + \frac{(\mu_{j, \hat{i}_j+6} - \mu_{j, \hat{i}_j+5}) \sqrt{m_j}}{\sqrt{2} c_s \varepsilon})] \right) \\
&\leq \frac{1}{16} \rho_m(\varepsilon; f)^2 + \mathbb{E}_l \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}-2} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \right. \\
&\quad \left. [\Pi_{j=j^*+1}^{j_1} \Phi(-2 + \frac{\frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} 8 m_j \sqrt{m_j}}{\sqrt{2} c_s \varepsilon})] \Phi(2 - \frac{\sqrt{m_{j_1}} \frac{\mu_{j_1, \hat{i}_{j_1}-2} - M(f)}{4}}{\sqrt{2} c_s \varepsilon}) \right) \\
&\quad + \mathbb{E}_l \left( \sum_{j_1=j^*+2}^{\infty} (\mu_{j_1, \hat{i}_{j_1}} - M(f))^2 \mathbb{1}\{\tilde{j} \geq j_1 + 1\} \Phi(2 - \frac{\sqrt{m_{j_1}} \frac{\mu_{j_1, \hat{i}_{j_1}} - M(f)}{2}}{\sqrt{2} c_s \varepsilon}) \right. \\
&\quad \left. [\Pi_{j=j^*+1}^{j_1-1} \Phi(-2 + \frac{\frac{\rho_m(\varepsilon; f)}{\rho_z(\varepsilon; f)} 8 m_j \sqrt{m_j}}{\sqrt{2} c_s \varepsilon})] \right) \\
&\leq \frac{1}{16} \rho_m(\varepsilon; f)^2 + \mathbb{E}_l \left( \sum_{j_1=j^*+2}^{\infty} \mathbb{1}\{\tilde{j} \geq j_1 + 1\} V \frac{32 c_s^2 \varepsilon^2}{m_{j_1}} \Phi(-1.75)^{j_1-j^*} \right) \\
&\quad + \mathbb{E}_l \left( \sum_{j_1=j^*+2}^{\infty} \mathbb{1}\{\tilde{j} \geq j_1 + 1\} V \frac{8 c_s^2 \varepsilon^2}{m_{j_1}} \Phi(-1.75)^{j_1-j^*-1} \right) \\
&\leq \frac{1}{16} \rho_m(\varepsilon; f)^2 + \mathbb{E}_l \left( \sum_{j_1=j^*+2}^{\infty} \mathbb{1}\{\tilde{j} \geq j_1 + 1\} V \times 32 \times 3 \times 8 \frac{\varepsilon^2}{\rho_z(\varepsilon; f)} \times 2^{j_1-j^*} \Phi(-1.75)^{j_1-j^*} \right) \\
&\quad + \mathbb{E}_l \left( \sum_{j_1=j^*+2}^{\infty} \mathbb{1}\{\tilde{j} \geq j_1 + 1\} V \times 24 \times 8 \frac{\varepsilon^2}{\rho_z(\varepsilon; f)} \times 2^{j_1-j^*} \Phi(-1.75)^{j_1-j^*-1} \right) \\
&< \frac{1}{16} \rho_m(\varepsilon; f)^2 + \rho_m(\varepsilon; f)^2 V \times 78.
\end{aligned}$$

Combining the two together, we get

$$\mathbb{E}_{l,s}((\hat{f} - M(f))^2 \mathbb{1}\{\tilde{j} \geq \hat{j} + 1\}) < ((5760V + 2) + 78V + \frac{1}{16}) \rho_m(\varepsilon; f)^2. \quad (\text{A.2.36})$$

□

*Proof of Lemma A.1.13.*

$$\begin{aligned}
& \mathbb{E}_{l,s} \left( ((\hat{f} - \mu_{\hat{j}, \hat{i}_{\hat{j}}})_+)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\} \right) = \mathbb{E}_{l,s} \left( (\hat{f} - \mu_{\hat{j}, \hat{i}_{\hat{j}}})^2 \mathbb{1}\{\tilde{j} \leq \hat{j}, \hat{f} > \mu_{\hat{j}, \hat{i}_{\hat{j}}}\} \right) \\
& \leq \mathbb{E}_{l,s} \left( (\mu_{\hat{j}, \hat{i}_{\hat{j}}+2} - \mu_{\hat{j}, \hat{i}_{\hat{j}}})^2 \mathbb{1}\{\tilde{j} \leq \hat{j}, \mu_{\hat{j}, \hat{i}_{\hat{j}}+2} > \mu_{\hat{j}, \hat{i}_{\hat{j}}}, \frac{\tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}+6} - \tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}+5}}{\sqrt{2}\sqrt{m_{\hat{j}}}c_s\varepsilon} \leq 2, \right. \\
& \quad \left. \frac{\tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}-6} - \tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}-5}}{\sqrt{2}\sqrt{m_{\hat{j}}}c_s\varepsilon} > 2 \text{ if } \hat{i}_{\hat{j}-6} \geq 1\} \right) \\
& \quad + \mathbb{E}_{l,s} \left( (\mu_{\hat{j}, \hat{i}_{\hat{j}}-2} - \mu_{\hat{j}, \hat{i}_{\hat{j}}})^2 \mathbb{1}\{\tilde{j} \leq \hat{j}, \mu_{\hat{j}, \hat{i}_{\hat{j}}-2} > \mu_{\hat{j}, \hat{i}_{\hat{j}}}, \frac{\tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}-6} - \tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}-5}}{\sqrt{2}\sqrt{m_{\hat{j}}}c_s\varepsilon} \leq 2, \right. \\
& \quad \left. \frac{\tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}+6} - \tilde{X}_{\hat{j}, \hat{i}_{\hat{j}}+5}}{\sqrt{2}\sqrt{m_{\hat{j}}}c_s\varepsilon} > 2 \text{ if } \hat{i}_{\hat{j}+6} \leq 1\} \right) \\
& \leq \sum_{j_1=2}^{\infty} \sum_{j_2=j_1}^{\infty} \left( \mathbb{E}_{l,s} \left( (\mu_{j_2, \hat{i}_{j_2}+2} - \mu_{j_2, \hat{i}_{j_2}})^2 \mathbb{1}\{\tilde{j} = j_1, \hat{j} = j_2, \frac{\tilde{X}_{j_2, \hat{i}_{j_2}+6} - \tilde{X}_{j_2, \hat{i}_{j_2}+5}}{\sqrt{2}\sqrt{m_{j_2}}}c_s\varepsilon \leq 2, \right. \right. \\
& \quad \left. \forall j^* + 2 \leq j \leq j_2 - 1, \frac{\tilde{X}_{j, \hat{i}_j+6} - \tilde{X}_{j, \hat{i}_j+5}}{\sqrt{2}\sqrt{m_j}c_s\varepsilon} > 2, \frac{\tilde{X}_{j, \hat{i}_j-6} - \tilde{X}_{j, \hat{i}_j-5}}{\sqrt{2}\sqrt{m_j}c_s\varepsilon} > 2, \mu_{j_2, \hat{i}_{j_2}+2} > \mu_{j_2, \hat{i}_{j_2}} \} \right) \\
& \quad + \mathbb{E}_{l,s} \left( (\mu_{j_2, \hat{i}_{j_2}-2} - \mu_{j_2, \hat{i}_{j_2}})^2 \mathbb{1}\{\tilde{j} = j_1, \hat{j} = j_2, \frac{\tilde{X}_{j_2, \hat{i}_{j_2}-6} - \tilde{X}_{j_2, \hat{i}_{j_2}-5}}{\sqrt{2}\sqrt{m_{j_2}}}c_s\varepsilon \leq 2, \right. \\
& \quad \left. \forall j^* + 2 \leq j \leq j_2 - 1, \frac{\tilde{X}_{j, \hat{i}_j+6} - \tilde{X}_{j, \hat{i}_j+5}}{\sqrt{2}\sqrt{m_j}c_s\varepsilon} > 2, \frac{\tilde{X}_{j, \hat{i}_j-6} - \tilde{X}_{j, \hat{i}_j-5}}{\sqrt{2}\sqrt{m_j}c_s\varepsilon} > 2, \mu_{j_2, \hat{i}_{j_2}-2} > \mu_{j_2, \hat{i}_{j_2}} \} \right) \Bigg) \\
& \leq \sum_{j_1=2}^{\infty} \sum_{j_2=j_1}^{\infty} \left( \mathbb{E}_l \left( (\mu_{j_2, \hat{i}_{j_2}+2} - \mu_{j_2, \hat{i}_{j_2}})^2 \mathbb{1}\{\tilde{j} = j_1, \mu_{j_2, \hat{i}_{j_2}+2} > \mu_{j_2, \hat{i}_{j_2}}\} \right. \right. \\
& \quad \left. \Phi \left( 2 - \frac{\mu_{j_2, \hat{i}_{j_2}+2} - \mu_{j_2, \hat{i}_{j_2}}}{2} \frac{\sqrt{m_{j_2}}}{\sqrt{2}c_s\varepsilon} \right) \right. \\
& \quad \left. \Pi_{j=j^*+2}^{j_2-1} \max \{ \Phi(-2), \Phi(-2 + (\frac{7}{16} + \frac{6m_j}{\rho_z(\varepsilon; f)})\rho_m(\varepsilon; f) \frac{\sqrt{m_j}}{\sqrt{2}c_s\varepsilon}) \} \right) \\
& \quad + \mathbb{E}_l \left( (\mu_{j_2, \hat{i}_{j_2}-2} - \mu_{j_2, \hat{i}_{j_2}})^2 \mathbb{1}\{\tilde{j} = j_1, \mu_{j_2, \hat{i}_{j_2}-2} > \mu_{j_2, \hat{i}_{j_2}}\} \Phi \left( 2 - \frac{\mu_{j_2, \hat{i}_{j_2}-2} - \mu_{j_2, \hat{i}_{j_2}}}{2} \frac{\sqrt{m_{j_2}}}{\sqrt{2}c_s\varepsilon} \right) \right. \\
& \quad \left. \left. \Pi_{j=j^*+2}^{j_2-1} \max \{ \Phi(-2), \Phi(-2 + (\frac{7}{16} + \frac{6m_j}{\rho_z(\varepsilon; f)})\rho_m(\varepsilon; f) \frac{\sqrt{m_j}}{\sqrt{2}c_s\varepsilon}) \} \right) \right). \tag{A.2.37}
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{j_1=2}^{\infty} \sum_{j_2=j_1}^{\infty} 2 \times \mathbb{E}_l \left( \mathbb{1}\{\tilde{j} = j_1\} \frac{8c_s^2 \varepsilon^2}{m_{j_2}} V \Phi(-1.85)^{(j_2-j^*-2)+} \right) \\
&\leq \sum_{j_1=2}^{\infty} \sum_{j_2=j_1}^{\infty} 2 \times \mathbb{E}_l \left( \mathbb{1}\{\tilde{j} = j_1\} \times 8c_s^2 \times 2^{j_2-j^*+4} \rho_m(\varepsilon; f)^2 V \Phi(-1.85)^{(j_2-j^*-2)+} \right) \\
&\leq \sum_{j_1=2}^{\infty} \sum_{j_2=j_1}^{\infty} \mathbb{E}_l \left( \mathbb{1}\{\tilde{j} = j_1\} \right) \times 2^{10} \times 3 \times \rho_m(\varepsilon; f)^2 V 2^{j_2-j^*-2} \Phi(-1.85)^{(j_2-j^*-2)+} \\
&\leq \sum_{j_1=2}^{\infty} \mathbb{E}_l \left( \mathbb{1}\{\tilde{j} = j_1\} \right) \times 2^{10} \times 3 \times \rho_m(\varepsilon; f)^2 V (2 \times \mathbb{1}\{j_1 \leq j^* + 2\} + \frac{2\Phi(-1.85)}{1 - 2\Phi(-1.85)}) \\
&\leq 2^{11} \times 3 \rho_m(\varepsilon; f)^2 V \times P(\tilde{j} \leq j^* + 2) + 2^{11} \times 3 \times \Phi(-1.85) \frac{\rho_m(\varepsilon; f)^2 V}{1 - 2\Phi(-1.85)} \\
&\leq 6355.2 V \rho_m(\varepsilon; f)^2
\end{aligned}$$

□

*Proof of Lemma A.1.14.* For simplicity, we denote the set for possible  $\hat{i}_{j_2}$  when  $\tilde{j} = j_2$  to be  $Op(j_2) = \{i_{j_2}^* - 4, i_{j_2}^* - 3, i_{j_2}^* - 2, i_{j_2}^* + 2, i_{j_2}^* + 3, i_{j_2}^* + 4\}$ . By the definition of  $\tilde{j}$ , it's easy to verify that  $\hat{i}_{\tilde{j}} \in Op(\tilde{j})$ . Further, for the convenience of notation, we define  $Ind(j, i) = sign(i - i_j^*)$ . Without loss of generality, we assume

$$sup\{t > Z(f) : f(t) \leq \rho_m(\varepsilon; f) + M(f)\} = \rho_z(\varepsilon; f) + Z(f).$$

$$\begin{aligned}
&\mathbb{E}_{l,s}((\mu_{\tilde{j}, \hat{i}_{\tilde{j}}} - M(f))^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\}) \\
&= \sum_{j_2=2}^{\infty} \sum_{j_1=j_2}^{\infty} \mathbb{E}_{l,s}((\mu_{j_2, \hat{i}_{j_2}} - M(f))^2 \mathbb{1}\{\tilde{j} = j_2, \hat{j} = j_1\}) \\
&= \sum_{j_2=2}^{\infty} \sum_{j_1=j_2}^{\infty} \sum_{i \in Op(j_2)} \mathbb{E}_{l,s}((\mu_{j_2, i} - M(f))^2 \mathbb{1}\{\tilde{j} = j_2, \hat{j} = j_1, \hat{i}_{j_2} = i\}) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j_1=j_2}^{\infty} \sum_{i \in Op(j_2)} \mathbb{E}_l \left( (\mu_{j_2, i} - M(f))^2 \mathbb{1}\{\tilde{j} = j_2, \hat{i}_{j_2} = i\} \mathbb{E}_s(\mathbb{1}\{\hat{j} = j_1\} | Y_l) \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j_1=j_2}^{\infty} \sum_{i \in Op(j_2)} \mathbb{E}_l \left( (\mu_{j_2, i} - M(f))^2 \mathbb{1}\{\tilde{j} = j_2, \hat{i}_{j_2} = i\} \left( \mathbb{E}_s(\mathbb{1}\{\hat{j} = j_1\} | Y_l) \mathbb{1}\{j_1 \leq j^* + 2\} \right. \right. \\
&\quad \left. \left. + \mathbb{1}\{j_1 \geq j^* + 3\} \Pi_{j=j^*+2}^{j_1-1} \max\{\Phi(-2), \Phi(-2 + (\frac{7}{16} + \frac{6m_j}{\rho_z(\varepsilon; f)}) \rho_m(\varepsilon; f) \frac{\sqrt{m_j}}{\sqrt{2}c_s\varepsilon})\} \right) \right)
\end{aligned} \tag{A.2.38}$$

$$\begin{aligned}
&\leq \sum_{j_2=2}^{\infty} \sum_{i \in Op(j_2)} \mathbb{E}_l \left( (\mu_{j_2,i} - M(f))^2 \mathbb{1}\{\tilde{j} = j_2, \hat{i}_{j_2} = i\} (\mathbb{1}\{j_2 \leq j^* + 2\} + \right. \\
&\quad \left. \mathbb{1}\{j_2 \geq j^* + 3\} \Phi(-1.85) \frac{\Phi(-2 + \frac{1}{12})^{j_2-j^*-3}}{1 - \Phi(-2 + \frac{1}{12})}) \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{i \in Op(j_2)} \mathbb{E}_l \left( (\mu_{j_2,i} - M(f))^2 \mathbb{1}\{X_{j_2,i} \leq X_{j_2, i_{j_2}^* + Ind(j_2,i)}\} (\mathbb{1}\{j_2 \leq j^* + 2\} + \right. \\
&\quad \left. \mathbb{1}\{j_2 \geq j^* + 3\} \Phi(-1.85) \frac{\Phi(-2 + \frac{1}{12})^{j_2-j^*-3}}{1 - \Phi(-2 + \frac{1}{12})}) \right) \\
&= \sum_{j_2=2}^{\infty} (\mathbb{1}\{j_2 \leq j^* + 2\} + \mathbb{1}\{j_2 \geq j^* + 3\} \Phi(-1.85) \frac{\Phi(-2 + \frac{1}{12})^{j_2-j^*-3}}{1 - \Phi(-2 + \frac{1}{12})}) \\
&\quad \sum_{i \in Op(j_2)} (\mu_{j_2,i} - M(f))^2 \Phi\left(\frac{\mu_{j_2, i_{j_2}^* + Ind(j_2,i)} - \mu_{j_2,i}}{\sqrt{2}c_l \varepsilon} \sqrt{m_{j_2}}\right) \\
&\leq \sum_{j_2=2}^{\infty} (\mathbb{1}\{j_2 \leq j^* + 2\} + \mathbb{1}\{j_2 \geq j^* + 3\} \Phi(-1.85) \frac{\Phi(-2 + \frac{1}{12})^{j_2-j^*-3}}{1 - \Phi(-2 + \frac{1}{12})}) \\
&\quad \sum_{i \in Op(j_2)} (\mu_{j_2,i} - M(f))^2 \Phi\left(-(\mu_{j_2,i} - M(f)) \frac{|i - i_{j_2}^*| - 1}{|i - i_{j_2}^*| + \frac{1}{2}} \frac{\sqrt{m_{j_2}}}{\sqrt{2}c_l \varepsilon}\right) \\
&\leq \sum_{j_2=2}^{\infty} (\mathbb{1}\{j_2 \leq j^* + 2\} + \mathbb{1}\{j_2 \geq j^* + 3\} \Phi(-1.85) \frac{\Phi(-2 + \frac{1}{12})^{j_2-j^*-3}}{1 - \Phi(-2 + \frac{1}{12})}) \\
&\quad \sum_{i \in Op(j_2)} \frac{2c_l^2 \varepsilon^2}{m_{j_2}} \left(\frac{|i - i_{j_2}^*| + \frac{1}{2}}{|i - i_{j_2}^*| - 1}\right)^2 Q \\
&< \sum_{j_2=2}^{\infty} (\mathbb{1}\{j_2 \leq j^* + 2\} + \mathbb{1}\{j_2 \geq j^* + 3\} \Phi(-1.85) \frac{\Phi(-2 + \frac{1}{12})^{j_2-j^*-3}}{1 - \Phi(-2 + \frac{1}{12})}) * \\
&\quad 3 \times 2^{4+j_2-j^*} \rho_m(\varepsilon; f)^2 (23\frac{1}{8}) Q \times 2 \\
&< 3 \times (2^8 + 2^8 \frac{\Phi(-1.85)}{(1 - \Phi(-2 + \frac{1}{12}))^2}) \rho_m(\varepsilon; f)^2 (23\frac{1}{8}) Q,
\end{aligned}$$

where

$$Q = \sup_{x \geq 0} x^2 \Phi(-x).$$

The reason for the fourth to last inequality is as follows. Without loss of generality, we can assume  $i \geq i_{j_2}^* + 2$ . Then  $\frac{\mu_{j_2,i} - \mu_{j_2, i_{j_2}^* + 1}}{\mu_{j_2,i} - M(f)} \geq 1$ ,  $\frac{f(t_{j_2, i - \frac{1}{2}}) - \mu_{j_2, i_{j_2}^* + 1}}{f(t_{j_2, i - \frac{1}{2}}) - M(f)} \geq 1$ . Since we also have



$\mu_{j_2,i} \geq f(t_{j_2,i} - \frac{1}{2})$ , we have

$$\begin{aligned} \frac{\mu_{j_2,i} - \mu_{j_2,i_{j_2}^*+1}}{\mu_{j_2,i} - M(f)} &\geq \frac{f(t_{j_2,i} - \frac{1}{2}) - \mu_{j_2,i_{j_2}^*+1}}{f(t_{j_2,i} - \frac{1}{2}) - M(f)} \geq \int_{0,1} \frac{t_{j_2,i} - \frac{1}{2} - t_{j_2,i_{j_2}^*} - x}{t_{j_2,i} - \frac{1}{2} - Z(f)} dx \\ &= \frac{|i - i_{j_2}^*| - 1}{t_{j_2,i} - \frac{1}{2} - Z(f)} \geq \frac{|i - i_{j_2}^*| - 1}{|i - i_{j_2}^*| + \frac{1}{2}}. \end{aligned} \quad (\text{A.2.39})$$

□

*Proof of Lemma A.1.15.* First, with a bit of abuse of notation, define the events  $A_r$ ,  $B_r$ ,  $C_r$ ,  $D_r$  to be the following (they only mean events but not constants in this proof):

$$\begin{aligned} A_r &= \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, t_{\tilde{j}+r+1, \hat{i}_{\tilde{j}+r+1}} = t_{\tilde{j}+r, \hat{i}_{\tilde{j}+r}} - m_{\tilde{j}+r}\} \\ &\quad \cup \{\omega : \hat{i}_{\tilde{j}+r} > i_{\tilde{j}+r}^*, t_{\tilde{j}+r+1, \hat{i}_{\tilde{j}+r+1}} = t_{\tilde{j}+r, \hat{i}_{\tilde{j}+r}} + m_{\tilde{j}+r+1}\} \\ B_r &= \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, t_{\tilde{j}+r+1, \hat{i}_{\tilde{j}+r+1}} = t_{\tilde{j}+r, \hat{i}_{\tilde{j}+r}} - m_{\tilde{j}+r+1}\} \\ &\quad \cup \{\omega : \hat{i}_{\tilde{j}+r} > i_{\tilde{j}+r}^*, t_{\tilde{j}+r+1, \hat{i}_{\tilde{j}+r+1}} = t_{\tilde{j}+r, \hat{i}_{\tilde{j}+r}}\} \\ C_r &= \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, t_{\tilde{j}+r+1, \hat{i}_{\tilde{j}+r+1}} = t_{\tilde{j}+r, \hat{i}_{\tilde{j}+r}}\} \\ &\quad \cup \{\omega : \hat{i}_{\tilde{j}+r} > i_{\tilde{j}+r}^*, t_{\tilde{j}+r+1, \hat{i}_{\tilde{j}+r+1}} = t_{\tilde{j}+r, \hat{i}_{\tilde{j}+r}} - m_{\tilde{j}+r+1}\} \\ D_r &= \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, t_{\tilde{j}+r+1, \hat{i}_{\tilde{j}+r+1}} = t_{\tilde{j}+r, \hat{i}_{\tilde{j}+r}} + m_{\tilde{j}+r+1}\} \\ &\quad \cup \{\omega : \hat{i}_{\tilde{j}+r} > i_{\tilde{j}+r}^*, t_{\tilde{j}+r+1, \hat{i}_{\tilde{j}+r+1}} = t_{\tilde{j}+r, \hat{i}_{\tilde{j}+r}} - m_{\tilde{j}+r}\} \end{aligned} \quad (\text{A.2.40})$$

Basically, these events indicates which interval the localization procedure picks at the step  $\tilde{j}+r+1$ , and from the highest average to the lowest average is A to D. These sets of notation for events are only used in this proof, and in the proof of other theorem, the same notation can denote different things.

Still, without loss of generality, we assume

$$\sup\{t > Z(f) : f(t) \leq \rho_m(\varepsilon; f) + M(f)\} = \rho_z(\varepsilon; f) + Z(f).$$

$$\begin{aligned}
& \mathbb{E}_{l,s} \left( ((\mu_{\hat{j}, \hat{i}_{\hat{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}} )_+)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\} \right) \\
&= \mathbb{E}_{l,s} \left( ((\mu_{\hat{j}, \hat{i}_{\hat{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}} )_+)^2 \mathbb{1}\{\tilde{j} \leq \hat{j} - 1\} \right) \\
&= \mathbb{E}_{l,s} \left( ((\mu_{\hat{j}, \hat{i}_{\hat{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}} )_+)^2 \mathbb{1}\{\tilde{j} \leq \hat{j} - 1, A_0 \cup B_0 \cup (C_0 \cap (A_1 \cup B_1))\} \right) \\
&= \mathbb{E}_{l,s} \left( ((\mu_{\hat{j}, \hat{i}_{\hat{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}} )_+)^2 \mathbb{1}\{\tilde{j} \leq \hat{j} - 1, A_0 \cup (B_0 \cap D_1^c) \cup (B_0 \cap D_1 \cap \{\hat{j} = \tilde{j} + 1\})\} \right) \\
&\quad + \mathbb{E}_{l,s} \left( ((\mu_{\hat{j}, \hat{i}_{\hat{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}} )_+)^2 \mathbb{1}\{\tilde{j} \leq \hat{j} - 2, C_0 \cap A_1\} \right) \\
&\quad + \mathbb{E}_{l,s} \left( ((\mu_{\hat{j}, \hat{i}_{\hat{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}} )_+)^2 \mathbb{1}\{\tilde{j} \leq \hat{j} - 3, (C_0 \cap B_1) \cup (B_0 \cap D_1)\} \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j_1=j_2+1}^{\infty} \mathbb{E}_{l,s} \left( \left( \sum_{j=j_2}^{j_1-1} (\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})_+ \right)^2 \right. \\
&\quad \left. \mathbb{1}\{\hat{j} = j_1, \tilde{j} = j_2, A_0 \cup (B_0 \cap D_1^c) \cup (B_0 \cap D_1 \cap \{j_1 = j_2 + 1\})\} \right) \\
&\quad + \sum_{j_2=2}^{\infty} \sum_{j_1=j_2+2}^{\infty} \mathbb{E}_{l,s} \left( \left( \sum_{j=j_2}^{j_1-1} (\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})_+ \right)^2 \mathbb{1}\{\hat{j} = j_1, \tilde{j} = j_2, C_0 \cap A_1\} \right) \\
&\quad + \mathbb{E}_{l,s} \left( ((\mu_{\hat{j}, \hat{i}_{\hat{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}} )_+)^2 \mathbb{1}\{\hat{j} \geq \tilde{j} + 3, (C_0 \cap B_1) \cup (B_0 \cap D_1)\} \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j_1=j_2+1}^{\infty} \mathbb{E}_{l,s} \left( 2 \sum_{j=j_2}^{j_1-1} 2^{j-j_2} ((\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})_+)^2 \right. \\
&\quad \left. \mathbb{1}\{\hat{j} = j_1, \tilde{j} = j_2, A_0 \cup (B_0 \cap D_1^c) \cup (B_0 \cap D_1 \cap \{j_1 = j_2 + 1\})\} \right) \\
&\quad + \sum_{j_2=2}^{\infty} \sum_{j_1=j_2+2}^{\infty} \mathbb{E}_{l,s} \left( 2 \sum_{j=j_2+1}^{j_1-1} 2^{j-j_2-1} ((\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})_+)^2 \mathbb{1}\{\hat{j} = j_1, \tilde{j} = j_2, C_0 \cap A_1\} \right) \\
&\quad + \mathbb{E}_{l,s} \left( ((\mu_{\hat{j}, \hat{i}_{\hat{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}} )_+)^2 \mathbb{1}\{\hat{j} \geq \tilde{j} + 3, (C_0 \cap B_1) \cup (B_0 \cap D_1)\} \right)
\end{aligned} \tag{A.2.41}$$

Now we will bound the sum of first two terms in Inequality (A.2.41) first. For the simplicity

of formula's expression, define  $\delta_0 = \mathbb{1}\{j_1 = j_2 + 1\}$ ,  $\delta = \mathbb{1}\{j = j_2\}$ , which will only be used for the inequalities below.

$$\begin{aligned}
&= \sum_{j_2=2}^{\infty} \sum_{j=j_2}^{\infty} 2^{j+1-j_2} \sum_{j_1=j+1}^{\infty} \mathbb{E}_{l,s} \left( (\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})^2 \mathbb{1}\{\mu_{j+1, \hat{i}_{j+1}} > \mu_{j, \hat{i}_j}\} \mathbb{1}\{\hat{j} = j_1\} \right. \\
&\quad \left. \left( \mathbb{1}\{\tilde{j} = j_2, A_0 \cup (B_0 \cap D_1^c)\} + \mathbb{1}\{\tilde{j} = j_2, j_1 = j_2 + 1, j = j_2, B_0 \cap D_1\} \right) \right) \\
&\quad + \sum_{j_2=2}^{\infty} \sum_{j=j_2+1}^{\infty} 2^{j-j_2} \sum_{j_1=j+1}^{\infty} \mathbb{E}_{l,s} \left( (\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})^2 \mathbb{1}\{\mu_{j+1, \hat{i}_{j+1}} > \mu_{j, \hat{i}_j}\} \right. \\
&\quad \left. \mathbb{1}\{\tilde{j} = j_2, C_0 \cap A_1\} \mathbb{1}\{\hat{j} = j_1\} \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j=j_2}^{\infty} 2^{j+1-j_2} \mathbb{E}_l \left( (\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})^2 \mathbb{1}\{\mu_{j+1, \hat{i}_{j+1}} > \mu_{j, \hat{i}_j}\} \right. \\
&\quad \sum_{j_1=j+1}^{\infty} \Phi(-1.85)^{(j_2-j^*-\delta_0)+} \Phi(-2)^{(j_1-j_2-2)+} \\
&\quad \left. \left( \mathbb{1}\{\tilde{j} = j_2, A_0 \cup (B_0 \cap D_1^c)\} + \mathbb{1}\{\tilde{j} = j_2, B_0 \cap D_1, j_1 = j_2 + 1, j = j_2\} \right) \right) \\
&\quad + \sum_{j_2=2}^{\infty} \sum_{j=j_2+1}^{\infty} 2^{j-j_2} \mathbb{E}_l \left( (\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})^2 \mathbb{1}\{\mu_{j+1, \hat{i}_{j+1}} > \mu_{j, \hat{i}_j}\} \mathbb{1}\{\tilde{j} = j_2, C_0 \cap A_1\} \right. \\
&\quad \left. \sum_{j_1=j+1}^{\infty} \Phi(-1.85)^{(j_2-j^*)+} \Phi(-2)^{(j_1-j_2-2)+} \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j=j_2}^{\infty} \Phi(-1.85)^{(j_2-j^*-\delta)+} 2^{j+1-j_2} \mathbb{E}_l \left( (\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})^2 \mathbb{1}\{\mu_{j+1, \hat{i}_{j+1}} > \mu_{j, \hat{i}_j}\} \mathbb{1}\{\tilde{j} = j_2\} \right) \\
&\quad \left( \mathbb{1}\{j = j_2, A_0 \cup B_0\} \left( 1 + \frac{1}{1 - \Phi(-2)} \right) + \mathbb{1}\{j \geq j_2 + 1, A_0 \cup (B_0 \cap D_1^c)\} \frac{\Phi(-2)^{j-j_2-1}}{1 - \Phi(-2)} \right) \\
&\quad + \sum_{j_2=2}^{\infty} \Phi(-1.85)^{(j_2-j^*)+} \sum_{j=j_2+1}^{\infty} 2^{j-j_2} \mathbb{E}_l \left( (\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})^2 \right. \\
&\quad \left. \mathbb{1}\{\mu_{j+1, \hat{i}_{j+1}} > \mu_{j, \hat{i}_j}\} \mathbb{1}\{\tilde{j} = j_2, C_0 \cap A_1\} \right) \left( \Phi(-2)^{j-j_2-1} \frac{1}{1 - \Phi(-2)} \right)
\end{aligned} \tag{A.2.42}$$

Now define the set  $C(j, k, k+1)$  to be the set of pairs  $(i_1, i_2)$  such that ,  $P(\hat{i}_{k+1} = i_2, \hat{i}_k = i_1 | \tilde{j} = j) > 0$ , then we know that  $|C(j, k, k+1)| \leq \min\{10 \times 2^{k-j} \times 4, 6 \times 4^{k+1-j}\}$ . Then,

continuing with the inequality we have

$$\begin{aligned}
&\leq \sum_{j_2=2}^{\infty} \sum_{j=j_2}^{\infty} \Phi(-1.85)^{(j_2-j^*-\delta)_+} \cdot 2^{j+1-j_2} \sum_{(i_1, i_2) \in C(j_2, j, j+1)} \\
&\quad \mathbb{E}_l \left( (\mu_{j+1, i_2} - \mu_{j, i_1})^2 \mathbb{1}\{\mu_{j+1, i_2} > \mu_{j, i_1}\} \mathbb{1}\{\tilde{j} = j_2, A_0 \cup B_0, \hat{i}_j = i_1, \hat{i}_{j+1} = i_2\} \right) \\
&\quad \left( \mathbb{1}\{j = j_2\} \left(1 + \frac{1}{1 - \Phi(-2)}\right) + \mathbb{1}\{j \geq j_2 + 1\} \frac{\Phi(-2)^{j-j_2-1}}{1 - \Phi(-2)} \right) \\
&+ \sum_{j_2=2}^{\infty} \Phi(-1.85)^{(j_2-j^*)_+} \sum_{j=j_2+1}^{\infty} 2^{j-j_2} \sum_{(i_1, i_2) \in C(j_2, j, j+1)} \\
&\quad \mathbb{E}_l \left( (\mu_{j+1, i_2} - \mu_{j, i_1})^2 \mathbb{1}\{\mu_{j+1, i_2} > \mu_{j, i_1}\} \mathbb{1}\{\tilde{j} = j_2, \hat{i}_{j+1} = i_2, \hat{i}_j = i_1, C_0 \cap A_1\} \right) \\
&\quad \left( \Phi(-2)^{j-j_2-1} \frac{1}{1 - \Phi(-2)} \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j=j_2}^{\infty} \Phi(-1.85)^{(j_2-j^*-\delta)_+} \cdot 2^{j+1-j_2} \sum_{(i_1, i_2) \in C(j_2, j, j+1)} \frac{2c_l^2 \varepsilon^2}{m_{j+1}} Q \mathbb{1}\{\mu_{j+1, i_2} > \mu_{j, i_1}\} \\
&\quad \left( \mathbb{1}\{j = j_2\} \left(1 + \frac{1}{1 - \Phi(-2)}\right) + \mathbb{1}\{j \geq j_2 + 1\} \frac{\Phi(-2)^{j-j_2-1}}{1 - \Phi(-2)} \right) \\
&+ \sum_{j_2=2}^{\infty} \Phi(-1.85)^{(j_2-j^*)_+} \sum_{j=j_2+1}^{\infty} 2^{j-j_2} \sum_{(i_1, i_2) \in C(j_2, j, j+1)} \frac{2c_l^2 \varepsilon^2}{m_{j+1}} Q \mathbb{1}\{\mu_{j+1, i_2} > \mu_{j, i_1}\} \\
&\quad \left( \Phi(-2)^{j-j_2-1} \frac{1}{1 - \Phi(-2)} \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j=j_2}^{\infty} \Phi(-1.85)^{(j_2-j^*-\delta)_+} \cdot 2^{j+1-j_2} \times \min\{10 \times 2^{j-j_2} \times 2, 6 \times 4^{j-j_2} \times 2\} \frac{2c_l^2 \varepsilon^2}{m_{j+1}} Q \\
&\quad \left( \mathbb{1}\{j = j_2\} \left(1 + \frac{1}{1 - \Phi(-2)}\right) + \mathbb{1}\{j \geq j_2 + 1\} \frac{\Phi(-2)^{j-j_2-1}}{1 - \Phi(-2)} \right) \\
&\quad + \sum_{j_2=2}^{\infty} \Phi(-1.85)^{(j_2-j^*)_+} \sum_{j=j_2+1}^{\infty} 2^{j-j_2} \times \min\{10 \times 2^{j-j_2} \times 2, 6 \times 4^{j-j_2} \times 2\} \frac{2c_l^2 \varepsilon^2}{m_{j+1}} Q \\
&\quad \left( \Phi(-2)^{j-j_2-1} \frac{1}{1 - \Phi(-2)} \right) \\
&= \frac{c_l^2 Q \varepsilon^2}{m_{j^*}} \sum_{j_2=2}^{\infty} 2^{j_2+3-j^*} \times \left(12 \times \left(1 + \frac{1}{1 - \Phi(-2)}\right) \times \Phi(-1.85)^{(j_2-j^*-1)_+} + \right. \\
&\quad \left. \Phi(-1.85)^{(j_2-j^*)_+} \times 160 \times \frac{1}{1 - \Phi(-2)} \times \frac{1}{1 - 8\Phi(-2)} \right) \\
&\quad + \frac{c_l^2 Q \varepsilon^2}{m_{j^*}} \sum_{j_2=2}^{\infty} \Phi(-1.85)^{(j_2-j^*)_+} 2^{7+j_2-j^*} \times 5 \times \frac{1}{1 - \Phi(-2)} \times \frac{1}{1 - 8\Phi(-2)} \\
&< \frac{c_l^2 Q \varepsilon^2}{m_{j^*}} 2790.303 \times \left( \frac{1}{1 - 2\Phi(-1.85)} + 2 - 1 \right) \leq Q \times 277075 \rho_m(\varepsilon; f)^2.
\end{aligned}$$

Now we will turn to the third term in Inequality (A.2.41).

$$\begin{aligned}
& \mathbb{E}_{l,s} \left( ((\mu_{\hat{j}, \hat{i}_{\hat{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}})_+)^2 \mathbb{1}\{\hat{j} \geq \tilde{j} + 3, (C_0 \cap B_1) \cup (B_0 \cap D_1)\} \right) \\
& \leq \sum_{j_2=2}^{\infty} \sum_{j_1=j_2+3}^{\infty} \mathbb{E}_{l,s} \left( \left( \sum_{j=j_2+2}^{j_1-1} (\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})_+ \right)^2 \right. \\
& \quad \left. \mathbb{1}\{\hat{j} = j_1, \tilde{j} = j_2, (C_0 \cap B_1) \cup (B_0 \cap D_1)\} \right) \\
& \leq \sum_{j_2=2}^{\infty} \sum_{j_1=j_2+3}^{\infty} \mathbb{E}_{l,s} \left( 2 \sum_{j=j_2+2}^{j_1-1} 2^{j-j_2-2} ((\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})_+)^2 \right. \\
& \quad \left. \mathbb{1}\{\hat{j} = j_1, \tilde{j} = j_2, (C_0 \cap B_1) \cup (B_0 \cap D_1)\} \right) \\
& \leq \sum_{j_2=2}^{\infty} \mathbb{E}_l \left( 2 \sum_{j=j_2+2}^{\infty} 2^{j-j_2-2} ((\mu_{j+1, \hat{i}_{j+1}} - \mu_{j, \hat{i}_j})_+)^2 \right. \\
& \quad \left. \mathbb{1}\{\tilde{j} = j_2, (C_0 \cap B_1) \cup (B_0 \cap D_1)\} \times \Phi(-1.85)^{(j_2+1-j^*)_+} \frac{\Phi(-2)^{j-j_2-2}}{1-\Phi(-2)} \right) \tag{A.2.43} \\
& \leq \sum_{j_2=2}^{\infty} \sum_{j=j_2+2}^{\infty} 2^{j-j_2-1} (2 \cdot 3 \cdot 2^{j-j_2-2} \cdot 2) \frac{2c_l^2 \varepsilon^2}{m_{j+1}} Q \Phi(-1.85)^{(j_2+1-j^*)_+} \frac{\Phi(-2)^{j-j_2-2}}{1-\Phi(-2)} \\
& = \frac{c_l^2 \varepsilon^2}{m_{j^*}} Q \sum_{j_2=2}^{\infty} \frac{192}{1-\Phi(-2)} \times 2^{j_2+1-j^*} \times \Phi(-1.85)^{(j_2+1-j^*)_+} \frac{1}{1-8\Phi(-2)} \\
& \leq \frac{c_l^2 \varepsilon^2}{m_{j^*}} Q \frac{192}{1-\Phi(-2)} \times \left( \frac{1}{1-2\Phi(-1.85)} + 2 - 1 \right) \frac{1}{1-8\Phi(-2)} \\
& \leq 48Q \times \frac{192}{1-\Phi(-2)} \times \left( \frac{1}{1-2\Phi(-1.85)} + 1 \right) \frac{1}{1-8\Phi(-2)} \rho_m(\varepsilon; f)^2 \\
& \leq 23850.1 \rho_m(\varepsilon; f)^2 Q.
\end{aligned}$$

The fourth inequality is because the number of possible pairs of  $(\hat{i}_j, \hat{i}_{j+1})$  such that  $(C_0 \cap B_1) \cup (B_0 \cap D_1)$ ,  $\mu_{j+1, \hat{i}_{j+1}} > \mu_{j, \hat{i}_j}$ ,  $\tilde{j} = j_2$ ,  $j \geq j_2 + 2$ , and  $\mu_{\hat{j}, \hat{i}_{\hat{j}}} > \mu_{\tilde{j}, \hat{i}_{\tilde{j}}}$  is at most  $2 \times 3 \times 2^{j-(j_2+2)} \times 2$ . Other analysis are similar to the previous one. Combining the two parts together,

$$\begin{aligned}
& \mathbb{E}_{l,s} \left( ((\mu_{\hat{j}, \hat{i}_{\hat{j}}} - \mu_{\tilde{j}, \hat{i}_{\tilde{j}}})_+)^2 \mathbb{1}\{\tilde{j} \leq \hat{j}\} \right) \\
& \leq Q \times 277075 \times \rho_m(\varepsilon; f)^2 + Q \times 23850.1 \times \rho_m(\varepsilon; f)^2.
\end{aligned} \tag{A.2.44}$$

□

*Proof of Lemma A.1.18.*

$$\begin{aligned}
& P(\hat{j} \leq j^* - 2 - \tilde{K}) \\
& \leq P(\hat{j} \leq j^* - 2 - \tilde{K}, |\hat{i}_{\hat{j}} - i_{\hat{j}^*}| \leq 4) + P(\hat{j} \leq j^* - 2 - \tilde{K}, |\hat{i}_{\hat{j}} - i_{\hat{j}}^*| \geq 5) \\
& \leq \sum_{j=1}^{j^*-2-\tilde{K}} P(|\hat{i}_{\hat{j}} - i_{\hat{j}^*}| \leq 4, X_{\hat{i}_{\hat{j}}+6} - X_{\hat{i}_{\hat{j}}+5} \leq 2c_s\sqrt{2\varepsilon}) + \\
& \quad P(|\hat{i}_{\hat{j}} - i_{\hat{j}^*}| \leq 4, X_{\hat{i}_{\hat{j}}-6} - X_{\hat{i}_{\hat{j}}-5} \leq 2c_s\sqrt{2\varepsilon}) + P(|\hat{i}_{j-1} - i_{j-1}^*| \geq 2) \\
& \leq \sum_{j=1}^{j^*-2-\tilde{K}} 2\Phi(2 - (\frac{m_j}{\rho_z(\varepsilon; f)})^{\frac{3}{2}} \frac{\rho_m(\varepsilon; f)\sqrt{\rho_z(\varepsilon; f)}}{\sqrt{2}c_s\varepsilon}) + 2\Phi(-(\frac{m_{j-1}}{\rho_z(\varepsilon; f)})^{\frac{3}{2}} \frac{\rho_m(\varepsilon; f)\sqrt{\rho_z(\varepsilon; f)}}{\sqrt{2}c_s\varepsilon}) \\
& \quad + 2\Phi(-2(\frac{m_{j-1}}{\rho_z(\varepsilon; f)})^{\frac{3}{2}} \frac{\rho_m(\varepsilon; f)\sqrt{\rho_z(\varepsilon; f)}}{\sqrt{2}c_s\varepsilon}) + 2\Phi(-3(\frac{m_{j-1}}{\rho_z(\varepsilon; f)})^{\frac{3}{2}} \frac{\rho_m(\varepsilon; f)\sqrt{\rho_z(\varepsilon; f)}}{\sqrt{2}c_s\varepsilon}) \\
& \tag{A.2.45}
\end{aligned}$$

$$\begin{aligned}
& < 2 \sum_{j=1}^{j^*-2-\tilde{K}} \left( \Phi(2 - 2^{\frac{3}{2}(j^*-j-4)-\frac{1}{2}}) + \Phi(-2^{\frac{3}{2}(j^*-j-3)-\frac{1}{2}}) + \Phi(-2^{\frac{3}{2}(j^*-j-3)+\frac{1}{2}}) \right. \\
& \quad \left. + \Phi(-3 \times 2^{\frac{3}{2}(j^*-j-3)-\frac{1}{2}}) \right) \\
& \leq 2 \sum_{k=\tilde{K}}^{\infty} \left( \Phi(2 - 2^{\frac{3}{2}(k-2)-\frac{1}{2}}) + \Phi(-2^{\frac{3}{2}(k-1)-\frac{1}{2}}) + \Phi(-2^{\frac{3}{2}k-1}) \right. \\
& \quad \left. + \Phi(-3 \times 2^{\frac{3}{2}(k-1)-\frac{1}{2}}) \right) \\
& \leq 2(\Phi(2 - 2^{\frac{3}{2}(\tilde{K}-2)-\frac{1}{2}}) \frac{1 + 3\exp(-44)}{1 - \exp(-44)}) \\
& \leq \frac{2}{1 - \exp(-40)} \Phi(2 - 2^{\frac{3}{2}(\tilde{K}-2)-\frac{1}{2}}).
\end{aligned}$$

The last three equation uses the fact that  $\Phi(-2\sqrt{2}x) \leq 2\sqrt{2}\exp(-\frac{7x^2}{2})\Phi(-x)$ , for  $x > 0$ .

□

*Proof of Lemma A.1.19.* For the ease of expression, we define  $\tilde{\mathcal{E}}_{j,i} = \frac{1}{\sqrt{m_j}}(Y_3(t_{j,i}) - Y_3(t_{j,i-1}) - \int_{t_{j,i-1}}^{t_{j,i}} f(x)dx)$ . Then  $\tilde{\mathcal{E}}_{j,i} \stackrel{i.i.d}{\sim} N(0, \varepsilon^2 c_e^2)$ ,  $i = 0, 1, \dots$

$$\begin{aligned}
& P\left(G \middle| Z(f) \in [t_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+, \hat{i}_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+} - 5}, t_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+, \hat{i}_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+} + 4}]\right) \\
&= P\left(\hat{f}_1 + S_{i_R - i_L, \frac{\alpha}{4}} \frac{c_e \varepsilon}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}} < M(f) \middle| \right. \\
&\quad \left. Z(f) \in [t_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+, \hat{i}_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+} - 5}, t_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+, \hat{i}_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+} + 4}]\right) \\
&\leq P\left(M(f) + \frac{1}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}} \min_{i_L < i \leq i_R} \tilde{\mathcal{E}}_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}, i} + S_{i_R - i_L, \frac{\alpha}{4}} \frac{c_e \varepsilon}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}} < M(f) \middle| \right. \\
&\quad \left. Z(f) \in [t_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+, \hat{i}_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+} - 5}, t_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+, \hat{i}_{(\hat{j}-K_{\frac{\alpha}{4}}-1)_+} + 4}]\right) \\
&\leq P\left(\min_{i_L < i \leq i_R} \tilde{\mathcal{E}}_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}, i} + c_e \varepsilon S_{i_R - i_L, \frac{\alpha}{4}} < 0\right) \leq \frac{\alpha}{4}.
\end{aligned} \tag{A.2.46}$$

□

*Proof of Lemma A.1.20.*

$$\begin{aligned}
& P(H|E^c \cap F^c) \\
&\leq P(\hat{f}_1 + \Phi^{-1}\left(\frac{\alpha}{4}\right) \frac{c_e \varepsilon}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}} - \frac{\sqrt{3}\varepsilon}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}} > M(f) | E^c \cap F^c) \\
&\leq P\left(\int_{t_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}, i_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}^*}}^{t_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}, i_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}^* + 1}} f(x) \frac{1}{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}} dx + \frac{1}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}} \tilde{\mathcal{E}}_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}, i_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}^* + 1} + \right. \\
&\quad \left. \Phi^{-1}\left(\frac{\alpha}{4}\right) \frac{c_e \varepsilon}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}} - \frac{\sqrt{3}\varepsilon}{\sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}}} > M(f) | E^c \cap F^c\right) \\
&\leq P(\tilde{\mathcal{E}}_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}, i_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}^*} \frac{1}{c_e} + \Phi^{-1}\left(\frac{\alpha}{4}\right) \varepsilon + \rho_m(\varepsilon; f) \sqrt{m_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}} - \sqrt{3}\varepsilon > 0 | E^c \cap F^c) \\
&\leq P(\tilde{\mathcal{E}}_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}, i_{\hat{j} + \tilde{K}_{\frac{\alpha}{4}}}^*} \frac{1}{c_e} + \Phi^{-1}\left(\frac{\alpha}{4}\right) \varepsilon + \rho_m(\varepsilon; f) \sqrt{\frac{1}{2} \rho_z(\varepsilon; f)} - \sqrt{3}\varepsilon > 0 | E^c \cap F^c) \\
&\leq \frac{\alpha}{4}.
\end{aligned} \tag{A.2.47}$$

□

*Proof of Lemma A.1.21.* Let  $i_l = \min\{i : g_{n,\sigma,h}(x_i) > f(x_i)\}$ ,  $i_r = \max\{i : g_{n,\sigma,h}(x_i) > f(x_i)\}$ .

We will first prove the lemma for the case  $\rho_z(\frac{\sigma}{\sqrt{6n}}; h) \geq 1/2n$ .

When  $\{i : g_{n,\sigma,h}(x_i) > f(x_i)\} = \emptyset$ , the lemma holds naturally.

When  $i_l = i_r$ , let  $x_l = \inf\{x : g_{n,\sigma,h}(x) > h(x)\}$ ,  $x_r = \sup\{x : g_{n,\sigma,h}(x) > h(x)\}$ , then we have

$$\begin{aligned} \frac{\sigma^2}{6n} &\geq \|h - g_{n,\sigma,h}\|_2^2 \geq \frac{1}{3}(x_r - x_l)\rho_m\left(\frac{\sigma}{\sqrt{6n}}; h\right)^2 \geq \frac{1}{6} \frac{\rho_m\left(\frac{\sigma}{\sqrt{6n}}; h\right)^2}{n} \\ &\geq \frac{1}{6} l_n(h, g_{n,\sigma,h})^2 = \frac{1}{6} l_n(f, g_{n,\sigma,h})^2. \end{aligned}$$

When  $i_l < i_r$ ,

$$\begin{aligned} \frac{\sigma^2}{6n} &\geq \|h - g_{n,\sigma,h}\|_2^2 \geq \sum_{k=i_l}^{i_r} \frac{1}{3} \frac{1}{2n} (h(x_k) - g_{n,\sigma,h}(x_k))^2 \geq \frac{1}{6} l_n(h, g_{n,\sigma,h})^2 \\ &= \frac{1}{6} l_n(f, g_{n,\sigma,h})^2. \end{aligned}$$

Now we turn to the second case  $\rho_z(\frac{\sigma}{\sqrt{6n}}; h) < 1/2n$ .

Since  $\rho_z(\frac{\sigma}{\sqrt{6n}}; h) < 1/2n$ , then  $|\{i : g_{n,\sigma,h}(x_i) > f(x_i)\}| \leq 1$ . When  $|\{i : g_{n,\sigma,h}(x_i) > f(x_i)\}| = 0$ , the lemma holds naturally. When  $|\{i : g_{n,\sigma,h}(x_i) > f(x_i)\}| = 1$ , we have

$$l_n(f, g_{n,\sigma,h})^2 = l_n(h, g_{n,\sigma,h})^2 \leq \frac{1}{n} \rho_m\left(\frac{\sigma}{\sqrt{6n}}; h\right)^2 \cdot 2n \rho_z\left(\frac{\sigma}{\sqrt{6n}}; h\right) \leq \sigma^2.$$

□



*Proof of Lemma A.1.22.*

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\hat{j} < \tilde{j}\} 1.5m_{\tilde{j}}) \\
& \leq \mathbb{E}(\mathbb{1}\{\hat{j} < \tilde{j}\} 1.5m_{\tilde{j}} \mathbb{1}\{\hat{j} \leq j^* - 3\}) + \mathbb{E}(\mathbb{1}\{\hat{j} < \tilde{j}\} 1.5m_{\tilde{j}} \mathbb{1}\{\hat{j} \geq j^* - 2\}) \quad (\text{A.2.48}) \\
& \leq 1.5\mathbb{E}(\mathbb{1}\{\hat{j} < \tilde{j}\} m_{\tilde{j}} \mathbb{1}\{\hat{j} \leq j^* - 3\}) + 1.5 \times \rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right)
\end{aligned}$$

Also we have

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\hat{j} < \tilde{j}\} m_{\tilde{j}} \mathbb{1}\{\hat{j} \leq j^* - 3\}) \\
& \leq \sum_{j=0}^{(j^*-3) \wedge (J-1)} \mathbb{E}(\mathbb{1}\{\hat{j} = j, \tilde{j} > j\} m_{\tilde{j}}) + \frac{1}{n} \mathbb{1}\{J \leq j^* - 3\} \\
& \leq \sum_{j=0}^{(j^*-3) \wedge (J-1)} m_j \left( \mathbb{E}(\mathbb{1}\{\tilde{j} > j, Y_{j, \hat{i}_j+6, s} - Y_{j, \hat{i}_j+5, s} \leq \gamma_s 2\sqrt{2}\sqrt{2^{J-j}}\sigma\}) \right. \\
& \quad \left. + \mathbb{E}(\mathbb{1}\{\tilde{j} > j, Y_{j, \hat{i}_j-6, s} - Y_{j, \hat{i}_j-5, s} \leq \gamma_s 2\sqrt{2}\sqrt{2^{J-j}}\sigma\}) \right) \\
& \quad + \frac{1}{n} \mathbb{1}\{J \leq j^* - 3\} \\
& \leq \sum_{j=0}^{(j^*-3) \wedge (J-1)} m_j \mathbb{E}\left(\mathbb{1}\{\tilde{j} > j, \frac{\sqrt{2^{J-j}}}{\gamma_s \sqrt{2}\sigma} (\text{ave}_f(j, \hat{i}_j + 6) - \text{ave}_f(j, \hat{i}_j + 5)) \leq \right. \\
& \quad \left. \frac{(\mathfrak{E}_{j, \hat{i}_j+5, s} - \mathfrak{E}_{j, \hat{i}_j+6, s})}{\sqrt{2}\sqrt{2^{J-j}}\gamma_s \sigma} + 2\}\right) + m_j \mathbb{E}\left(\mathbb{1}\{\tilde{j} > j, \right. \\
& \quad \left. \frac{\sqrt{2^{J-j}}}{\gamma_s \sqrt{2}\sigma} (\text{ave}_f(j, \hat{i}_j - 6) - \text{ave}_f(j, \hat{i}_j - 5)) \leq \frac{(\mathfrak{E}_{j, \hat{i}_j-5, s} - \mathfrak{E}_{j, \hat{i}_j-6, s})}{\gamma_s \sqrt{2}\sigma \sqrt{2^{J-j}}} + 2\}\right) \\
& \quad + \frac{1}{n} \mathbb{1}\{J \leq j^* - 3\} \\
& \leq \sum_{j=0}^{(j^*-3) \wedge (J-1)} m_j \mathbb{E}(\mathbb{1}\{\tilde{j} > j\} \Phi(2 - \frac{\rho_m(\frac{\sigma}{\sqrt{n}}; f)}{\rho_z(\frac{\sigma}{\sqrt{n}}; f)} m_j^{\frac{3}{2}} \frac{\sqrt{n}}{\gamma_s \sqrt{2}\sigma})) \times 2 + \frac{1}{n} \mathbb{1}\{J \leq j^* - 3\} \\
& \leq \sum_{j=0}^{(j^*-3) \wedge (J-1)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right) 2^{j^*-j-2} \mathbb{E}(\mathbb{1}\{\tilde{j} > j\}) \times 2\Phi(2 - \frac{1}{2\gamma_s} 2^{\frac{3}{2}(j^*-j-3)}) + \frac{1}{n} \mathbb{1}\{J \leq j^* - 3\} \\
& \leq c_{z0} \rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right) + \frac{1}{n} \mathbb{1}\{J \leq j^* - 3\}. \quad (\text{A.2.49})
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} < \tilde{\mathbf{j}}\}1.5m_{\tilde{\mathbf{j}}}) &\leq 1.5(c_{z0} + 1)\rho_z(\frac{\sigma}{\sqrt{n}}; f) + \frac{3}{2}\frac{1}{n}\mathbb{1}\{J \leq \mathbf{j}^* - 3\} \\
&= c_{z1}\rho_z(\frac{\sigma}{\sqrt{n}}; f) + \frac{3}{2}\frac{1}{n}\mathbb{1}\{J \leq \mathbf{j}^* - 3\}.
\end{aligned} \tag{A.2.50}$$

□

*Proof of Lemma A.1.23.*

$$\begin{aligned}
&\mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} \geq \tilde{\mathbf{j}}\}|\hat{Z} - Z(f)|) \\
&\leq \mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} \geq \tilde{\mathbf{j}}\}6m_{\tilde{\mathbf{j}}}) \\
&\leq 6 \sum_{j=3}^{(\mathbf{j}^*-3) \wedge J} \rho_z(\frac{\sigma}{\sqrt{n}}; f) 2^{\mathbf{j}^*-j-2} \Phi(-\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; f)}{\rho_z(\frac{\sigma}{\sqrt{n}}; f)} m_{\mathbf{j}^*}^{\frac{3}{2}} 2^{\frac{3}{2}(\mathbf{j}^*-j)} \frac{\sqrt{n}}{\gamma_l \sigma \sqrt{2}}) \\
&\quad + 6\mathbb{1}\{J \geq \mathbf{j}^* - 2\}\rho_z(\frac{\sigma}{\sqrt{n}}; f) \\
&\leq c_{z2}\rho_z(\frac{\sigma}{\sqrt{n}}; f)
\end{aligned} \tag{A.2.51}$$

□

*Proof of Lemma A.1.24.*

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\check{j} < \infty\} \mathbb{1}\{\check{j} < \tilde{j}\} |\hat{Z} - Z(\tilde{h})|) \\
& \leq \mathbb{E}(\mathbb{1}\{\check{j} < \infty\} \mathbb{1}\{\check{j} < \tilde{j}\} 1.5m_{\check{j}}) \\
& \leq \sum_{j=3}^J 1.5 \frac{2^{J-j}}{n} \cdot 2\Phi\left(2 - \frac{\rho_m(\frac{\sigma}{\sqrt{n}}; h)}{\rho_z(\frac{\sigma}{\sqrt{n}}; h)} \frac{2^{J-j}}{n} \frac{1}{\gamma_s \sigma \sqrt{2} \sqrt{2^{J-j}}}\right) \\
& \leq \sum_{j=3}^J 3 \frac{2^{J-j}}{n} \cdot \Phi\left(2 - \frac{\frac{1}{\sqrt{2n}} \sigma}{\rho_z(\frac{\sigma}{\sqrt{n}}; h) \sqrt{\rho_z(\frac{\sigma}{\sqrt{n}}; h)}} \frac{2^{\frac{3}{2}(J-j)}}{n} \frac{1}{\sqrt{2} \sigma \gamma_s}\right) \\
& = \sum_{j=3}^J 3 \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right) \frac{2^{J-j}}{n \rho_z(\frac{\sigma}{\sqrt{n}}; h)} \cdot \Phi\left(2 - \left(\frac{2^{J-j}}{n \rho_z(\frac{\sigma}{\sqrt{n}}; h)}\right)^{\frac{3}{2}} \frac{1}{2 \gamma_s}\right) \\
& \leq \sum_{j=3}^J 3 \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)} 2^{\frac{j-J}{2}} \cdot 2 \gamma_s \check{C} \\
& \leq \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)} \check{c}_{z1},
\end{aligned} \tag{A.2.52}$$

where  $\check{C} = \sup_{x>0} x \Phi(2-x)$ . □

*Proof of Lemma A.1.25.*

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\hat{j} \geq \tilde{j}\} |\hat{Z} - Z(\tilde{h})|) \leq \mathbb{E}(\mathbb{1}\{\hat{j} \geq \tilde{j}\} 6m_{\tilde{j}}) \\
& \leq \sum_{j=1}^J 6 \frac{2^{J-j}}{n} \cdot 6\Phi\left(-\frac{2^{J-j} \frac{\rho_m(\frac{\sigma}{\sqrt{n}}; h)}{\rho_z(\frac{\sigma}{\sqrt{n}}; h)} \frac{2^{J-j}}{n}}{\sqrt{2} \gamma_s \sigma \sqrt{2^{J-j}}}\right) \leq \sum_{j=1}^J 6 \frac{2^{J-j}}{n} \cdot 6\Phi\left(-\frac{\left(\frac{2^{J-j}}{n \rho_z(\frac{\sigma}{\sqrt{n}}; h)}\right)^{\frac{3}{2}}}{2 \gamma_s}\right) \\
& \leq \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)} \check{c}_{z2}
\end{aligned} \tag{A.2.53}$$

□

*Proof of Lemma A.1.29.*

$$\begin{aligned}
& \sum_{j=1}^{j^*-1} \mathbb{E}(2^{-j} \mathbb{1}\{\tilde{j} = j, \tilde{j} > j\}) \leq 2^{-J} \mathbb{1}\{J \leq j^* - 1\} + \\
& \sum_{j=1}^{\min\{j^*-1, J\}} \mathbb{E} \left( 2^{-j} \left( \mathbb{1}\{Y_{j, \hat{i}_j+6, s} - Y_{j, \hat{i}_j+5, s} \leq 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j}}\} \right. \right. \\
& \quad \left. \left. + \mathbb{1}\{Y_{j, \hat{i}_j-6} - Y_{j, \hat{i}_j-5} \leq 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j}}\} \right) \mathbb{1}\{\tilde{j} > j\} \right) \\
& \leq \sum_{j=1}^{\min\{j^*-1, J\}} 2^{-j+1} \Phi \left( 2 - \frac{\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; f)}{\rho_z(\frac{\sigma}{\sqrt{n}}; f)} \frac{2^{J-j}}{n} 2^{J-j}}{\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j}}} \right) \mathbb{E}(\mathbb{1}\{\tilde{j} > j\}) \\
& \quad + 2^{-J} \mathbb{1}\{J \leq j^* - 1\}. \\
& \leq \sum_{j=1}^{\min\{j^*-1, J\}} 2^{-j^*} \cdot 2^{(j^*-j)+1} \Phi \left( 2 - \frac{1}{2\gamma_s} 2^{\frac{3}{2}(j^*-j-3)} \right) + 2^{-J} \mathbb{1}\{J \leq j^* - 1\} \\
& \leq c_2 3^{2^{-j^*}} + 2^{-J} \mathbb{1}\{J \leq j^* - 1\}.
\end{aligned} \tag{A.2.54}$$

□

*Proof of Lemma A.1.30.*

$$\begin{aligned}
& \sum_{j=1}^{j^*-1} \mathbb{E}(2^{-j} \mathbb{1}\{\hat{j} = j, \tilde{j} \leq j\}) \\
& \leq \sum_{j=1}^{(j^*-3) \wedge (J-1)} \mathbb{E}(2^{-j} \mathbb{1}\{\tilde{j} = j\}) \\
& \leq \sum_{j=1}^{(j^*-3) \wedge (J-1)} 2^{-j} \cdot 6\Phi\left(-\frac{\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; f)}{\rho_z(\frac{\sigma}{\sqrt{n}}; f)} \cdot \frac{2^{J-j}}{n} 2^{J-j}}{\sqrt{2}\gamma_l \sigma \sqrt{2^{J-j}}}\right) \\
& \leq \sum_{j=1}^{(j^*-3) \wedge (J-1)} 2^{-j} \cdot 6\Phi\left(-\frac{1}{2\gamma_l} 2^{\frac{3}{2}(j^*-j-3)}\right) \tag{A.2.55} \\
& \leq \sum_{j=1}^{(j^*-3) \wedge (J-1)} 2^{-j^*} \cdot 2^{j^*-j} \cdot 6\Phi\left(-\frac{1}{2\gamma_l} 2^{\frac{3}{2}(j^*-j-3)}\right) \\
& \leq 2^{-j^*} \sum_{j=1}^{\infty} 6 \cdot 2^j \Phi\left(-\frac{1}{2\gamma_l} 2^{\frac{3}{2}(j-3)}\right) \\
& \leq 2^{-j^*} c_{z4}.
\end{aligned}$$

□

*Proof of Lemma A.1.31.*

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\check{j} < \infty\} L(\mathbf{CI}_{z,\alpha}(Y))) \leq (12 \cdot 2^{K_{\alpha/2}+1} + 1) \mathbb{E}\left(\frac{2^{J-\check{j}}}{n} \mathbb{1}\{\check{j} < \infty\}\right) \\
& = (12 \cdot 2^{K_{\alpha/2}+1} + 1) \sum_{j=3}^J \mathbb{E}(\mathbb{1}\{\check{j} = j\}) \frac{2^{J-j}}{n} \\
& = (12 \cdot 2^{K_{\alpha/2}+1} + 1) \times \tag{A.2.56} \\
& \quad \left( \sum_{j=3}^J \mathbb{E}(\mathbb{1}\{\check{j} = j\} \mathbb{1}\{\tilde{j} \leq j\}) \frac{2^{J-j}}{n} + \sum_{j=3}^J \mathbb{E}(\mathbb{1}\{\check{j} = j\} \mathbb{1}\{\tilde{j} > j\}) \frac{2^{J-j}}{n} \right).
\end{aligned}$$

We bound the two terms separately and we start with the first term.

$$\begin{aligned}
& \sum_{j=3}^J \mathbb{E}(\mathbb{1}\{\check{j} = j\} \mathbb{1}\{\check{j} \leq j\}) \frac{2^{J-j}}{n} \leq \sum_{j=3}^J \mathbb{E} \left( \mathbb{1}\{\check{j} = j\} \mathbb{1}\{\check{j} \leq j\} \frac{2^{J-\check{j}}}{n} \right) \\
& \leq \sum_{j=1}^J \mathbb{E}(\mathbb{1}\{\check{j} = j\}) \frac{2^{J-j}}{n} \leq \sum_{j=1}^J \frac{2^{J-j}}{n} 6\Phi \left( -\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})}{\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{2^{J-j}}{n} \frac{2^{J-j}}{\sqrt{2^{J-j}} \sqrt{2} \gamma_l \sigma} \right) \\
& \leq \sum_{j=1}^J \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} 2^{\frac{j-J}{2}} \left( \frac{2^{J-j}}{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \right)^{\frac{3}{2}} \cdot 6 \cdot \Phi \left( -\left( \frac{2^{J-j}}{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \right)^{\frac{3}{2}} \frac{1}{2\gamma_l} \right) \quad (\text{A.2.57}) \\
& \leq 6 \frac{1}{1 - \sqrt{1/2}} \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \cdot 2\gamma_l \cdot \check{C} \\
& \leq 6 \frac{1}{1 - \sqrt{1/2}} \check{C} \cdot 2\gamma_l \cdot \sup_{h \in \mathcal{G}_n(f)} \rho_z(\frac{\sigma}{\sqrt{n}}; h) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; h)},
\end{aligned}$$

where  $\check{C} = \sup_{t>0} t\Phi(-t)$ .

For the second term, we have

$$\begin{aligned}
& \sum_{j=3}^J \mathbb{E}(\mathbb{1}\{\check{j} = j\} \mathbb{1}\{\check{j} > j\}) \frac{2^{J-j}}{n} \\
& \leq \sum_{j=3}^J \frac{2^{J-j}}{n} 2\Phi \left( 2 - \frac{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})}{\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{2^{J-j}}{n} \frac{2^{J-j}}{\sqrt{2^{J-j}} \sqrt{2} \gamma_s \sigma} \right) \\
& \leq \sum_{j=3}^J 2\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} 2^{\frac{j-J}{2}} \left( \frac{2^{J-j}}{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \right)^{\frac{3}{2}} \Phi \left( 2 - \left( \frac{2^{J-j}}{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \right)^{\frac{3}{2}} \frac{1}{2\gamma_s} \right) \quad (\text{A.2.58}) \\
& \leq 2 \cdot 2\gamma_s \cdot \check{Q} \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{1}{1 - \sqrt{1/2}},
\end{aligned}$$

where  $\check{Q} = \sup_{t>0} t\Phi(2-t)$ .

Let  $\check{c}_{1,\alpha} = (6 \frac{1}{1-\sqrt{1/2}} 2\gamma_l \check{C} + 4\gamma_s \cdot \check{Q} \frac{1}{1-\sqrt{1/2}}) \cdot (12 \cdot 2^{K_{\alpha/2}+1} + 1)$  gives the statement of Lemma A.1.31.

□

*Proof of Lemma A.1.32.* When  $2 \leq i_m \leq n-2$ ,  $t_{hi} - t_{lo} \geq \frac{3}{n}$  implies that  $i_l \leq i_m - 1$  or

$i_r \geq i_m$ . When  $i_m \leq 1$ ,  $t_{hi} - t_{lo} \geq \frac{3}{n}$  implies that  $i_r \geq i_m$ . When  $i_m \geq n - 1$ ,  $t_{hi} - t_{lo} \geq \frac{3}{n}$  implies that  $i_l \leq i_m - 1$ . Therefore, we have

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{t_{hi} - t_{lo} \geq \frac{3}{n}\} L(\text{CI}_{z,\alpha}(Y))) \\
& \leq \frac{1 + 12 \cdot 2^{K_{\alpha/2}+1}}{n} \mathbb{E}\left(\mathbb{1}\{i_l \leq i_m - 1\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{i_m \geq 2\} \right. \\
& \quad \left. + \mathbb{1}\{i_r \geq i_m\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{i_m \leq n - 2\}\right) \\
& = \frac{1 + 12 \cdot 2^{K_{\alpha/2}+1}}{n} \mathbb{E}\left(\mathbb{1}\{i_l \leq i_m - 1\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{U \leq i_m - 1, i_m \geq 2\} + \right. \\
& \quad \mathbb{1}\{i_l \leq i_m - 1\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{U \geq i_m, i_m \geq 2\} \\
& \quad + \mathbb{1}\{i_r \geq i_m\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{L \geq i_m + 1, i_m \leq n - 2\} \\
& \quad \left. + \mathbb{1}\{i_r \geq i_m\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{L \leq i_m, i_m \leq n - 2\}\right). \tag{A.2.59}
\end{aligned}$$

Since  $\{U \leq i_m - 1, i_m \geq 2\} \cup \{L \geq i_m + 1, i_m \leq n - 2\}$  implies that  $\check{j} < n$ , and  $\{U \leq i_m - 1\} \cap \{L \geq i_m + 1\} = \emptyset$ , we have

$$\begin{aligned}
& \mathbb{E}\left(\mathbb{1}\{i_l \leq i_m - 1\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{U \leq i_m - 1\} + \mathbb{1}\{i_r \geq i_m\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{L \geq i_m + 1\}\right) \\
& \leq \mathbb{E}(\mathbb{1}\{\check{j} < n\}) = \sum_{j=2}^J P(\check{j} = j) \leq \sum_{j=2}^J \Phi\left(-\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})}{\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{2^{J-j}}{n} \frac{2^{J-j}}{\sqrt{2^{J-j}} \sqrt{2} \gamma_l \sigma}\right) \\
& \leq \sum_{j=2}^J \Phi\left(-\left(\frac{2^{J-j}}{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})}\right)^{\frac{3}{2}} \frac{1}{2 \gamma_l}\right) \\
& \leq n \rho_z\left(\frac{\sigma}{\sqrt{n}}; \tilde{h}\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; \tilde{h}\right)} \frac{1}{1 - \sqrt{\frac{1}{8}}} 2 \gamma_l \check{C}, \tag{A.2.60}
\end{aligned}$$

where  $\check{C} = \sup_{t>0} t \Phi(-t)$ .

And for  $\mathbb{E}(\mathbb{1}\{i_l \leq i_m - 1\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{U \geq i_m, i_m \geq 2\} + \mathbb{1}\{i_r \geq i_m\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{L \leq$

$i_m, i_m \leq n - 2\}$ ), we have

$$\begin{aligned}
& \mathbb{E}\left(\mathbb{1}\{i_l \leq i_m - 1\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{U \geq i_m, i_m \geq 2\} \right. \\
& \quad \left. + \mathbb{1}\{i_r \geq i_m\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{L \leq i_m, i_m \leq n - 2\}\right) \\
&= \mathbb{E}\left(\mathbb{E}(\mathbb{1}\{i_l \leq i_m - 1\} | \mathbf{Y}_l, \mathbf{Y}_s) \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{U \geq i_m, i_m \geq 2\} \right. \\
& \quad \left. + \mathbb{E}(\mathbb{1}\{i_r \geq i_m\} | \mathbf{Y}_l, \mathbf{Y}_s) \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{L \leq i_m, i_m \leq n - 2\}\right) \\
&\leq 2(U - L) \Phi\left(-\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})}{\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{1}{n2\sqrt{3}\sigma} + z_{\alpha_1}\right) \\
&\leq 2(U - L) \Phi\left(-\left(\frac{1}{n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})}\right)^{\frac{3}{2}} \frac{1}{\sqrt{24}} + z_{\alpha_1}\right) \\
&\leq n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})\sqrt{24}} \cdot 2 \cdot \check{Q}_2(U - L),
\end{aligned} \tag{A.2.61}$$

where  $\check{Q}_2 = \sup_{t>0} x\Phi(z_{\alpha_1} - x)$ .

Note that  $U - L$  only depends on  $\alpha$ , therefore,

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{t_{hi} - t_{lo} \geq \frac{3}{n}\} L(\mathbf{CI}_{z,\alpha}(Y))) \\
&\leq \check{c}_{2,\alpha} \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \\
&\leq \check{c}_{2,\alpha} \sup_{h \in \mathcal{G}_n(f)} \rho_z(\frac{\sigma}{\sqrt{n}}; h) \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; h)}.
\end{aligned} \tag{A.2.62}$$

□

*Proof of Lemma A.1.33.* Note that when  $0 < t_{hi} - t_{lo} < \frac{3}{n}$ , one of the following holds:  
 $i_l = n = U = i_r + 1$ ,  $i_r = -1 = L - 1 = i_l - 1$ ,  $L + 1 \leq i_l = i_r + 1 \leq U - 1$ ,  $i_l = L = i_r$ ,  
 $i_r = U - 1 = i_l$ . We denote event

$$H_1 = \{i_l = n = U = i_r + 1\} \cup \{i_r = -1 = L - 1 = i_l - 1\} \cup \{L + 1 \leq i_l = i_r + 1 \leq U - 1\},$$

$$H_2 = \{i_l = L = i_r\} \cup \{i_r = U - 1 = i_l\}.$$



Therefore,

$$\begin{aligned} & \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{t_{hi} - t_{lo} < \frac{3}{n}\} L(\mathbf{CI}_{z,\alpha}(Y))) \\ &= \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} L(\mathbf{CI}_{z,\alpha}(Y)) \mathbb{1}\{H_1\}) + \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} L(\mathbf{CI}_{z,\alpha}(Y)) \mathbb{1}\{H_2\}). \end{aligned} \quad (\text{A.2.63})$$

We start with the second term

$$\begin{aligned} & \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} L(\mathbf{CI}_{z,\alpha}(Y)) \mathbb{1}\{H_2\}) \\ & \leq \mathbb{E}\left(\mathbb{1}\{\check{j} = \infty\} (t_{hi} - t_{lo}) (\mathbb{1}\{i_m \leq L-1\} + \mathbb{1}\{i_m \geq L\} \mathbb{1}\{i_l = L = i_r\} \right. \\ & \quad \left. + \mathbb{1}\{i_m \geq U+1\} + \mathbb{1}\{i_m \leq U\} \mathbb{1}\{i_r = U-1 = i_l\})\right) \\ & \leq \frac{2}{n} \left( \sum_{i=2}^{J-1} 6\Phi\left(-\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})}{\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{2^{J-j}}{n} \frac{2^{J-j}}{\sqrt{2^{J-j}} \gamma_l \sigma \sqrt{2}}\right) + 2\Phi\left(-\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})}{\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{1}{n} \frac{1}{2\sqrt{3}\sigma} + z_{\alpha_1}\right) \right) \quad (\text{A.2.64}) \\ & \leq \frac{2}{n} n \rho_z\left(\frac{\sigma}{\sqrt{n}}; \tilde{h}\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; \tilde{h}\right)} (12\gamma_l 2^{-\frac{3}{2}} \frac{1}{1 - \sqrt{\frac{1}{8}}} \check{C} + 2 \cdot 2\sqrt{6}\check{Q}_2) \\ & = \rho_z\left(\frac{\sigma}{\sqrt{n}}; \tilde{h}\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; \tilde{h}\right)} (24\gamma_l 2^{-\frac{3}{2}} \frac{1}{1 - \sqrt{\frac{1}{8}}} \check{C} + 4 \cdot 2\sqrt{6}\check{Q}_2). \end{aligned}$$

Now we turn to the first term

$$\begin{aligned} & \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} L(\mathbf{CI}_{z,\alpha}(Y)) \mathbb{1}\{H_1\}) \\ & \leq \mathbb{E}\left(\mathbb{1}\{\check{j} = \infty\} (t_{hi} - t_{lo}) (\mathbb{1}\{i_l = n = U = i_r + 1\} \right. \\ & \quad \left. + \mathbb{1}\{i_r = -1 = L-1 = i_l - 1\} + \mathbb{1}\{L+1 \leq i_l = i_r + 1 \leq U-1\})\right) \\ & \leq \mathbb{E}\left(\mathbb{1}\{\check{j} = \infty\} (t_{hi} - t_{lo}) \mathbb{1}\{i_l = i_m\} (\mathbb{1}\{i_l = n = U = i_r + 1\} \right. \\ & \quad \left. + \mathbb{1}\{i_r = -1 = L-1 = i_l - 1\} + \mathbb{1}\{L+1 \leq i_l = i_r + 1 \leq U-1\})\right) \quad (\text{A.2.65}) \\ & + \mathbb{E}\left(\mathbb{1}\{\check{j} = \infty\} (t_{hi} - t_{lo}) \mathbb{1}\{i_l \neq i_m\} (\mathbb{1}\{i_l = n = U = i_r + 1\} \right. \\ & \quad \left. + \mathbb{1}\{i_r = -1 = L-1 = i_l - 1\} + \mathbb{1}\{L+1 \leq i_l = i_r + 1 \leq U-1\})\right). \end{aligned}$$

We will bound the two terms separately, we start with the second term.

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (t_{hi} - t_{lo}) \mathbb{1}\{i_l \neq i_m\} (\mathbb{1}\{i_l = n = U = i_r + 1\} \right. \\
& \quad \left. + \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\}) \right) \\
& < \frac{3}{n} \left( \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{i_l = n = U = i_r + 1\}) \mathbb{1}\{i_m \leq n - 1\} \right. \\
& \quad + \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\}) \mathbb{1}\{i_m \geq 1\} \\
& \quad + \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{i_l \leq i_m - 1\} \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\}) \\
& \quad \left. + \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{i_l \geq i_m + 1\} \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\}) \right) \\
& \leq \frac{3}{n} \left( \mathbb{E} \left( \Phi \left( -\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})}{\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{1}{n\sqrt{12}\sigma} \right) \mathbb{1}\{U = n\} \mathbb{1}\{\check{j} = \infty\} \right) \mathbb{1}\{i_m \leq n - 1\} \right. \\
& \quad + \mathbb{E} \left( \Phi \left( -\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})}{\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{1}{n\sqrt{12}\sigma} \right) \mathbb{1}\{L = 0\} \mathbb{1}\{\check{j} = \infty\} \right) \mathbb{1}\{i_m \geq 1\} \\
& \quad + \mathbb{1}\{2 \leq i_m \leq n - 2\} (P(U \leq i_m, \check{j} = \infty) \\
& \quad + P(U > i_m, \check{j} = \infty) \Phi \left( -\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})}{\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{1}{n\sqrt{12}\sigma} \right) \\
& \quad + P(L \geq i_m, \check{j} = \infty) + P(L < i_m, \check{j} = \infty) \Phi \left( -\frac{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})}{\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{1}{n\sqrt{12}\sigma} \right)) \Big) \\
& \leq \frac{3}{n} \left( 4\Phi \left( -\left( \frac{1}{n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \right)^{\frac{3}{2}} \frac{1}{\sqrt{24}} \right) + \sum_{j=2}^{J-1} 6\Phi \left( -\left( \frac{2^{J-j}}{n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \right)^{\frac{3}{2}} \frac{1}{2\gamma_l} \right) \right) \\
& \leq \rho_z \left( \frac{\sigma}{\sqrt{n}}; \tilde{h} \right) \sqrt{n\rho_z \left( \frac{\sigma}{\sqrt{n}}; \tilde{h} \right)} \cdot 3 \cdot (4\sqrt{24} + 12\gamma_l) \check{C}.
\end{aligned} \tag{A.2.66}$$

Now we turn to the first term. We discuss the four settings:  $i_m = 0$ ,  $i_m = n$ ,  $2 \leq i_m \leq n - 2$ ,

$(i_m - 1)(i_m - n + 1) = 0$ . Note that when  $(i_m - 1)(i_m - n + 1) = 0$ ,  $\mathfrak{D}_z(n, f) \geq \frac{1}{n}$ . Therefore,

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (t_{hi} - t_{lo}) \mathbb{1}\{i_l = i_m\} \left( \mathbb{1}\{i_l = n = U = i_r + 1\} \right. \right. \\
& \quad \left. \left. + \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\} \right) \right) \\
&= \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} \frac{2}{n} \mathbb{1}\{i_l = i_m\} \mathbb{1}\{(i_m - 1)(i_m - n + 1) = 0\} \right. \\
& \quad \left( \mathbb{1}\{i_l = n = U = i_r + 1\} + \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} \right. \\
& \quad \left. \left. + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\} \right) \right) \\
&\leq 2\mathfrak{D}_z(n, f).
\end{aligned} \tag{A.2.67}$$

Now we turn to the cases  $(i_m - 1)(i_m - n + 1) \neq 0$ . Note that under the event  $\{i_l = n = U = i_r + 1\} \cup \{i_r = -1 = L - 1 = i_l - 1\} \cup \{L + 1 \leq i_l = i_r + 1 \leq U - 1\}$ ,  $t_{lo} \leq i_l/n \leq t_{hi}$ .

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (t_{hi} - t_{lo}) \mathbb{1}\{i_l = i_m\} \left( \mathbb{1}\{i_l = n = U = i_r + 1\} \right. \right. \\
& \quad \left. \left. + \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\} \right) \right) \\
&= \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (t_{hi} - i_l/n) \mathbb{1}\{i_l = i_m\} \left( \mathbb{1}\{i_l = n = U = i_r + 1\} \right. \right. \\
& \quad \left. \left. + \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\} \right) \right) \\
&+ \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (i_l/n - t_{lo}) \mathbb{1}\{i_l = i_m\} \left( \mathbb{1}\{i_l = n = U = i_r + 1\} \right. \right. \\
& \quad \left. \left. + \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\} \right) \right).
\end{aligned} \tag{A.2.68}$$

Due to the symmetric nature of the procedure, the case  $(i_m - 1)(i_m - n + 1) \neq 0$ , and the event  $\{i_l = i_m\} \cap \{i_l = n = U = i_r + 1\} \cup \{i_r = -1 = L - 1 = i_l - 1\} \cup \{L + 1 \leq i_l = i_r + 1 \leq U - 1\}$ , we only need to bound the first term, and the second term shares the similar (symmetric) bound.

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (t_{hi} - i_l/n) \mathbb{1}\{i_l = i_m\} (\mathbb{1}\{i_l = n = U = i_r + 1\} \right. \\
& \quad \left. + \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\}) \right) \\
& = \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (t_{hi} - i_l/n) \mathbb{1}\{i_l = i_m\} \right. \\
& \quad \left. (\mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\}) \right).
\end{aligned} \tag{A.2.69}$$

Note that  $i_{hi} = i_l + 1, i_{lo} = i_l - 1$  when  $\{L + 1 \leq i_l = i_r + 1 \leq U - 1\}$ , and if  $i_{hi}$  is also so defined when  $\{i_r = -1 = L - 1 = i_l - 1\}$ , then the definition of  $t_{hi}$  defined under  $\{i_r = -1 = L - 1 = i_l - 1\}$  has the same form with that defined under the case  $\{L + 1 \leq i_l = i_r + 1 \leq U - 1\}$ .

Let

$$t_{raw}(i) = \frac{y_{e,i-1} - y_{e,i} - \sqrt{3}\sigma(z_{3,i-1} - z_{3,i}) + 2\sqrt{6}\sigma z_{\alpha_2}}{n(y_{e,i+1} - y_{e,i} - \sqrt{3}\sigma(z_{3,i+1} - z_{3,i}) + 2\sqrt{6}\sigma z_{\alpha_2})}.$$

Then  $t_{hi} = \left( (t_{raw}(i_{hi}) + \frac{i_{hi}}{n}) \vee \frac{i_{hi}-1}{n} \right) \wedge \frac{i_{hi}}{n}$ .

And let

$$q(i) = n(y_{e,i+1} - y_{e,i} - \sqrt{3}\sigma(z_{3,i+1} - z_{3,i}) + 2\sqrt{6}\sigma z_{\alpha_2}).$$

Then we have

$$\begin{aligned}
&= \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (t_{hi} - i_l/n) \mathbb{1}\{i_l = i_m\} \right. \\
&\quad \left. (\mathbb{1}\{i_r = -1 = L-1 = i_l - 1\} + \mathbb{1}\{L+1 \leq i_l = i_r + 1 \leq U-1\}) \right) \\
&= \mathbb{E} \left( \mathbb{E} \left( \left( \left( t_{raw}(i_{hi}) + \frac{1}{n} \right) \wedge \frac{1}{n} \right) \mathbb{1}\{q(i_{hi}) > 0, -\frac{1}{n} \leq t_{raw}(i_{hi})\} \middle| \mathbf{Y}_l, \mathbf{Y}_s, Y_{e,1} \right) \right. \\
&\quad \left. \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{i_l = i_m\} (\mathbb{1}\{i_r = -1 = L-1 = i_l - 1\} \right. \\
&\quad \left. + \mathbb{1}\{L+1 \leq i_l = i_r + 1 \leq U-1\}) \right) \\
&= \mathbb{E} \left( \mathbb{E} \left( \left( \left( t_{raw}(i_m + 1) + \frac{1}{n} \right) \wedge \frac{1}{n} \right) \mathbb{1}\{q(i_m + 1) > 0, -\frac{1}{n} \leq t_{raw}(i_m + 1)\} \middle| \mathbf{Y}_l, \mathbf{Y}_s, Y_{e,1} \right) \right. \\
&\quad \left. \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{i_l = i_m\} (\mathbb{1}\{i_r = -1 = L-1 = i_l - 1\} \right. \\
&\quad \left. + \mathbb{1}\{L+1 \leq i_l = i_r + 1 \leq U-1\}) \right) \\
&= P \left( \check{j} = \infty, i_l = i_m, \text{ and } (i_r = -1 = L-1 = i_l - 1 \text{ or } L+1 \leq i_l = i_r + 1 \leq U-1) \right) \\
&\quad \mathbb{1}\{i_m \leq n-2\} \mathbb{E} \left( \left( \left( t_{raw}(i_m + 1) + \frac{1}{n} \right) \wedge \frac{1}{n} \right) \mathbb{1}\{q(i_m + 1) > 0, -\frac{1}{n} \leq t_{raw}(i_m + 1)\} \right).
\end{aligned} \tag{A.2.70}$$

Note that only when  $i_m \leq n-2$  the above quantity is not 0 ( when  $i_m = n$ , it's 0, and by  $(i_m - n + 1)(i_m - 1) \neq 0$ ,  $i_m \neq n-1$ ), so we take  $i_m \leq n-2$  by default from now before finished bounding this quantity.

Note that, when we denote  $\zeta_i = y_{e,i} - f(x_i) - \sqrt{3}\sigma z_{3,i}$ , then  $\{\frac{\zeta_i}{\sqrt{6}\sigma}\} \stackrel{i.i.d.}{\sim} N(0, 1)$ , and

$$t_{raw}(i_m + 1) + \frac{1}{n} = \frac{f(x_{i_m}) - 2f(x_{i_m+1}) + f(x_{i_m+2}) + \zeta_{i_m} - 2\zeta_{i_m+1} + \zeta_{i_m+2} + 4\sqrt{6}\sigma z_{\alpha_2}}{n(f(x_{i_m+2}) - f(x_{i_m+1}) + \zeta_{i_m+2} - \zeta_{i_m+1} + 2\sqrt{6}\sigma z_{\alpha_2})}. \tag{A.2.71}$$

Therefore, when we, with a bit abuse of the notation, denote the event  $A_0$  only in this proof

to be the following event:

$$\begin{aligned}
A_0 = \left\{ \zeta_{i_m+2} \geq -\frac{f(x_{i_m+2}) - f(x_{i_m+1})}{6} - \sqrt{6}\sigma z_{\alpha_2}, \right. \\
\zeta_{i_m+1} \leq \frac{f(x_{i_m+2}) - f(x_{i_m+1})}{6} + \sqrt{6}\sigma z_{\alpha_2}, \\
\left. \zeta_{i_m} \geq -\frac{f(x_{i_m+2}) - f(x_{i_m+1})}{6} - \sqrt{6}\sigma z_{\alpha_2} \right\}
\end{aligned} \tag{A.2.72}$$

, we have, on event  $A_0$ ,

$$\begin{aligned}
t_{raw}(i_m + 1) + \frac{1}{n} &\geq -\frac{1}{n}, \\
f(x_{i_m+2}) - f(x_{i_m+1}) + \zeta_{i_m+2} - \zeta_{i_m+1} + 2\sqrt{6}\sigma z_{\alpha_2} &\geq \frac{2(f(x_{i_m+2}) - f(x_{i_m+1}))}{3}.
\end{aligned}$$

With a bit abuse of notation, denote event  $B$  only in this proof to be

$$B = \{\zeta_{i_m} - 2\zeta_{i_m+1} + \zeta_{i_m+2} + f(x_{i_m}) - 2f(x_{i_m+1}) + f(x_{i_m+2}) + 4\sqrt{6}\sigma \geq 0\}. \tag{A.2.73}$$

Then on  $B^c \cap A_0$ ,  $t_{raw}(i_m + 1) + \frac{1}{n} < 0$ ; on  $B \cap A_0$ ,  $t_{raw}(i_m + 1) + \frac{1}{n} \geq 0$ .

Further, we have

$$\begin{aligned}
&P(A_0^c) \\
&\leq P(\zeta_{i_m+2} < -\frac{f(x_{i_m+2}) - f(x_{i_m+1})}{6} - \sqrt{6}\sigma z_{\alpha_2}) \\
&\quad + P(\zeta_{i_m+1} > \frac{f(x_{i_m+2}) - f(x_{i_m+1})}{6} + \sqrt{6}\sigma z_{\alpha_2}) \\
&\quad + P(\zeta_{i_m} < -\frac{f(x_{i_m+2}) - f(x_{i_m+1})}{6} - \sqrt{6}\sigma z_{\alpha_2}) \\
&= 3\Phi(-\frac{f(x_{i_m+2}) - f(x_{i_m+1})}{6\sqrt{6}\sigma} - z_{\alpha_2}) \leq 3\Phi(-\left(\frac{1}{n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})}\right)^{\frac{3}{2}} \frac{1}{6\sqrt{12}} - z_{\alpha_2}) \\
&\leq n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} 18\sqrt{12}\check{Q}_3,
\end{aligned} \tag{A.2.74}$$

where  $\check{Q}_3 = \sup_{x>0} x\Phi(-x - z_{\alpha_2})$ .

Therefore, we have

$$\begin{aligned}
& \mathbb{E} \left( \left( \left( t_{raw}(i_m + 1) + \frac{1}{n} \right) \wedge \frac{1}{n} \right) \mathbb{1}\{q(i_m + 1) > 0, -\frac{1}{n} \leq t_{raw}(i_m + 1)\} \right) \\
&= \mathbb{E} \left( \left( \left( t_{raw}(i_m + 1) + \frac{1}{n} \right) \wedge \frac{1}{n} \right) \mathbb{1}\{q(i_m + 1) > 0, -\frac{1}{n} \leq t_{raw}(i_m + 1)\} \right. \\
&\quad \left. (\mathbb{1}\{A_0 \cap B\} + \mathbb{1}\{A_0 \cap B^c\} + \mathbb{1}\{A_0^c\}) \right) \tag{A.2.75} \\
&\leq \mathbb{E} \left( \left( t_{raw}(i_m + 1) + \frac{1}{n} \right) \mathbb{1}\{A_0 \cap B\} \right) + \frac{1}{n} P(A_0^c) \\
&\leq \mathbb{E} \left( \left( t_{raw}(i_m + 1) + \frac{1}{n} \right) \mathbb{1}\{A_0 \cap B\} \right) + \rho_z \left( \frac{\sigma}{\sqrt{n}}; \tilde{h} \right) \sqrt{n \rho_z \left( \frac{\sigma}{\sqrt{n}}; \tilde{h} \right)} 6\sqrt{12} \check{Q}_3.
\end{aligned}$$

Further, given the convexity, we know that

$$\sup\{Z(h) : h(x_i) = f(x_i), 0 \leq i \leq n\} - \frac{i_m}{n} = \frac{f(x_{i_m}) - f(x_{i_m+1})}{n(f(x_{i_m+2}) - f(x_{i_m+1}))} + \frac{1}{n}.$$

Therefore, we have

$$\begin{aligned}
& \mathbb{E} \left( \left( t_{raw}(i_m + 1) + \frac{1}{n} \right) \mathbb{1}\{A_0 \cap B\} \right) \\
&= \sup\{Z(h) : h(x_i) = f(x_i), 0 \leq i \leq n\} - \frac{i_m}{n} + \\
&\quad \mathbb{E} \left( \left( t_{raw}(i_m + 1) - \frac{f(x_{i_m}) - f(x_{i_m+1})}{n(f(x_{i_m+2}) - f(x_{i_m+1}))} \right) \mathbb{1}\{A_0 \cap B\} \right). \tag{A.2.76}
\end{aligned}$$

Further, since on event  $A_0$ , we have

$$\begin{aligned}
& t_{raw}(i_m + 1) - \frac{f(x_{i_m}) - f(x_{i_m+1})}{n(f(x_{i_m+2}) - f(x_{i_m+1}))} = \\
& \frac{\zeta_{i_m}(f(x_{i_m+2}) - f(x_{i_m+1})) + \zeta_{i_m+1}(f(x_{i_m}) - f(x_{i_m+2}))}{n(f(x_{i_m+2}) - f(x_{i_m+1}) + \zeta_{i_m+2} - \zeta_{i_m+1} + 2\sqrt{6}\sigma z_{\alpha_2})(f(x_{i_m+2}) - f(x_{i_m+1}))} \\
& + \frac{\zeta_{i_m+2}(f(x_{i_m+1}) - f(x_{i_m})) + 2\sqrt{6}\sigma z_{\alpha_2}(f(x_{i_m+2}) - f(x_{i_m}))}{n(f(x_{i_m+2}) - f(x_{i_m+1}) + \zeta_{i_m+2} - \zeta_{i_m+1} + 2\sqrt{6}\sigma z_{\alpha_2})(f(x_{i_m+2}) - f(x_{i_m+1}))} \\
& \leq \left( |\zeta_{i_m}|(f(x_{i_m+2}) - f(x_{i_m+1})) + |\zeta_{i_m+1}|(f(x_{i_m}) - f(x_{i_m+2})) \right. \\
& \quad \left. + |\zeta_{i_m+2}|(f(x_{i_m+1}) - f(x_{i_m})) + 2\sqrt{6}\sigma z_{\alpha_2}(f(x_{i_m+2}) - f(x_{i_m})) \right) \\
& \quad \frac{1}{\frac{2}{3}n(f(x_{i_m+2}) - f(x_{i_m+1}))^2} \\
& \leq \sqrt{6}\sigma \frac{3}{2n} (|\zeta_{i_m}| + 2|\zeta_{i_m+1}| + |\zeta_{i_m+2}| + 4z_{\alpha_2}) \frac{1}{f(x_{i_m+2}) - f(x_{i_m+1})}.
\end{aligned} \tag{A.2.77}$$

Therefore,

$$\begin{aligned}
& \mathbb{E} \left( \left( t_{raw}(i_m + 1) - \frac{f(x_{i_m}) - f(x_{i_m+1})}{n(f(x_{i_m+2}) - f(x_{i_m+1}))} \right) \mathbb{1}_{\{A_0 \cap B\}} \right) \\
& \leq \mathbb{E} \left( \sqrt{6}\sigma \frac{3}{2n} (|\zeta_{i_m}| + 2|\zeta_{i_m+1}| + |\zeta_{i_m+2}| + 4z_{\alpha_2}) \frac{1}{f(x_{i_m+2}) - f(x_{i_m+1})} \mathbb{1}_{\{A_0 \cap B\}} \right) \\
& \leq \mathbb{E} \left( \sqrt{6}\sigma \frac{3}{2n} (|\zeta_{i_m}| + 2|\zeta_{i_m+1}| + |\zeta_{i_m+2}| + 4z_{\alpha_2}) \frac{1}{f(x_{i_m+2}) - f(x_{i_m+1})} \right) \\
& \leq \sqrt{6}\sigma \frac{3}{2n} (4\check{Q}_3 + 4z_{\alpha_2}) \frac{1}{f(x_{i_m+2}) - f(x_{i_m+1})} \\
& \leq \frac{\sigma}{\rho_m(\frac{\sigma}{\sqrt{n}}; \tilde{h})/n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \frac{1}{n} \sqrt{6}(6\check{Q}_3 + 6z_{\alpha_2}) \\
& \leq \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n\rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \sqrt{12}(6\check{Q}_3 + 6z_{\alpha_2}),
\end{aligned} \tag{A.2.78}$$

where  $\check{Q}_3 = \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx$ .



Going back to Equation (A.2.69), we have

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (t_{hi} - i_l/n) \mathbb{1}\{i_l = i_m\} (\mathbb{1}\{i_l = n = U = i_r + 1\} \right. \\
& \quad \left. + \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\}) \right) \\
& \leq \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \sqrt{12(6\check{Q}_3 + 6z_{\alpha_2})} \\
& \quad + \sup\{Z(h) : h(x_i) = f(x_i), 0 \leq i \leq n\} - \frac{i_m}{n}.
\end{aligned} \tag{A.2.79}$$

The first term is bounded for under the case  $(i_m - n + 1)(i_m) \neq 0$ .

Similarly, for the second term in Equation (A.2.68), we have

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (i_l/n - t_{lo}) \mathbb{1}\{i_l = i_m\} (\mathbb{1}\{i_l = n = U = i_r + 1\} \right. \\
& \quad \left. + \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\}) \right) \\
& \leq \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} \sqrt{12(6\check{Q}_3 + 6z_{\alpha_2})} \\
& \quad + \frac{i_m}{n} - \inf\{Z(h) : h(x_i) = f(x_i), 0 \leq i \leq n\}.
\end{aligned} \tag{A.2.80}$$

Therefore, under the case  $(i_m - n + 1)(i_m) \neq 0$ ,

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{\check{j} = \infty\} (t_{hi} - t_{lo}) \mathbb{1}\{i_l = i_m\} (\mathbb{1}\{i_l = n = U = i_r + 1\} \right. \\
& \quad \left. + \mathbb{1}\{i_r = -1 = L - 1 = i_l - 1\} + \mathbb{1}\{L + 1 \leq i_l = i_r + 1 \leq U - 1\}) \right) \\
& \leq \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} 2\sqrt{12(6\check{Q}_3 + 6z_{\alpha_2})} + \\
& \quad \sup\{Z(h) : h(x_i) = f(x_i), 0 \leq i \leq n\} - \inf\{Z(h) : h(x_i) = f(x_i), 0 \leq i \leq n\} \\
& = \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h}) \sqrt{n \rho_z(\frac{\sigma}{\sqrt{n}}; \tilde{h})} 2\sqrt{12(6\check{Q}_3 + 6z_{\alpha_2})} + \mathfrak{D}_z(n, f).
\end{aligned} \tag{A.2.81}$$

All the cases analyzed, and all the terms added up

$$\begin{aligned}
& \mathbb{E}(\mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{t_{hi} - t_{lo} < \frac{3}{n}\} L(\mathbf{CI}_{z,\alpha}(Y))) \\
& \leq \check{c}_{3,\alpha} \sup_{h \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right) \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)} + 2\mathfrak{D}_z(n, f).
\end{aligned} \tag{A.2.82}$$

□

*Proof of Lemma A.1.34.*

$$\begin{aligned}
P(\hat{\mathbf{j}} \geq \mathbf{j}^{\mathfrak{w}} + K + 1) &= \mathbb{E}(\mathbb{1}\{\hat{\mathbf{j}} \geq \mathbf{j}^{\mathfrak{w}} + K + 1\} \mathbb{1}\{\mathbf{j}^{\mathfrak{w}} < \infty\}) \\
&\leq \mathbb{E}(\mathbb{1}\{\forall \mathbf{j}^{\mathfrak{w}} + 1 \leq j \leq \mathbf{j}^{\mathfrak{w}} + K, \\
&\quad \min\{Y_{j, \hat{\mathbf{i}}_j - 6, s} - Y_{j, \hat{\mathbf{i}}_j - 5, s}, Y_{j, \hat{\mathbf{i}}_j + 6, s} - Y_{j, \hat{\mathbf{i}}_j + 5, s}\} > 2\gamma_s \sqrt{2\sigma} \sqrt{2^{J-j}}\} \mathbb{1}\{\mathbf{j}^{\mathfrak{w}} < \infty\}) \\
&\leq \Phi(-2)^K \mathbb{E}(\mathbb{1}\{\mathbf{j}^{\mathfrak{w}} < \infty\}) \leq \Phi(-2)^K
\end{aligned} \tag{A.2.83}$$

The second inequality is by taking conditional expectation on the localization copy of the observation (i.e.  $\mathbf{Y}_l$ ), and the fact that for the iteration steps  $j$  such that  $\mathbf{j}^{\mathfrak{w}} + 1 \leq j \leq \mathbf{j}^{\mathfrak{w}} + K$  the target interval is more than 6 blocks away from the estimated one. □

*Proof of Lemma A.1.35.* Given the symmetric nature of our procedure, we only need to prove

$$\mathbb{E}(\mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1^c\}) \leq \alpha_1. \tag{A.2.84}$$

Note that, when  $\check{j} = \infty$ ,

$$\begin{aligned}
E &= \{Z(f) \in [(\hat{\mathbf{i}}_{\check{\mathbf{j}}} - (6 \cdot 2^{K_{\alpha/2+1}} - 2) - 1) \frac{2^{J-\check{\mathbf{j}}}}{n} - \frac{1}{2n}, \\
&\quad (\hat{\mathbf{i}}_{\check{\mathbf{j}}} + (6 \cdot 2^{K_{\alpha/2+1}} - 2)) \frac{2^{J-\check{\mathbf{j}}}}{n} - \frac{1}{2n}] \cap [0, 1]\} \\
&\subset \left\{ \frac{\hat{\mathbf{i}}_{\check{\mathbf{j}}} - (6 \cdot 2^{K_{\alpha/2+1}} - 2) - 2}{n} < Z(f) < \frac{\hat{\mathbf{i}}_{\check{\mathbf{j}}} + 6 \cdot 2^{K_{\alpha/2+1}} - 2}{n} \right\}
\end{aligned}$$

Let  $L_0 = \hat{\mathbf{i}}_{\check{\mathbf{j}}} - (6 \cdot 2^{K_{\alpha/2+1}} - 2) - 2$ ,  $U_0 = \hat{\mathbf{i}}_{\check{\mathbf{j}}} + 6 \cdot 2^{K_{\alpha/2+1}} - 2$ . Hence we know that when

$L_0 \geq 1$ ,  $L = L_0 - 1$ ; when  $U_0 \leq n - 1$ ,  $U = U_0 + 1$ .

Let  $i_m = \min\{k : f(x_k) = \min\{f(x_i) : 0 \leq i \leq n\}\}$ . Then we know that, on  $E$ ,  $L_0 \leq i_m \leq U_0$ . And also  $i_m = n$  implies  $F_1$ , hence we only need to consider the case  $i_m \leq n - 1$  to compute  $F_1^c$ . And  $\{i_m \leq n - 1\} \cap \{L_0 \leq i_m \leq U_0\}$  implies that  $i_m < U$ .

We also know that  $\{y_{e,i} + \sqrt{3}\sigma z_{3,i} : 0 \leq i \leq n\}$ ,  $\{y_{e,i} - \sqrt{3}\sigma z_{3,i} : 0 \leq i \leq n\}$ ,  $\{y_{s,i} : 0 \leq i \leq n\}$ ,  $\{y_{l,i} : 0 \leq i \leq n\}$  are independent random variables.

Therefore,

$$\begin{aligned} \mathbb{E}(\mathbb{1}\{E\}\mathbb{1}\{\check{j} = \infty\}\mathbb{1}\{F_1^c\}) &\leq \mathbb{E}\left(\mathbb{E}\left(\mathbb{1}\{E\}\mathbb{1}\{\check{j} = \infty\}\mathbb{1}\{i_m < U\}\right.\right. \\ &\quad \left.\left.\mathbb{1}\{y_{e,i_m} + \sqrt{3}\sigma z_{3,i_m} - (y_{e,i_m+1} + \sqrt{3}\sigma z_{3,i_m+1}) > 2\sqrt{3}\sigma z_{\alpha_1}\}\middle|\mathbf{y}_s, \mathbf{y}_l\right)\right) \leq \alpha_1. \end{aligned} \quad (\text{A.2.85})$$

□

*Proof of Lemma A.1.36.* The event  $E \cap \{\check{j} = \infty\} \cap F_1 \cap F_2 \cap \{(i_l - U)(i_r - L + 1) = 0\}$  is the subset of the union of the following four events:

$$\begin{aligned} G_1 &= E \cap \{\check{j} = \infty\} \cap F_1 \cap F_2 \cap \{i_l = U, U \neq n\}, \\ G_2 &= E \cap \{\check{j} = \infty\} \cap F_1 \cap F_2 \cap \{i_l = U, U = n\}, \\ G_3 &= E \cap \{\check{j} = \infty\} \cap F_1 \cap F_2 \cap \{i_r = L - 1, L = 0\}, \\ G_4 &= E \cap \{\check{j} = \infty\} \cap F_1 \cap F_2 \cap \{i_r = L - 1, L \neq 0\}. \end{aligned} \quad (\text{A.2.86})$$

Since  $\{U \neq n\} \cap \{\check{j} = \infty\}$  means  $U_0 \leq U - 1 \leq n - 2$ ; and on  $E \cap \{\check{j} = \infty\} \cap F_1 \cap F_2$  we have  $i_l \leq \min\{k : f(x_k) = \min\{f(x_i)\}\}$  and  $\min\{k : f(x_k) = \min\{f(x_i)\}\} \leq U_0$ , we know that  $G_1 = \emptyset$ . Similarly, we have  $G_4 = \emptyset$ . Also, on  $E \cap \{\check{j} = \infty\} \cap F_1 \cap F_2$ , we know that  $i_l \leq i_r + 1$ , hence we have  $G_2 \cap G_3 = \emptyset$ .

Also, on  $G_2$ , we know that  $f(x_n) = \min\{f(x_i)\}$  and  $f(x_k) > \min\{f(x_i) : 0 \leq i \leq n\}$  for all

k, which implies that  $Z(f) \geq \frac{f(x_n) - f(x_{n-1})}{n(f(x_{n-2}) - f(x_{n-1}))} + \frac{n-1}{n}$ .

Suppose  $Y_{e,1} = \{y_{e,i} + \sqrt{3}\sigma z_{3,i} : (L-1) \vee 0 \leq i \leq (U+1) \wedge n\}$ ,  $Y_{e,2} = \{y_{e,i} - \sqrt{3}\sigma z_{3,i} : (L-1) \vee 0 \leq i \leq (U+1) \wedge n\}$ . Then we know that  $\mathbf{Y}_l, \mathbf{Y}_s, Y_{e,1}, Y_{e,2}$  are independent.

If we denote  $\kappa_{i,1} = y_{e,i} + \sqrt{3}\sigma z_{3,i} - f(x_i)$ ,  $\kappa_{i,2} = y_{e,i} - \sqrt{3}\sigma z_{3,i} - f(x_i)$ , then we know that on  $G_2$  when we further have  $\kappa_{n,2} \geq -\sqrt{6}\sigma z_{\alpha_2}$ ,  $\kappa_{n-1,2} \leq \sqrt{6}\sigma z_{\alpha_2}$ ,  $\kappa_{n-2,2} \geq -\sqrt{6}\sigma z_{\alpha_2}$ , then  $t_{lo} \leq Z(f)$ .

We have similar analysis for  $G_3$ .

Hence we know that

$$\begin{aligned}
& \mathbb{E}\left(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) = 0\}\right) \\
&= \mathbb{E}\left(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{G_2\}\right) \\
&\quad + \mathbb{E}\left(\mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{G_3\}\right) \\
&= \mathbb{E}\left(\mathbb{E}(\mathbb{1}\{t_{lo} > Z(f)\} | \mathbf{Y}_l, \mathbf{Y}_s, Y_{e,1}) \mathbb{1}\{G_2\}\right) + \mathbb{E}\left(\mathbb{E}(\mathbb{1}\{t_{hi} < Z(f)\} | \mathbf{Y}_l, \mathbf{Y}_s, Y_{e,1}) \mathbb{1}\{G_3\}\right) \\
&\leq 3\alpha_2 P(G_2) + 3\alpha_2 P(G_3) \\
&\leq 3\alpha_2 P(\mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) = 0\}).
\end{aligned}
\tag{A.2.87}$$

□

*Proof of Lemma A.1.37.*

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
& \quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \right) \\
& \leq \mathbb{E} \left( \mathbb{1}\{Z(f) > t_{hi}\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
& \quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \right) \\
& + \mathbb{E} \left( \mathbb{1}\{Z(f) < t_{lo}\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
& \quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \right). \tag{A.2.88}
\end{aligned}$$

Given the symmetric nature of the procedure, we only need to bound the first term, the second term shares the same bound.

Suppose  $Y_{e,1} = \{y_{e,i} + \sqrt{3}\sigma z_{3,i} : (L-1) \vee 0 \leq i \leq (U+1) \wedge n\}$ ,  $Y_{e,2} = \{y_{e,i} - \sqrt{3}\sigma z_{3,i} : (L-1) \vee 0 \leq i \leq (U+1) \wedge n\}$ . Then we know that  $Y_l, Y_s, Y_{e,1}, Y_{e,2}$  are independent.

On the event  $E \cap \{\check{j} = \infty\} \cap F_1 \cap F_2 \cap \{(i_l - U)(i_r - L + 1) \neq 0\} \cap \{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\}$ , we know that  $|\{k : f(x_k) = \min\{f(x_i) : 0 \leq i \leq n\}\}| = 1$ , we denote this unique element to be  $i_m$ . Also, on this event, we know that  $2 \leq i_m \leq n - 2$ . Hence we know that  $Z(f) \leq \frac{f(x_{i_m}) - f(x_{i_m+1})}{(f(x_{i_m+2}) - f(x_{i_m+1}))/\frac{1}{n}} + \frac{i_m+1}{n}$ . If we denote  $\kappa_{i,1} = y_{e,i} + \sqrt{3}\sigma z_{3,i} - f(x_i)$ ,  $\kappa_{i,2} = y_{e,i} - \sqrt{3}\sigma z_{3,i} - f(x_i)$ , then we know that on event  $E \cap \{\check{j} = \infty\} \cap F_1 \cap F_2 \cap \{(i_l - U)(i_r - L + 1) \neq 0\} \cap \{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\}$ , if we further have  $\kappa_{i_m+2,2} \geq -\sqrt{6}\sigma z_{\alpha_2}$ ,  $\kappa_{i_m+1,2} \leq \sqrt{6}\sigma z_{\alpha_2}$ ,  $\kappa_{i_m,2} \geq -\sqrt{6}\sigma z_{\alpha_2}$ , then  $Z(f) \leq t_{hi}$ .

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{Z(f) > t_{hi}\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
& \quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \right) \\
&= \mathbb{E} \left( \mathbb{E}(\mathbb{1}\{Z(f) > t_{hi}\} | \mathbf{Y}_l, \mathbf{Y}_s, Y_{e,1}) \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \right. \\
& \quad \left. \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \right) \\
&= \mathbb{E} \left( \mathbb{E}(\mathbb{1}\{Z(f) > t_{hi}\} | \mathbf{Y}_l, \mathbf{Y}_s, Y_{e,1}) \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
& \quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \mathbb{1}\{i_{lo} + 1 = i_{hi} - 1 = i_m\} \right) \\
&\leq \mathbb{E} \left( \mathbb{E}(\mathbb{1}\{\kappa_{i_m+2,2} < -\sqrt{6}\sigma z_{\alpha_2} \text{ or } \kappa_{i_m+1,2} > \sqrt{6}\sigma z_{\alpha_2} \text{ or } \kappa_{i_m,2} < -\sqrt{6}\sigma z_{\alpha_2}\} | \mathbf{Y}_l, \mathbf{Y}_s, Y_{e,1}) \right. \\
& \quad \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \\
& \quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \mathbb{1}\{i_{lo} + 1 = i_{hi} - 1 = i_m\} \right) \\
&\leq 3\alpha_2 \mathbb{E}(\mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \\
& \quad \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \mathbb{1}\{i_{lo} + 1 = i_{hi} - 1 = i_m\}).
\end{aligned} \tag{A.2.89}$$

Therefore,

$$\begin{aligned}
& \mathbb{E} \left( \mathbb{1}\{Z(f) \notin \mathbf{CI}_{z,\alpha}(Y)\} \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
& \quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \right) \\
&\leq 6\alpha_2 \mathbb{E} \left( \mathbb{1}\{E\} \mathbb{1}\{\check{j} = \infty\} \mathbb{1}\{F_1 \cap F_2\} \mathbb{1}\{(i_l - U)(i_r - L + 1) \neq 0\} \right. \\
& \quad \left. \mathbb{1}\{i_{hi} - i_{lo} \leq 2, 0 < i_{lo}, i_{hi} < n\} \mathbb{1}\{i_{lo} + 1 = i_{hi} - 1 = i_m\} \right).
\end{aligned} \tag{A.2.90}$$

□

*Proof of Lemma A.1.38.*

$$\begin{aligned}
& \mathbb{E}((\mathfrak{E}_{\check{j}, \check{i}_{\check{j}}, e} \frac{1}{2^{J-\check{j}}})^2 \mathbb{1}\{\check{j} < \infty\}) \\
&= \mathbb{E}(\frac{1}{2^{J-\check{j}}} \sigma^2 \gamma_e^2 \mathbb{1}\{\check{j} < \infty\}) = \sigma^2 \gamma_e^2 2^{j^*-J} \mathbb{E}(2^{-j^*+j} \mathbb{1}\{\check{j} < \infty\}) \\
&= \sigma^2 \gamma_e^2 2^{j^*-J} (\sum_{j=1}^{j^*+2} \mathbb{E}(2^{-j^*+j} \mathbb{1}\{\check{j} = j\}) + \sum_{j=j^*+3}^{\infty} \mathbb{E}(2^{-j^*+j} \mathbb{1}\{\check{j} = j\})) \\
&\leq \sigma^2 \gamma_e^2 2^{j^*-J} \left( 4 + \sum_{j=j^*+3}^{\infty} 2^{-j^*+j} \Phi\left(-2 + \frac{\frac{13}{16} \rho_m(\frac{\sigma}{\sqrt{n}}; f) \sqrt{2^{J-j^*-2}}}{\sigma \gamma_s \sqrt{2}}\right) \right. \\
&\quad \left. \Phi\left(-2 + \frac{\frac{13}{32} \rho_m(\frac{\sigma}{\sqrt{n}}; f) \sqrt{2^{J-j^*-3}}}{\sigma \gamma_s \sqrt{2}}\right)^{(j-j^*-3)+} \right) \\
&\leq \sigma^2 \gamma_e^2 2^{j^*-J} \left( 4 + \sum_{j=j^*+3}^{\infty} 2^{-j^*+j} \Phi\left(-2 + \frac{13\sqrt{3}}{\gamma_s 16\sqrt{2}} 2^{\frac{-4}{2}}\right) \Phi\left(-2 + \frac{13\sqrt{3}}{\gamma_s 32\sqrt{2}} 2^{\frac{-5}{2}}\right)^{(j-j^*-3)+} \right. \\
&\leq \sigma^2 \gamma_e^2 2^{j^*-J} \left( 4 + \frac{8\Phi\left(-2 + \frac{13\sqrt{3}}{\gamma_s 16\sqrt{2}} 2^{\frac{-4}{2}}\right)}{1 - \Phi\left(-2 + \frac{13\sqrt{3}}{\gamma_s 32\sqrt{2}} 2^{\frac{-5}{2}}\right)} \right) \\
&\leq 2n\rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2 \rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right) \gamma_e^2 \frac{8}{n\rho_z\left(\frac{\sigma}{\sqrt{n}}; f\right)} \left( 4 + 8 \frac{\Phi\left(-2 + \frac{13\sqrt{3}}{\gamma_s 16\sqrt{2}} 2^{\frac{-4}{2}}\right)}{1 - \Phi\left(-2 + \frac{13\sqrt{3}}{\gamma_s 32\sqrt{2}} 2^{\frac{-5}{2}}\right)} \right) \\
&= c_{m1} \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2.
\end{aligned}
\tag{A.2.91}$$

□

*Proof of Lemma A.1.39.*

$$\begin{aligned}
& \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\check{j} < \infty\}) \\
&= \mathbb{E}\left((\hat{\mathbf{f}} - M(f))^2 (\mathbb{1}\{\tilde{j} > \check{j}\} + \mathbb{1}\{\tilde{j} \leq \check{j}\}) \mathbb{1}\{\check{j} < \infty\}\right) \\
&= \sum_{j_1=2}^{j^*+1} \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\tilde{j} > \check{j} = j_1\}) + \sum_{j_1=j^*+2}^{\infty} \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\tilde{j} > \check{j} = j_1\}) \\
&\quad + \mathbb{E}\left(\left((\hat{\mathbf{f}} - \text{ave}_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}))_+ + (\text{ave}_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - M(f))\right)^2 \mathbb{1}\{\tilde{j} \leq \check{j} = j_1\} \mathbb{1}\{\check{j} < \infty\}\right) \quad (\text{A.2.92}) \\
&\leq \sum_{j_1=2}^{j^*+1} \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\tilde{j} > \check{j} = j_1\}) + \sum_{j_1=j^*+2}^{\infty} \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\tilde{j} > \check{j} = j_1\}) \\
&\quad + 2\mathbb{E}\left(\left((\hat{\mathbf{f}} - \text{ave}_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}))_+\right)^2 \mathbb{1}\{\tilde{j} \leq \check{j}\} \mathbb{1}\{\check{j} < \infty\}\right) + \\
&\quad 2\mathbb{E}\left(\left(\text{ave}_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - M(f)\right)^2 \mathbb{1}\{\tilde{j} \leq \check{j}\} \mathbb{1}\{\check{j} < \infty\}\right).
\end{aligned}$$

We will have four lemmas to bound each term respectively. To avoid distraction, we will defer the proofs of the lemmas to later part.

**Lemma A.2.1.**

$$\sum_{j_1=2}^{j^*+1} \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\tilde{j} > \check{j} = j_1\}) \leq c_{m3} \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2. \quad (\text{A.2.93})$$

**Lemma A.2.2.**

$$\sum_{j_1=j^*+2}^{\infty} \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\tilde{j} > \check{j} = j_1\}) \leq c_{m4} \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2. \quad (\text{A.2.94})$$

**Lemma A.2.3.**

$$\mathbb{E}\left(\left((\hat{\mathbf{f}} - \text{ave}_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}))_+\right)^2 \mathbb{1}\{\tilde{j} \leq \check{j}\} \mathbb{1}\{\check{j} < \infty\}\right) \leq c_{m5} \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2. \quad (\text{A.2.95})$$

**Lemma A.2.4.**

$$\mathbb{E}\left(\left(\text{ave}_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - M(f)\right)^2 \mathbb{1}\{\tilde{j} \leq \check{j}\} \mathbb{1}\{\check{j} < \infty\}\right) \leq c_{m6} \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2. \quad (\text{A.2.96})$$



With these four lemmas, we know that

$$\mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}_{\{\check{j} < \infty\}}) \leq (c_{m3} + c_{m4} + 2c_{m5} + 2c_{m6})\rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2 = c_{m2}\rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2. \quad (\text{A.2.97})$$

Now we will prove these four lemmas, For simplicity, we will not repeatedly write  $\mathbb{1}_{\{\check{j} < \infty\}}$ , in the expectation, but that is the default assumption whenever  $\check{j}$  appears.

**Proof of Lemma A.2.1** Similarly to the white noise model, we have

$$\begin{aligned} & \sum_{j_1=2}^{j^*+1} \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}_{\{\tilde{\mathbf{j}} > \check{j} = j_1\}}) \\ & \leq \sum_{j_1=2}^{j^*+1} \mathbb{E} \left( \frac{3}{2} ((ave_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) - M(f))^2 \mathbb{1}_{\{Y_{j_1, \hat{\mathbf{i}}_{j_1}+6,s} - Y_{j_1, \hat{\mathbf{i}}_{j_1}+5,s} \leq 2\sqrt{2}\gamma_s\sigma\sqrt{2^{J-j_1}}\}} \right. \\ & \quad \left. + (ave_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) - M(f))^2 \mathbb{1}_{\{Y_{j_1, \hat{\mathbf{i}}_{j_1}+6,s} - Y_{j_1, \hat{\mathbf{i}}_{j_1}+5,s} \leq 2\sqrt{2}\gamma_s\sigma\sqrt{2^{J-j_1}}\}}) \mathbb{1}_{\{\tilde{\mathbf{j}} > j_1\}} \right) \\ & \leq \sum_{j_1=2}^{j^*+1} \frac{3}{2} ((ave_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) - M(f))^2 \Phi(2 - \frac{(ave_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) - M(f))2^{\frac{1}{2}(J-j_1)}}{3.5\sigma\gamma_s\sqrt{2}}) \\ & \quad + (ave_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) - M(f))^2 \Phi(2 - \frac{(ave_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) - M(f))2^{\frac{1}{2}(J-j_1)}}{3.5\sigma\gamma_s\sqrt{2}})) \mathbb{E}(\mathbb{1}_{\{\tilde{\mathbf{j}} > j_1\}}) \\ & \leq \sum_{j_1=2}^{j^*+1} 3 \cdot 2^{j_1-J} (3.5\sqrt{2}\sigma\gamma_s)^2 V \\ & \leq 6 \times 2^{j^*+1-J} \sigma^2 \frac{49}{2} \gamma_s^2 V \\ & \leq 48 \times 49 \times 2\gamma_s^2 V \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2 = c_{m3}\rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2. \end{aligned} \quad (\text{A.2.98})$$

The  $V$  in the inequalities are still the same as the  $V$  in the white noise model:

$$V = \max_{x>0} x^2 \Phi(2-x).$$

□

*Proof of Lemma A.2.2.*

$$\begin{aligned}
& \sum_{j_1=j^*+2}^{\infty} \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\tilde{\mathbf{j}} > \check{\mathbf{j}} = j_1\}) \\
& \leq \sum_{j_1=j^*+2}^{\infty} \mathbb{E} \left( \frac{3}{2} \left( (ave_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) - M(f))^2 \mathbb{1}\{Y_{j_1, \hat{\mathbf{i}}_{j_1}+6, s} - Y_{j_1, \hat{\mathbf{i}}_{j_1}+5, s} \leq 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j_1}}\} \right. \right. \\
& \quad \mathbb{1}\{\forall j^* + 1 \leq j \leq j_1 - 1, \min\{Y_{j, \hat{\mathbf{i}}_j+6, s} - Y_{j, \hat{\mathbf{i}}_j+5, s}, Y_{j, \hat{\mathbf{i}}_j-6, s} - Y_{j, \hat{\mathbf{i}}_j-5, s}\} > 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j_1}}\} \\
& \quad + (ave_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) - M(f))^2 \mathbb{1}\{Y_{j_1, \hat{\mathbf{i}}_{j_1}+6, s} - Y_{j_1, \hat{\mathbf{i}}_{j_1}+5, s} \leq 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j_1}}\} \\
& \quad \left. \left. \mathbb{1}\{\forall j^* + 1 \leq j \leq j_1 - 1, \min\{Y_{j, \hat{\mathbf{i}}_j+6, s} - Y_{j, \hat{\mathbf{i}}_j+5, s}, Y_{j, \hat{\mathbf{i}}_j-6, s} - Y_{j, \hat{\mathbf{i}}_j-5, s}\} > 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j_1}}\} \right) \right) \mathbb{1}\{\tilde{\mathbf{j}} > j_1\} \Big) \\
& \leq \sum_{j_1=j^*+2}^{\infty} \mathbb{E} \left( \frac{3}{2} \left( (ave_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) - M(f))^2 \Phi \left( 2 - \frac{(ave_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) - M(f)) 2^{\frac{1}{2}(J-j_1)}}{3.5\sigma\gamma_s\sqrt{2}} \right) \right. \right. \\
& \quad \Phi \left( -2 + \frac{\rho_m(\frac{\sigma}{\sqrt{n}}; f) 2^{\frac{1}{2}(J-j^*-1)}}{\sigma\sqrt{2}\gamma_s} \right) \Phi \left( -2 + \frac{\rho_m(\frac{\sigma}{\sqrt{n}}; f) 2^{\frac{1}{2}(J-j^*-2)}}{\sigma 2\sqrt{2}\gamma_s} \right)_{(j_1-j^*-2)+} \\
& \quad + (ave_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) - M(f))^2 \Phi \left( 2 - \frac{(ave_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) - M(f)) 2^{\frac{1}{2}(J-j_1)}}{3.5\sigma\gamma_s\sqrt{2}} \right) \\
& \quad \Phi \left( -2 + \frac{\rho_m(\frac{\sigma}{\sqrt{n}}; f) 2^{\frac{1}{2}(J-j^*-1)}}{\sigma\sqrt{2}\gamma_s} \right) \Phi \left( -2 + \frac{\rho_m(\frac{\sigma}{\sqrt{n}}; f) 2^{\frac{1}{2}(J-j^*-2)}}{\sigma 2\sqrt{2}\gamma_s} \right)_{(j_1-j^*-2)+} \\
& \quad \left. \right) \mathbb{E}(\mathbb{1}\{\tilde{\mathbf{j}} > j_1\}) \Big) \\
& \leq \sum_{j_1=j^*+2}^{\infty} 3 \cdot 2^{j_1-J} (3.5\sqrt{2}\sigma\gamma_s)^2 V \cdot \Phi \left( -2 + \frac{\sqrt{3}}{4\gamma_s} \right) \Phi \left( -2 + \frac{\sqrt{3}}{8\sqrt{2}\gamma_s} \right)_{(j_1-j^*-2)+} \\
& \leq 3 \cdot 2^{j^*+2-J} \frac{49}{2} \sigma^2 \gamma_s^2 V \Phi \left( -2 + \frac{1}{4} \right) \frac{1}{1 - 2\Phi(-1.9)} \\
& \leq c_{m4} \rho_m \left( \frac{\sigma}{\sqrt{n}}; f \right)^2.
\end{aligned}$$

(A.2.99)

□

*Proof of Lemma A.2.3.*

$$\begin{aligned}
& \mathbb{E}(((\hat{\mathbf{f}} - \text{ave}_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}))_+)^2 \mathbb{1}\{\check{j} \leq j < \infty\}) \\
&= \sum_{j_2=2}^J \sum_{j_1=j_2}^J \mathbb{E} \left( ((\hat{\mathbf{f}} - \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1}))_+)^2 \mathbb{1}\{\check{j} = j_2\} \mathbb{1}\{\check{j} = j_1\} \right) \\
&\leq \sum_{j_2=2}^J \sum_{j_1=j_2}^J \mathbb{E} \left( \mathbb{1}\{\check{j} = j_2\} ((\text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) - \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1}))_+)^2 \right. \\
&\quad \mathbb{1}\{Y_{j_1, \hat{\mathbf{i}}_{j_1+6}, s} - Y_{j_1, \hat{\mathbf{i}}_{j_1+5}, s} \leq 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j_1}}\} \\
&\quad \mathbb{1}\{\forall j^* + 2 \leq j \leq j_1 - 1, Y_{j_1, \hat{\mathbf{i}}_{j_1-6}, s} - Y_{j_1, \hat{\mathbf{i}}_{j_1-5}, s} > 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j_1}}, \\
&\quad Y_{j_1, \hat{\mathbf{i}}_{j_1+6}, s} - Y_{j_1, \hat{\mathbf{i}}_{j_1+5}, s} > 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j_1}}, \text{ if exists}\} \\
&\quad + \mathbb{1}\{\check{j} = j_2\} ((\text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) - \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1}))_+)^2 \\
&\quad \mathbb{1}\{Y_{j_1, \hat{\mathbf{i}}_{j_1-6}, s} - Y_{j_1, \hat{\mathbf{i}}_{j_1-5}, s} \leq 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j_1}}\} \\
&\quad \mathbb{1}\{\forall j^* + 2 \leq j \leq j_1 - 1, Y_{j_1, \hat{\mathbf{i}}_{j_1-6}, s} - Y_{j_1, \hat{\mathbf{i}}_{j_1-5}, s} > 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j_1}}, \\
&\quad Y_{j_1, \hat{\mathbf{i}}_{j_1+6}, s} - Y_{j_1, \hat{\mathbf{i}}_{j_1+5}, s} > 2\sqrt{2}\gamma_s \sigma \sqrt{2^{J-j_1}}, \text{ if exists}\} \Big) \\
&\leq \sum_{j_2=2}^J \sum_{j_1=j_2}^J \mathbb{E} \left( (\text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) - \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1}))^2 \mathbb{1}\{\text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) - \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1}) > 0\} \right. \\
&\quad \Phi \left( 2 - \frac{(\text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1+2}) - \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1})) \sqrt{2^{J-j_1}}}{2\sqrt{2}\gamma_s \sigma} \right) \Phi \left( -2 + \frac{\frac{13}{16}\rho_m(\frac{\sigma}{\sqrt{n}}; f) \sqrt{2^{J-j^*-2}}}{\sigma\gamma_s \sqrt{2}} \right)_{(j_1-j^*-2)+} \\
&\quad \mathbb{1}\{\check{j} = j_2\} \\
&\quad + (\text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) - \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1}))^2 \mathbb{1}\{\text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) - \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1}) > 0\} \\
&\quad \Phi \left( 2 - \frac{(\text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1-2}) - \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1})) \sqrt{2^{J-j_1}}}{2\sqrt{2}\gamma_s \sigma} \right) \Phi \left( -2 + \frac{\frac{13}{16}\rho_m(\frac{\sigma}{\sqrt{n}}; f) \sqrt{2^{J-j^*-2}}}{\sigma\gamma_s \sqrt{2}} \right)_{(j_1-j^*-2)+} \\
&\quad \mathbb{1}\{\check{j} = j_2\} \Big)
\end{aligned}$$

(A.2.100)

$$\begin{aligned}
&\leq \sum_{j_2=2}^J \sum_{j_1=j_2}^J \mathbb{E} \left( \mathbb{1}\{ave_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) > ave_f(j_1, \hat{\mathbf{i}}_{j_1})\} \right. \\
&\quad 2^{3+j_1-J} \gamma_s^2 \sigma^2 V \mathbb{1}\{\tilde{\mathbf{j}} = j_2\} \Phi(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s})^{(j_1-\tilde{\mathbf{j}}^*-2)_+} \\
&\quad + \mathbb{1}\{ave_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) > ave_f(j_1, \hat{\mathbf{i}}_{j_1})\} \\
&\quad \left. 2^{3+j_1-J} \gamma_s^2 \sigma^2 V \mathbb{1}\{\tilde{\mathbf{j}} = j_2\} \Phi(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s})^{(j_1-\tilde{\mathbf{j}}^*-2)_+} \right) \\
&\leq 2 \sum_{j_2=2}^J \mathbb{E}(\mathbb{1}\{\tilde{\mathbf{j}} = j_2\}) \gamma_s^2 \sigma^2 V 2^{5+\tilde{\mathbf{j}}^*-J} (1 + \frac{1}{1 - 2\Phi(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s})}) \\
&\leq c_{m5} \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2.
\end{aligned}$$

□

*Proof of Lemma A.2.4.* For simplicity, in this proof, we take  $\check{j} < \infty$  by default. However, this is not a key condition, we only need this to establish that  $\tilde{\mathbf{j}} \leq J$  and  $\tilde{\mathbf{j}} \leq \hat{\mathbf{j}}$ .

$$\begin{aligned}
&\mathbb{E}((ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - M(f))^2 \mathbb{1}\{\tilde{\mathbf{j}} \leq \check{j}\}) \\
&\leq 2\mathbb{E} \left( \left( (ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}))_+ \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} \leq \check{j}\} \right) \\
&\quad + 2\mathbb{E} \left( \left( (ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}) - M(f))_+ \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} \leq \check{j}\} \right).
\end{aligned} \tag{A.2.101}$$

Now we introduce two lemmas that we will prove later.

**Lemma A.2.5.**

$$\mathbb{E} \left( \left( (ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}))_+ \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} \leq \check{j}\} \right) \leq c_{m7} \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2. \tag{A.2.102}$$

**Lemma A.2.6.**

$$\mathbb{E}\left(\left((ave_f(\tilde{j}, \hat{i}_{\tilde{j}}) - M(f))_+\right)^2 \mathbb{1}\{\tilde{j} \leq \check{j}\}\right) \leq c_{m8} \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2. \quad (\text{A.2.103})$$

With these two lemmas, we have

$$\mathbb{E}((ave_f(\check{j}, \hat{i}_{\check{j}}) - M(f))^2 \mathbb{1}\{\check{j} \leq \tilde{j}\}) \leq 2(c_{m7} + c_{m8}) \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2 = c_{m6} \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2. \quad (\text{A.2.104})$$

□

*Proof of Lemma A.2.5 .* Similar to the white noise model, we will first define the following events to describe the relative location of one iteration further compared to the current one at step  $\tilde{j} + r$ :

$$\begin{aligned} \tilde{A}_r &= \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, \hat{i}_{\tilde{j}+r+1} = 2\hat{i}_{\tilde{j}+r+1} - 2\} \\ &\cup \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, \hat{i}_{\tilde{j}+r+1} = 2\hat{i}_{\tilde{j}+r+1} + 1\}, \\ \tilde{B}_r &= \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, \hat{i}_{\tilde{j}+r+1} = 2\hat{i}_{\tilde{j}+r+1} - 1\} \\ &\cup \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, \hat{i}_{\tilde{j}+r+1} = 2\hat{i}_{\tilde{j}+r+1}\}, \\ \tilde{C}_r &= \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, \hat{i}_{\tilde{j}+r+1} = 2\hat{i}_{\tilde{j}+r+1}\} \\ &\cup \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, \hat{i}_{\tilde{j}+r+1} = 2\hat{i}_{\tilde{j}+r+1} - 1\}, \\ \tilde{D}_r &= \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, \hat{i}_{\tilde{j}+r+1} = 2\hat{i}_{\tilde{j}+r+1} + 1\} \\ &\cup \{\omega : \hat{i}_{\tilde{j}+r} < i_{\tilde{j}+r}^*, \hat{i}_{\tilde{j}+r+1} = 2\hat{i}_{\tilde{j}+r+1} - 2\}. \end{aligned} \quad (\text{A.2.105})$$

Basically, from  $\tilde{A}_r$  to  $\tilde{D}_r$ , the average of the signal are from the highest to the lowest.

Then we have

$$\begin{aligned}
& \mathbb{E} \left( \left( (ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}))_+ \right)^2 \mathbb{1}\{\tilde{j} \leq \check{j}\} \right) \\
&= \mathbb{E} \left( \left( (ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}))_+ \right)^2 \mathbb{1}\{\tilde{j} + 1 \leq \check{j}\} \right) \\
&\leq \mathbb{E} \left( \left( (ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}))_+ \right)^2 \mathbb{1}\{\tilde{j} + 1 \leq \check{j}\} \right. \\
&\quad \left( \mathbb{1}\{\tilde{A}_0 \cup (\tilde{B}_0 \cap \tilde{D}_1^c) \cup (\tilde{B}_0 \cap \tilde{D}_1 \cap \{\tilde{j} = \tilde{j} + 1\})\} \right. \\
&\quad \left. \left. + \mathbb{1}\{(\tilde{B}_0 \cap \tilde{D}_1) \cup (\tilde{C}_0 \cap \tilde{B}_1)\} \mathbb{1}\{\check{j} \geq \tilde{j} + 3\} + \mathbb{1}\{\tilde{C}_0 \cap \tilde{A}_1\} \right) \right). \tag{A.2.106}
\end{aligned}$$

We will bound the three terms separately, before that, only for bounding these three terms, denote  $\delta = \mathbb{1}\{j_1 = j_2 + 1\}$ ,  $\delta_0 = \mathbb{1}\{j = j_2\}$ .

$$\begin{aligned}
& \mathbb{E} \left( (ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}))_+ \mathbb{1}\{\tilde{j} + 1 \leq \check{j}\} (\mathbb{1}\{\tilde{A}_0 \cup (\tilde{B}_0 \cap \tilde{D}_1^c) \cup (\tilde{B}_0 \cap \tilde{D}_1 \cap \{\tilde{j} = \tilde{j} + 1\})\}) \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j_1=j_2+1}^{\infty} \mathbb{E} \left( 2 \sum_{j=j_2}^{j_1-1} 2^{j-j_2} \left( (ave_f(j+1, \hat{\mathbf{i}}_{j+1}) - ave_f(j, \hat{\mathbf{i}}_j))_+ \right)^2 \right. \\
&\quad \left. \mathbb{1}\{\check{j} = j_1, \tilde{j} = j_2\} (\mathbb{1}\{\tilde{A}_0 \cup (\tilde{B}_0 \cap \tilde{D}_1^c)\} + \mathbb{1}\{\tilde{B}_0 \cap \tilde{D}_1\} \mathbb{1}\{j_1 = j_2 + 1\}) \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j_1=j_2+1}^{\infty} \sum_{j=j_2}^{j_1-1} 2^{j+1-j_2} \mathbb{E} \left( \mathbb{1}\{\check{j} = j_1\} \mathbb{1}\{\tilde{j} = j_2\} (ave_f(j+1, \hat{\mathbf{i}}_{j+1}) - ave_f(j, \hat{\mathbf{i}}_j))^2 \right. \\
&\quad \left. \mathbb{1}\{ave_f(j+1, \hat{\mathbf{i}}_{j+1}) > ave_f(j, \hat{\mathbf{i}}_j)\} (\mathbb{1}\{\tilde{A}_0 \cup (\tilde{B}_0 \cap \tilde{D}_1^c)\} + \mathbb{1}\{\tilde{B}_0 \cap \tilde{D}_1\} \mathbb{1}\{j = j_2, j_1 = j+1\}) \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j=j_2}^{\infty} 2^{j+1-j_2} \mathbb{E} \left( (ave_f(j+1, \hat{\mathbf{i}}_{j+1}) - ave_f(j, \hat{\mathbf{i}}_j))^2 \mathbb{1}\{ave_f(j+1, \hat{\mathbf{i}}_{j+1}) > ave_f(j, \hat{\mathbf{i}}_j)\} \right. \\
&\quad \mathbb{1}\{\tilde{j} = j_2\} \sum_{j_1=j+1}^{\infty} \Phi(-2)^{(j_1-j_2-2)+} \Phi(-2 + \frac{13}{16} \rho_m(\frac{\sigma}{\sqrt{n}}; f) \sqrt{2^{J-j^*-2}}}{\gamma_s \sigma \sqrt{2}})^{(j_2-j^*-\delta)+} \\
&\quad \left. (\mathbb{1}\{\tilde{A}_0 \cup (\tilde{B}_0 \cap \tilde{D}_1^c)\} + \mathbb{1}\{\tilde{B}_0 \cap \tilde{D}_1\} \mathbb{1}\{j = j_2, j_1 = j+1\}) \right) \tag{A.2.107}
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{j_2=2}^{\infty} \sum_{j=j_2}^{\infty} 2^{j+1-j_2} \mathbb{E} \left( \left( ave_f(j+1, \hat{\mathbf{i}}_{j+1}) - ave_f(j, \hat{\mathbf{i}}_j) \right)^2 \mathbb{1}\{ave_f(j+1, \hat{\mathbf{i}}_{j+1}) > ave_f(j, \hat{\mathbf{i}}_j)\} \right. \\
&\quad \mathbb{1}\{\tilde{\mathbf{j}} = j_2, \tilde{A}_0 \cup \tilde{B}_0\} \left( \mathbb{1}\{j = j_2\} \left( 1 + \frac{1}{1 - \Phi(-2)} \right) + \mathbb{1}\{j \geq j_2 + 1\} \frac{\Phi(-2)^{j-j_2-1}}{1 - \Phi(-2)} \right) \\
&\quad \left. \Phi \left( -2 + \frac{\frac{13}{16} \rho_m(\frac{\sigma}{\sqrt{n}}; f) \sqrt{2^{J-j^*-2}}}{\gamma_s \sigma \sqrt{2}} \right)^{(j_2-j^*-\delta_0)_+} \right) \\
&\leq \sum_{j_2=2}^{\infty} \sum_{j=j_2}^{\infty} 2^{j+1-j_2} \left( \mathbb{1}\{j = j_2\} \left( 1 + \frac{1}{1 - \Phi(-2)} \right) + \mathbb{1}\{j \geq j_2 + 1\} \frac{\Phi(-2)^{j-j_2-1}}{1 - \Phi(-2)} \right) \\
&\quad \Phi \left( -2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s} \right)^{(j_2-j^*-\delta_0)_+} \\
&\quad \mathbb{E} \left( \left( ave_f(j+1, \hat{\mathbf{i}}_{j+1}) - ave_f(j, \hat{\mathbf{i}}_j) \right)^2 \mathbb{1}\{ave_f(j+1, \hat{\mathbf{i}}_{j+1}) > ave_f(j, \hat{\mathbf{i}}_j)\} \mathbb{1}\{\tilde{\mathbf{j}} = j_2, \tilde{A}_0 \cup \tilde{B}_0\} \right)
\end{aligned}$$

Now define the set  $\mathcal{C}(j, k)$  to be the set of pairs  $(i_1, i_2)$  such that,  $P(\hat{\mathbf{i}}_{k+1} = i_2, \hat{\mathbf{i}}_k = i_1 | \tilde{\mathbf{j}} = j) > 0$  and  $ave_f(j+1, i_2) > ave_f(j, i_1)$ . Then we know that  $|\mathcal{C}(j, k)| \leq \min\{10 \cdot 2^{k-j} \cdot 2, 3 \cdot 4^{k+1-j}\}$ .

$$\begin{aligned}
&\mathbb{E} \left( \left( ave_f(j+1, \hat{\mathbf{i}}_{j+1}) - ave_f(j, \hat{\mathbf{i}}_j) \right)^2 \mathbb{1}\{ave_f(j+1, \hat{\mathbf{i}}_{j+1}) > ave_f(j, \hat{\mathbf{i}}_j)\} \mathbb{1}\{\tilde{\mathbf{j}} = j_2, \tilde{A}_0 \cup \tilde{B}_0\} \right) \\
&\leq \sum_{(i_1, i_2) \in \mathcal{C}(j_2, j)} \mathbb{E} \left( \left( ave_f(j+1, i_2) - ave_f(j, i_1) \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} = j_2, \tilde{A}_0 \cup \tilde{B}_0\} \mathbb{1}\{\hat{\mathbf{i}}_{j+1} = i_2, \hat{\mathbf{i}}_j = i_1\} \right) \\
&\leq \sum_{(i_1, i_2) \in \mathcal{C}(j_2, j)} \mathbb{E} \left( \left( ave_f(j+1, i_2) - ave_f(j, i_1) \right)^2 \mathbb{1}\{\hat{\mathbf{i}}_{j+1} = i_2, \hat{\mathbf{i}}_j = i_1\} \right) \\
&\leq \sum_{(i_1, i_2) \in \mathcal{C}(j_2, j)} \left( ave_f(j+1, i_2) - ave_f(j, i_1) \right)^2 \Phi \left( -\frac{(ave_f(j+1, i_2) - ave_f(j, i_1)) \sqrt{2^{J-j-1}}}{\gamma_l \sigma \sqrt{2}} \right) \\
&\leq \sum_{(i_1, i_2) \in \mathcal{C}(j_2, j)} 2^{j+1-J} \cdot 2\sigma^2 \gamma_l^2 Q \\
&\leq \min\{10 \cdot 2^{j-j_2} \cdot 2, 3 \cdot 4^{j+1-j_2}\} 2^{j+1-J} \cdot 2\sigma^2 \gamma_l^2 Q.
\end{aligned} \tag{A.2.108}$$

Still,  $Q = \sup_{x>0} x^2 \Phi(-x)$ .

Continue with inequality (A.2.107), we have

$$\begin{aligned}
& \mathbb{E} \left( (ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}))_+ \mathbb{1}\{\tilde{j} + 1 \leq \check{j}\} \mathbb{1}\{\tilde{A}_0 \cup \tilde{B}_0\} \right) \\
& \leq \sum_{j_2=2}^{\infty} \sum_{j=j_2}^{\infty} 2^{j+1-j_2} (\mathbb{1}\{j = j_2\} (1 + \frac{1}{1 - \Phi(-2)}) + \mathbb{1}\{j \geq j_2 + 1\} \frac{\Phi(-2)^{j-j_2-1}}{1 - \Phi(-2)}) \\
& \quad \Phi(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s})^{(j_2-j^*-\delta_0)_+} \min\{10 \cdot 2^{j-j_2} \cdot 2, 3 \cdot 4^{j+1-j_2}\} 2^{j+1-J} \cdot 2\sigma^2\gamma_l^2 Q \\
& = \sum_{j_2=2}^{\infty} \left( 24(1 + \frac{1}{1 - \Phi(-2)}) + \frac{4}{1 - \Phi(-2)} \frac{80}{1 - 8\Phi(-2)} \right) \Phi(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s})^{(j_2-j^*-1)_+} 2^{j_2+2-J} \sigma^2\gamma_l^2 Q \\
& \leq 2^{j^*-J} \sigma^2\gamma_l^2 Q \sum_{j_2=2}^{\infty} \left( 24(1 + \frac{1}{1 - \Phi(-2)}) + \frac{80}{1 - 8\Phi(-2)} \frac{4}{1 - \Phi(-2)} \right) \Phi(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s})^{(j_2-j^*-1)_+} 2^{j_2-j^*+2} \\
& \leq \frac{8\sigma^2\gamma_l^2 Q}{n\rho_z(\frac{\sigma}{\sqrt{n}}; f)} \tilde{c}_{m_9} \leq \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2 \cdot 16\gamma_l^2 Q \tilde{c}_{m_9} = c_{m_9} \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2
\end{aligned}$$

(A.2.109)



Now let's turn to the second term

$$\begin{aligned}
& \mathbb{E} \left( \left( ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}) - ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}) \right)_+ \mathbb{1}\{\tilde{j} + 3 \leq j\} \mathbb{1}\{(\tilde{B}_0 \cap \tilde{D}_1) \cup (\tilde{C}_0 \cap \tilde{B}_1)\} \right) \\
& \leq \sum_{j_2=2}^{\infty} \sum_{j_1=j_2+3}^{\infty} 2 \sum_{j=j_2+2}^{j_1-1} \mathbb{E} \left( 2^{j-j_2-2} \left( (ave_f(j+1, \hat{\mathbf{i}}_{j+1}) - ave_f(j, \hat{\mathbf{i}}_j))_+ \right)^2 \right. \\
& \quad \left. \mathbb{1}\{\tilde{j} = j_1, \tilde{j} = j_2\} \mathbb{1}\{(\tilde{B}_0 \cap \tilde{D}_1) \cup (\tilde{C}_0 \cap \tilde{B}_1)\} \right) \\
& \leq \sum_{j_2=2}^{\infty} \sum_{j=j_2+2}^{\infty} 2^{j-j_2-1} \sum_{j_1=j+1}^{\infty} \sum_{(i_1, i_2) \in \mathcal{C}(j_2, j)} \mathbb{E} \left( \left( ave_f(j+1, i_2) - ave_f(j, i_1) \right)^2 \right. \\
& \quad \left. \mathbb{1}\{\tilde{j} = j_1\} \mathbb{1}\{\tilde{j} = j_2\} \mathbb{1}\{\hat{\mathbf{i}}_{j+1} = i_2, \hat{\mathbf{i}}_j = i_1\} \mathbb{1}\{(\tilde{B}_0 \cap \tilde{D}_1) \cup (\tilde{C}_0 \cap \tilde{B}_1)\} \right) \\
& \leq \sum_{j_2=2}^{\infty} \sum_{j=j_2+2}^{\infty} 2^{j-j_2-1} \sum_{(i_1, i_2) \in \mathcal{C}(j_2, j)} \sum_{j_1=j+1}^{\infty} \mathbb{E} \left( \left( ave_f(j+1, i_2) - ave_f(j, i_1) \right)^2 \right. \\
& \quad \left. \mathbb{1}\{\tilde{j} = j_2\} \mathbb{1}\{\hat{\mathbf{i}}_{j+1} = i_2, \hat{\mathbf{i}}_j = i_1\} \mathbb{1}\{(\tilde{B}_0 \cap \tilde{D}_1) \cup (\tilde{C}_0 \cap \tilde{B}_1)\} \right) \tag{A.2.110} \\
& \quad \Phi(-2)^{j_1-j_2-3} \Phi\left(-2 + \frac{\frac{13}{16}\rho_m(\frac{\sigma}{\sqrt{n}}; f)\sqrt{2^{J-j^*-2}}}{\gamma_s \sigma \sqrt{2}}\right)^{(j_2-j^*)_+} \\
& \leq \sum_{j_2=2}^{\infty} \sum_{j=j_2+2}^{\infty} 2^{j-j_2-1} \min\{20 \cdot 2^{j-j_2}, 3 \cdot 4^{j+1-j_2}\} 2^{j+2-J} \sigma^2 \gamma_t^2 Q \\
& \quad \frac{\Phi(-2)^{j-j_2-2}}{1 - \Phi(-2)} \Phi\left(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s}\right)^{(j_2-j^*)_+} \\
& = 2^{j^*-J} \sigma^2 \sum_{j_2=2}^{\infty} 2^{j_2-j^*} \Phi\left(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s}\right)^{(j_2-j^*)_+} \tilde{c}_{m10} \\
& \leq c_{m10} \rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2.
\end{aligned}$$

Finally, let's look at the third term

$$\begin{aligned}
& \mathbb{E} \left( \left( ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) \right)_+ \mathbb{1}\{\check{j} + 1 \leq \check{j}\} \mathbb{1}\{\tilde{C}_0 \cap \tilde{A}_1\} \right) \\
& \leq \sum_{j_2=2}^{\infty} \sum_{j=j_2+1}^{\infty} 2^{j-j_2} \sum_{j_1=j+1}^{\infty} \sum_{(i_1, i_2) \in \mathcal{C}(j_2, j)} \mathbb{E} \left( \left( ave_f(j+1, i_2) - ave_f(j, i_1) \right)^2 \right. \\
& \quad \left. \mathbb{1}\{\check{j} = j_1\} \mathbb{1}\{\check{j} = j_2\} \mathbb{1}\{\hat{\mathbf{i}}_{j+1} = i_2, \hat{\mathbf{i}}_j = i_1\} \mathbb{1}\{(\tilde{C}_0 \cap \tilde{A}_1)\} \right) \\
& \leq \sum_{j_2=2}^{\infty} \sum_{j=j_2+1}^{\infty} 2^{j-j_2} \min\{20 \cdot 2^{j-j_2}, 3 \cdot 4^{j+1-j_2}\} 2^{j+2-J} \sigma^2 \gamma_l^2 Q \frac{\Phi(-2)^{j-j_2-2}}{1 - \Phi(-2)} \Phi(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s})^{(j_2-j^*)+} \\
& \leq 2^{j^*-J} \sigma^2 \tilde{c}_{m11} \leq c_{m11} \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2
\end{aligned} \tag{A.2.111}$$

Therefore,

$$\begin{aligned}
& \mathbb{E} \left( \left( \left( ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) \right)_+ \right)^2 \mathbb{1}\{\check{j} \leq \check{j}\} \right) \\
& \leq (c_{m9} + c_{m10} + c_{m11}) \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2 = c_{m7} \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2
\end{aligned} \tag{A.2.112}$$

□

*Proof of Lemma A.2.6.* First we define the following notation:

$$IH(j) = \{\mathbf{i}_j^* - 4, \mathbf{i}_j^* - 3, \mathbf{i}_j^* - 2, \mathbf{i}_j^* + 2, \mathbf{i}_j^* + 3, \mathbf{i}_j^* + 4\},$$

which denotes the possible values of  $\hat{\mathbf{i}}_j$  if  $j = \check{j}$ .

$$\begin{aligned}
& \mathbb{E} \left( \left( \text{ave}_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}} - M(f)) \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} \leq \check{\mathbf{j}}\} \right) \\
&= \sum_{j_2=2}^J \sum_{j_1=j_2}^J \sum_{i \in IH(j_2)} \mathbb{E} \left( \left( \text{ave}_f(j_2, i) - M(f) \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} = j_2, \check{\mathbf{j}} = j_1, \hat{\mathbf{i}}_{j_2} = i\} \right) \\
&\leq \sum_{j_2=2}^J \sum_{i \in IH(j_2)} \sum_{j_1=j_2}^J \mathbb{E} \left( \left( \text{ave}_f(j_2, i) - M(f) \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} = j_2, \hat{\mathbf{i}}_{j_2} = i\} \right. \\
&\quad \left( \mathbb{E}(\mathbb{1}\{\check{\mathbf{j}} = j_1\} | \mathbf{Y}_l) \mathbb{1}\{j_1 \leq \mathbf{j}^* + 2\} + \mathbb{1}\{j_1 \geq \mathbf{j}^* + 3\} \left( \mathbb{E}(\mathbb{1}\{\check{\mathbf{j}} = j_1\} | \mathbf{Y}_l) \wedge \right. \right. \\
&\quad \left. \left. \Pi_{j=\mathbf{j}^*+2}^{j_1-1} \max\{\Phi(-2), \Phi(-2 + (\frac{7}{16} + \frac{6m_j}{\rho_z(\frac{\sigma}{\sqrt{n}}; f)})\rho_m(\frac{\sigma}{\sqrt{n}}; f) \frac{\sqrt{2^{J-j}}}{\sqrt{2}\gamma_s\sigma})\} \right) \right) \Bigg) \\
&\leq \sum_{j_2=2}^J \sum_{i \in IH(j_2)} \mathbb{E} \left( \left( \text{ave}_f(j_2, i) - M(f) \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} = j_2, \hat{\mathbf{i}}_{j_2} = i\} \right. \\
&\quad \left. \left( \mathbb{1}\{j_2 \leq \mathbf{j}^* + 2\} + \mathbb{1}\{j_2 \geq \mathbf{j}^* + 3\} \frac{\Phi(-2 + \frac{13}{16} \frac{\sqrt{3}}{4\sqrt{2}\gamma_s})^{j_2-\mathbf{j}^*-2}}{1 - \Phi(-2 + \frac{13}{16} \frac{\sqrt{3}}{4\sqrt{2}\gamma_s})} \right) \right) \\
&\leq \sum_{j_2=2}^J \sum_{i \in IH(j_2)} \left( \text{ave}_f(j_2, i) - M(f) \right)^2 \Phi \left( - \frac{(\text{ave}_f(j_2, i) - \text{ave}_f(j_2, \mathbf{i}_{j_2}^* + \text{sign}(i - \mathbf{i}_{j_2}^*))) \sqrt{2^{J-j_2}}}{\sqrt{2}\gamma_l\sigma} \right) \\
&\quad \left( \mathbb{1}\{j_2 \leq \mathbf{j}^* + 2\} + \mathbb{1}\{j_2 \geq \mathbf{j}^* + 3\} \frac{\Phi(-2 + \frac{1}{6})^{j_2-\mathbf{j}^*-2}}{1 - \Phi(-2 + \frac{1}{6})} \right) \\
&\leq \sum_{j_2=2}^J (23 \frac{1}{8}) 2\gamma_l^2 \sigma^2 2^{j_2-J} Q \frac{\Phi(-2 + \frac{1}{6})^{(j_2-\mathbf{j}^*-2)+}}{1 - \Phi(-2 + \frac{1}{6})} \leq 2^{\mathbf{j}^*+2-J} \sigma^2 \tilde{c}_{m8} \\
&\leq c_{m8} \rho_m(\frac{\sigma}{\sqrt{n}}; f)^2.
\end{aligned}$$

(A.2.113)

□

*Proof of Lemma A.1.40.*

$$\begin{aligned}
& \mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\check{\mathbf{j}} = \infty\}) \\
&= (\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2 \\
&\quad + \mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\check{\mathbf{j}} = \infty\} \mathbb{1}\{|\hat{\mathbf{i}}_J - \mathbf{i}_J^*| \geq 2\}) \\
&\leq (\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2 \\
&\quad + \mathbb{E}((f_{\hat{\mathbf{i}}_J} - M(f))^2 \mathbb{1}\{\check{\mathbf{j}} = \infty\} \mathbb{1}\{|\hat{\mathbf{i}}_J - \mathbf{i}_J^*| \geq 2\})
\end{aligned}$$

(A.2.114)

In the proof of Lemma A.2.4, all the argument using properties of  $\check{j}$  only uses that  $T_j > 2\tilde{\sigma}_j$  for  $j < \check{j}$ , so for the second term, all the argument can also go through here in the case  $\check{j} = \infty$ . So we have

$$\mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\check{j} = \infty\}) \leq c_{m6}\rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2 + (\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2 \quad (\text{A.2.115})$$

□

*Proof of Lemma A.1.41.*

$$\begin{aligned} & \sigma^2 \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}) \\ & \leq \sigma^2 \mathbb{1}\{J \leq j^* + 1\} + \sigma^2 \mathbb{E}(\mathbb{1}\{\check{j} = \infty\}) \mathbb{1}\{J \geq j^* + 2\} \\ & < \sigma^2 \frac{16}{n\rho_z(\frac{\sigma}{\sqrt{n}}; f)} \mathbb{1}\{J \leq j^* + 1\} + \sigma^2 \Phi(-2 + \frac{1}{6})^{J-j^*-1} \\ & \leq 32\rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2 \mathbb{1}\{J \leq j^* + 1\} + \sigma^2 \frac{\frac{1}{n}}{\frac{2^{J-j^*-1}}{n}} (2\Phi(-2 + \frac{1}{6}))^{J-j^*-1} \mathbb{1}\{J \geq j^* + 2\} \\ & \leq 32\rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2 \mathbb{1}\{J \leq j^* + 1\} + 32\rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2 \cdot 2\Phi(-2 + \frac{1}{6}) \mathbb{1}\{J \geq j^* + 2\} \\ & \leq 32\rho_m\left(\frac{\sigma}{\sqrt{n}}; f\right)^2 \end{aligned} \quad (\text{A.2.116})$$

□

*Proof of Lemma A.1.42.* Similar to the proof of Lemma A.1.39, we have

$$\begin{aligned}
& \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\tilde{j} < \infty\}) \\
& \leq \sum_{j_1=2}^J \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\tilde{j} > \tilde{j} = j_1\}) \\
& \quad + 2\mathbb{E}\left(\left((\hat{\mathbf{f}} - \text{ave}_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}))_+\right)^2 \mathbb{1}\{\tilde{j} \leq \tilde{j}\} \mathbb{1}\{\tilde{j} < \infty\}\right) \\
& \quad + 2\mathbb{E}\left(\left(\text{ave}_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}) - M(f)\right)^2 \mathbb{1}\{\tilde{j} \leq \tilde{j}\} \mathbb{1}\{\tilde{j} < \infty\}\right).
\end{aligned} \tag{A.2.117}$$

Similar to the arguments in the proof of Lemma A.2.1, we have

$$\begin{aligned}
& \sum_{j_1=2}^J \mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\tilde{j} > \tilde{j} = j_1\}) \\
& \leq \sum_{j_1=2}^J 3 \cdot 2^{j_1-J} (3.5\sqrt{2}\sigma\gamma_s)^2 V \\
& \leq 6 \cdot \frac{49}{2} \gamma_s^2 V \sigma^2 = \check{c}_{m4} \sigma^2,
\end{aligned} \tag{A.2.118}$$

where  $V = \max_{x>0} x^2 \Phi(2-x)$ .

Similar to the arguments in the proof of Lemma A.2.3

$$\begin{aligned}
& 2\mathbb{E}\left(\left((\hat{\mathbf{f}} - \text{ave}_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}))_+\right)^2 \mathbb{1}\{\tilde{j} \leq \tilde{j}\} \mathbb{1}\{\tilde{j} < \infty\}\right) \\
& \leq 2 \sum_{j_2=2}^J \sum_{j_1=j_2}^J \\
& \mathbb{E}\left(\mathbb{1}\{\text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1} + 2) > \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1})\} 2^{3+j_1-J} \gamma_s^2 \sigma^2 V \mathbb{1}\{\tilde{j} = j_2\} \Phi\left(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s}\right)^{(j_1-j^*-2)+} \right. \\
& \quad \left. + \mathbb{1}\{\text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1} - 2) > \text{ave}_f(j_1, \hat{\mathbf{i}}_{j_1})\} 2^{3+j_1-J} \gamma_s^2 \sigma^2 V \mathbb{1}\{\tilde{j} = j_2\} \Phi\left(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s}\right)^{(j_1-j^*-2)+} \right) \\
& \leq 4 \sum_{j_2=2}^J \mathbb{E}(\mathbb{1}\{\tilde{j} = j_2\}) \gamma_s^2 V 2^4 \sigma^2 \\
& \leq 4 \gamma_s^2 V 2^4 \sigma^2 = \check{c}_{m5} \sigma^2,
\end{aligned} \tag{A.2.119}$$

where  $V = \max_{x>0} x^2 \Phi(2-x)$ .

For the third term, we have

$$\begin{aligned}
& 2\mathbb{E}\left(\left(ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - M(f)\right)^2 \mathbb{1}\{\check{j} \leq \check{j}\} \mathbb{1}\{\check{j} < \infty\}\right) \\
& \leq 4\mathbb{E}\left(\left(\left(ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}})\right)_+\right)^2 \mathbb{1}\{\check{j} \leq \check{j} < \infty\}\right) \\
& \quad + 4\mathbb{E}\left(\left(\left(ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}) - M(f)\right)_+\right)^2 \mathbb{1}\{\check{j} \leq \check{j} < \infty\}\right).
\end{aligned} \tag{A.2.120}$$

Now we have the following lemmas which we will prove later:

**Lemma A.2.7.**

$$\mathbb{E}\left(\left(\left(ave_f(\check{j}, \hat{\mathbf{i}}_{\check{j}}) - ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}})\right)_+\right)^2 \mathbb{1}\{\check{j} \leq \check{j} < \infty\}\right) \leq \check{c}_{m6} \sigma^2. \tag{A.2.121}$$

**Lemma A.2.8.**

$$\mathbb{E}\left(\left(\left(ave_f(\tilde{j}, \hat{\mathbf{i}}_{\tilde{j}}) - M(f)\right)_+\right)^2 \mathbb{1}\{\check{j} \leq \check{j} < \infty\}\right) \leq \check{c}_{m7} \sigma^2. \tag{A.2.122}$$

Now we can conclude that

$$\mathbb{E}((\hat{\mathbf{f}} - M(f))^2 \mathbb{1}\{\check{j} < \infty\}) \leq (\check{c}_{m4} + \check{c}_{m5} + 4\check{c}_{m6} + 4\check{c}_{m7}) \sigma^2 = \check{c}_{m2}^2 \sigma^2. \tag{A.2.123}$$

□

*Proof of A.2.7.* Note that, in this lemma, we have  $\check{j} < \infty$ , however, it's not the essential, all is needed from it is that  $\tilde{j} \leq J$ , which comes from  $\tilde{j} \leq \check{j} < \infty$ . The proof of this lemma use many arguments from the proof of Lemma A.2.5 and lemmas proving it. And it can be seen that all the use of  $\check{j}$  there are that  $T_j > 2\bar{\sigma}_j$  for  $j < \check{j}$ , and  $\tilde{j} \leq \check{j}$ . Similar to the arguments in the proof of Lemma A.2.5, and suppose we take all the notation there, then

we have

$$\begin{aligned}
& \mathbb{E} \left( \left( (ave_f(\hat{\mathbf{j}}, \hat{\mathbf{i}}_{\hat{\mathbf{j}}}) - ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}))_+ \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} \leq \hat{\mathbf{j}}\} \right) \\
& \leq \mathbb{E} \left( \left( (ave_f(\hat{\mathbf{j}}, \hat{\mathbf{i}}_{\hat{\mathbf{j}}}) - ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}))_+ \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} + 1 \leq \hat{\mathbf{j}}\} (\mathbb{1}\{\tilde{A}_0 \cup (\tilde{B}_0 \cap \tilde{D}_1^c) \cup (\tilde{B}_0 \cap \tilde{D}_1 \cap \{\hat{\mathbf{j}} = \tilde{\mathbf{j}} + 1\})\} \right. \\
& \quad \left. + \mathbb{1}\{(\tilde{B}_0 \cap \tilde{D}_1) \cup (\tilde{C}_0 \cap \tilde{B}_1)\} \mathbb{1}\{\hat{\mathbf{j}} \geq \tilde{\mathbf{j}} + 3\} + \mathbb{1}\{\tilde{C}_0 \cap \tilde{A}_1\} \right),
\end{aligned} \tag{A.2.124}$$

$$\begin{aligned}
& \mathbb{E} \left( \left( (ave_f(\hat{\mathbf{j}}, \hat{\mathbf{i}}_{\hat{\mathbf{j}}}) - ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}))_+ \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} + 1 \leq \hat{\mathbf{j}}\} \mathbb{1}\{\tilde{A}_0 \cup (\tilde{B}_0 \cap \tilde{D}_1^c) \cup (\tilde{B}_0 \cap \tilde{D}_1 \cap \{\hat{\mathbf{j}} = \tilde{\mathbf{j}} + 1\})\} \right) \\
& \leq \sum_{j_2=2}^J \sum_{j_1=j_2+1}^J \mathbb{E} \left( 2 \sum_{j=j_2}^{j_1-1} 2^{j-j_2} \left( (ave_f(j+1, \hat{\mathbf{i}}_{j+1}) - ave_f(j, \hat{\mathbf{i}}_j))_+ \right)^2 \right) \\
& \leq \sum_{j_2=2}^J \left( 24 \left( 1 + \frac{1}{1 - \Phi(-2)} \right) + \frac{4}{1 - \Phi(-2)} \frac{80}{1 - 8\Phi(-2)} \right) \Phi \left( -2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s} \right)^{(j_2 - j^* - 1) + 2j_2 + 2 - J} \sigma^2 \gamma_l^2 Q \\
& \leq \left( 24 \left( 1 + \frac{1}{1 - \Phi(-2)} \right) + \frac{4}{1 - \Phi(-2)} \frac{80}{1 - 8\Phi(-2)} \right) 2^3 \sigma^2 \gamma_l^2 Q = \check{c}_{m8} \sigma^2,
\end{aligned} \tag{A.2.125}$$

$$\begin{aligned}
& \mathbb{E} \left( \left( (ave_f(\hat{\mathbf{j}}, \hat{\mathbf{i}}_{\hat{\mathbf{j}}}) - ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}))_+ \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} + 3 \leq \hat{\mathbf{j}}\} \mathbb{1}\{(\tilde{B}_0 \cap \tilde{D}_1) \cup (\tilde{C}_0 \cap \tilde{B}_1)\} \right) \\
& \leq \sum_{j_2=2}^{J-3} \sum_{j_1=j_2+3}^J 2 \sum_{j=j_2+2}^{j_1-1} \mathbb{E} \left( 2^{j-j_2-2} \left( (ave_f(j+1, \hat{\mathbf{i}}_{j+1}) - ave_f(j, \hat{\mathbf{i}}_j))_+ \right)^2 \right. \\
& \quad \left. \mathbb{1}\{\hat{\mathbf{j}} = j_1, \tilde{\mathbf{j}} = j_2\} \mathbb{1}\{(\tilde{B}_0 \cap \tilde{D}_1) \cup (\tilde{C}_0 \cap \tilde{B}_1)\} \right) \\
& \leq 2^{-J} \sigma^2 \sum_{j_2=2}^{J-3} 2^{j_2} \Phi \left( -2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s} \right)^{(j_2 - j^*) +} \check{c}_{m10} \\
& \leq \check{c}_{m9} \sigma^2,
\end{aligned} \tag{A.2.126}$$

$$\begin{aligned}
& \mathbb{E} \left( \left( (ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}) - ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}))_+ \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} + 1 \leq \hat{\mathbf{j}}\} \mathbb{1}\{\tilde{C}_0 \cap \tilde{A}_1\} \right) \\
& \leq \sum_{j_2=2}^{J-2} \sum_{j=j_2+1}^{J-1} 2^{j-j_2} \sum_{j_1=j+1}^J \sum_{(i_1, i_2) \in \mathcal{C}(j_2, j)} \mathbb{E} \left( \left( ave_f(j+1, i_2) - ave_f(j, i_1) \right)^2 \right. \\
& \quad \left. \mathbb{1}\{\hat{\mathbf{j}} = j_1\} \mathbb{1}\{\tilde{\mathbf{j}} = j_2\} \mathbb{1}\{\hat{\mathbf{i}}_{j+1} = i_2, \hat{\mathbf{i}}_j = i_1\} \mathbb{1}\{(\tilde{C}_0 \cap \tilde{A}_1)\} \right) \\
& \leq \sum_{j_2=2}^{J-2} \sum_{j=j_2+1}^{J-1} \\
& \quad 2^{j-j_2} \min\{20 \cdot 2^{j-j_2}, 3 \cdot 4^{j+1-j_2}\} 2^{j+2-J} \sigma^2 \gamma_l^2 Q \frac{\Phi(-2)^{j-j_2-2}}{1 - \Phi(-2)} \Phi(-2 + \frac{13\sqrt{3}}{64\sqrt{2}\gamma_s})^{(j_2-\mathbf{j}^*)+} \\
& \leq \check{c}_{m10} \sigma^2.
\end{aligned} \tag{A.2.127}$$

Therefore,

$$\begin{aligned}
& \mathbb{E} \left( \left( (ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}) - ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}))_+ \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} \leq \hat{\mathbf{j}}\} \right) \\
& \leq (\check{c}_{m8} + \check{c}_{m9} + \check{c}_{m10}) \sigma^2.
\end{aligned} \tag{A.2.128}$$

□

*Proof of A.2.8.* The arguments in the proof of Lemma A.2.6 hold, and we only need to change the last two inequalities to come to statement of this lemma.

More specifically

$$\begin{aligned}
& \mathbb{E} \left( \left( ave_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}) - M(f) \right)^2 \mathbb{1}\{\tilde{\mathbf{j}} \leq \hat{\mathbf{j}}\} \right) \\
& \leq \sum_{j_2=2}^J (23 \frac{1}{8}) 2 \gamma_l^2 \sigma^2 2^{j_2-J} Q \frac{\Phi(-2 + \frac{1}{6})^{(j_2-\mathbf{j}^*-2)+}}{1 - \Phi(-2 + \frac{1}{6})} \\
& \leq \check{c}_{m7} \sigma^2.
\end{aligned} \tag{A.2.129}$$

□



*Proof of Lemma A.1.43.* Similar to the arguments in Lemma A.1.40, we have

$$\begin{aligned}
& \mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\hat{\mathbf{j}} = \infty\}) \\
&= (\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2 \\
&\quad + \mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\hat{\mathbf{j}} = \infty\} \mathbb{1}\{|\hat{\mathbf{i}}_J - \mathbf{i}_J^*| \geq 2\}) \\
&\leq (\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2 \\
&\quad + \mathbb{E}((f_{\hat{\mathbf{i}}_J} - M(f))^2 \mathbb{1}\{\hat{\mathbf{j}} = \infty\} \mathbb{1}\{|\hat{\mathbf{i}}_J - \mathbf{i}_J^*| \geq 2\}).
\end{aligned} \tag{A.2.130}$$

Also, since in the proof of Lemma A.2.4, and Lemma A.1.42, Lemma A.1.43, all the argument using properties of  $\hat{\mathbf{j}}$  only uses that  $T_j > 2\tilde{\sigma}_j$  for  $j < \hat{\mathbf{j}}$ , so for the second term, all the argument can also go through here in the case  $\hat{\mathbf{j}} = \infty$ . So we have

$$\begin{aligned}
& \mathbb{E}((f_{\hat{\mathbf{i}}_J} - M(f))^2 \mathbb{1}\{\hat{\mathbf{j}} = \infty\} \mathbb{1}\{|\hat{\mathbf{i}}_J - \mathbf{i}_J^*| \geq 2\}) \\
&\leq 2\mathbb{E}\left(\left((\text{ave}_f(\hat{\mathbf{j}}, \hat{\mathbf{i}}_{\hat{\mathbf{j}}}) - \text{ave}_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}))_+\right)^2 \mathbb{1}\{\tilde{\mathbf{j}} \leq \hat{\mathbf{j}}\}\right) + 2\mathbb{E}\left((\text{ave}_f(\tilde{\mathbf{j}}, \hat{\mathbf{i}}_{\tilde{\mathbf{j}}}) - M(f))^2 \mathbb{1}\{\tilde{\mathbf{j}} \leq \hat{\mathbf{j}}\}\right) \\
&\leq 2(\check{c}_{m6} + \check{c}_{m7})\sigma^2.
\end{aligned} \tag{A.2.131}$$

Therefore,

$$\mathbb{E}((f_{\mathbf{i}} - M(f))^2 \mathbb{1}\{\hat{\mathbf{j}} = \infty\}) \leq (\min\{f(x_i) : 0 \leq i \leq n\} - M(f))^2 + 2(\check{c}_{m6} + \check{c}_{m7})\sigma^2. \tag{A.2.132}$$

Let  $\check{c}_{m3} = \sqrt{2(\check{c}_{m6} + \check{c}_{m7})}$  gives the statement of the lemma.

□

### A.3. Comparison with CLS Methods and Connections with Classical Minimax Framework for Chapter 2

In this section, we compare our procedures with the convexity-constrained least squares methods for the minimizer and discuss the connections with the classical minimax framework. In particular, we prove that the CLS confidence interval for the minimizer proposed in Deng et al. (2020) is sub-optimal under the local minimax framework.

#### A.3.1. Sub-optimality of the CLS Confidence Interval

We start with the proof of Proposition 2.4.1. It suffices to prove the following proposition as we can set the  $r(n)$  in the following proposition to be arbitrary large.

**Proposition A.3.1.** *For any function  $r(n) \geq 1$ , for any integer  $n \geq 5$ ,  $\exists f_n \in \mathcal{F}_2$  such that*

$$\frac{\mathbb{E}_{f_n} L(\text{CLSCI}_\alpha)}{\mathbb{E}_{f_n} L(\text{CI}_{z,\alpha})} \geq r(n). \quad (\text{A.3.1})$$

*Proof.* Recall that we have established

$$\mathbb{E}_f L(\text{CI}_{z,\alpha}) \leq C_{2,\alpha} \tilde{L}_{z,\alpha,n}(\sigma; f), \quad \text{for all } f \in \mathcal{F}$$

in Theorem 2.4.2, and further, in the proof of Theorem 2.4.2, we have

$$\mathbb{E}_f L(\text{CI}_{z,\alpha}) \leq C_{2,\alpha} \left( \sup_{h \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right) \left(1 \wedge \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; h\right)}\right) + \frac{(1-2\alpha)}{2} \mathfrak{D}_z(n, f) \right),$$

where the definition of  $\mathcal{G}_n(f)$  is given in Equation (A.1.99). This combined with the lower bound of local minimax length of confidence interval that we established in Proposition A.1.4, namely

$$\tilde{L}_{z,\alpha,n}(\sigma; f) \geq \tilde{C}_{z,\alpha} \left( \sup_{g \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right) \left(1 \wedge \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right)}\right) + \frac{(1-2\alpha)}{2} \mathfrak{D}_z(n, f) \right),$$

indicates that it suffices to show that for any  $r(n) > 0$ , there exists  $f \in \mathcal{F}_2$  such that

$$\frac{\mathbb{E}_f L(CLSCI_\alpha)}{\left( \sup_{g \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right) \left(1 \wedge \sqrt{n \rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right)}\right) + \frac{(1-2\alpha)}{2} \mathfrak{D}_z(n, f) \right)} \geq r(n).$$

Note that  $L(CLSCI_\alpha) \geq \frac{1}{n}$ , we only need to find  $f \in \mathcal{F}_2$  such that

$$\mathfrak{D}_z(n, f) \leq \frac{1}{2nr(n)} \text{ and } \sup_{g \in \mathcal{G}_n(f)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right) \leq \frac{1}{2n(r(n) + 1)}. \quad (\text{A.3.2})$$

Consider function  $f_0(x) = 4n(r(n) + 1)^{\frac{3}{2}}(\sigma + 1)|x - \frac{\lfloor n/2 \rfloor}{n}|$ , for which we have

$$\mathfrak{D}_z(n, f_0) = 0, \quad \sup_{g \in \mathcal{G}_n(f_0)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right) \leq \frac{\left(\frac{3}{4}\right)^{\frac{1}{3}}}{2} \frac{1}{n(r(n) + 1)}.$$

The conditions mentioned in Inequality (A.3.2) are met, but  $f_0$  is not in  $\mathcal{F}_2$ . Now we will proceed to construct  $f_1(x) \in \mathcal{F}_2$  such that the conditions in Inequality (A.3.2) are still met.

Let

$$\tilde{f}_0(x) = \begin{cases} f_0(x), & x \in [0, 1] \\ f_0(1) + \sup_{t \rightarrow 1^-} \frac{f(1) - f(t)}{1 - t} (x - 1), & x > 1 \\ f_0(0) + \sup_{t \rightarrow 0^+} \frac{f(t) - f(0)}{t} x, & x < 0 \end{cases}. \quad (\text{A.3.3})$$

Then consider the following class of transformations of  $f_0$ :

$$T(f_0; \delta)(x) = \int \tilde{f}_0(t) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2\delta^2}\right) dt. \quad (\text{A.3.4})$$

It's easy to check that when  $f_0$  is a convex function on  $[0, 1]$ ,  $T(f_0; \delta)$  is a convex function on  $\mathbb{R}$ , and that  $T(f_0; \delta)$  is twice differentiable with continuous positive second order derivative around the minimizer. Also, when  $f_0$  is fixed,  $\lim_{\delta \rightarrow 0^+} \sup_{x \in [0, 1]} |T(f_0; \delta)(x) - f_0(x)| = 0$ .

Therefore,

$$\lim_{\delta \rightarrow 0^+} \sup_{g \in \mathcal{G}_n(T(f_0; \delta))} \rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right) = \sup_{g \in \mathcal{G}_n(f_0)} \rho_z\left(\frac{\sigma}{\sqrt{n}}; g\right) < \frac{1}{2n(r(n) + 1)},$$

and  $\lim_{\delta \rightarrow 0^+} \mathfrak{D}_z(n, T(f_0; \delta)) = \mathfrak{D}_z(n, f_0) = 0$ .

Thus there exists  $\delta(f_0) > 0$  such that  $T(f_0; \delta(f_0))$  satisfies conditions in Inequality (A.3.2), which concludes the proof. □

From the proof we can see the power of non-asymptotic and non-localized results.  $\rho_z(\frac{\sigma}{\sqrt{n}}; f)$  is not a localized quantity while second order derivative is a localized quantity. Therefore, the asymptotic result based on localized quantity will encounter the problem that no matter how large  $n$  is, it's still outside the realm of being local for some functions.

Before delving into a discussion in more depth, let's look at a more intuitive example where the convex least squares based confidence interval introduced in Deng et al. (2020) suffers from a long length empirically,  $f(x) = 100|2x - 1|$ . Its length remains roughly a constant while the benchmark apparently goes to zero as sample size goes to infinity. Note that the empirical performance for the estimation of the minimizer is reasonable, and  $f$  lies in the function class of CLS estimation (of the entire function), meaning that the “oracle” CLS estimator of the entire function would be  $f$  itself. An explanation of this long length is on the construction of the confidence interval after CLS. The length of the confidence interval in Deng et al. (2020) is a constant multiplier (depending on confidence level) of the distance of the two neighboring kinks around the minimizer based on the CLS estimation of the entire function. Note that for  $f$ , the perfect estimation of  $f$ , the neighboring kinks of the minimizer are always 0 and 1, regardless of the sample size. So it's not surprising to have a long length using kinks around the minimizer, which highly relies on second order derivatives rather than exploiting more of the convexity.

This extreme example, together with the example we provide in the proof of sub-optimality, shows that the convex least squares ingredient of the confidence interval construction is not only reason for sub-optimality. And Algorithm 1 provides a way to fully exploit the convexity of the true function.

On the other hand, for the behavior of our methods under the asymptotic sense for smooth functions (defined in Section A.3.2), we attain same rates optimal  $n^{-\frac{1}{2k+1}}$  for minimizer, which we will discuss more in Section A.3.2.

### **A.3.2. Connections with Classical Minimax Framework: Lower Bounds, Optimality, and Characteristics**

In this part we relate local minimax rates to classical minimax rates, which captures the worst case for a certain function class.

Before going into details, we elaborate on general comparison. Regarding the comparison with the classical minimax lower bound over a certain function class, the lower bound provided by our non-asymptotic local minimax framework (applied to that function class) is no larger than the classical one. Because in the classical minimax framework, the Le Cam two-point reduction, in a way, can be considered as a two-point case of Assouads or Fanos Lemma. This makes it a stricter criterion, and it preserves more information before taking supreme over the function class (individual functions are treated individually). A major difficulty of our non-asymptotic local minimax framework lies in the existence (and construction) of an adaptive procedure that attain this potentially smaller benchmark. And a key difference from the classical minimax framework is that the local minimax framework enables the characterization of the difficulty for estimating individual functions, and makes establishing the non-superefficiency type of results conceptually possible.

To illustrate through an example, we focus on convex function class with additional smoothness conditions, as in literature the classical minimax rates for both smooth functions and

smooth convex functions are extensively investigated. We walk through the procedures translating local minimax rates to classical minimax lower bounds, which has following (additional) implications that we highlight.

- For the same class of functions, all optimal procedures under non-asymptotic local minimax benchmarks are optimal in the classical sense.
- The local minimax rates established for one class of functions (e.g. convex functions) can be useful for establishing classical minimax lower bounds for another function class (e.g. smooth functions).

Last but not least, we show that the classical minimax rates for convex function class are meaningless, which shows the advantage of non-asymptotic local minimax framework.

The smoothness condition we consider is local smoothness defined around the minimizer. For  $k > 1$  and  $B \geq B_1 > 0$ , the locally smooth convex function class  $\Gamma_1(k; B_1, B)$  is defined as

$$\Gamma_1(k; B_1, B) = \{f \in \mathcal{F} : B_1 \leq \liminf_{t \rightarrow Z(f)} \frac{|f(t) - f(Z(f))|}{|t - Z(f)|^k} \leq \overline{\lim}_{t \rightarrow Z(f)} \frac{|f(t) - f(Z(f))|}{|t - Z(f)|^k} \leq B\}. \quad (\text{A.3.5})$$

Similar type of smoothness class has been considered in Shoung et al. (2001) except that their smoothness requires the limit to exist and be exactly  $B$  (i.e.  $B_1 = B$ ). We will also briefly discuss a global version of smoothness later. For the function class  $\mathcal{F} \cap \Gamma_1(k; B_1, B)$  the corresponding moduli of continuity is given by, for  $f \in \Gamma_1(k; B_1, B)$ ,

$$\begin{aligned} \hat{\omega}_z(\varepsilon; f) &= \sup\{|Z(f) - Z(g)| : \|f - g\|_2 \leq \varepsilon, g \in \Gamma_1(k; B_1, B)\}, \\ \hat{\omega}_m(\varepsilon; f) &= \sup\{|M(f) - M(g)| : \|f - g\|_2 \leq \varepsilon, g \in \Gamma_1(k; B_1, B)\}. \end{aligned}$$

Further, similar to the proof of Proposition 2.2.2 we can show that

$$\hat{\omega}_z(\varepsilon; f) \geq \rho_z(\varepsilon; f), \hat{\omega}_m(\varepsilon; f) \geq \rho_m(\varepsilon; f). \quad (\text{A.3.6})$$

We defer the proof of this inequality to the last part of this section.

Consider function  $f_1 = \frac{B}{2} |t - \frac{1}{2}|^k$ , which is in  $\Gamma_1(k; B_1, B)$ . Then we have that the classical minimax rate of estimating minimum for the function class  $\Gamma_1(k; B_1, B)$  is lower bounded by

$$\begin{aligned} & \inf_{\hat{M}} \sup_{f \in \Gamma_1(k; B_1, B)} E_f |\hat{M} - M(f)| \\ & \geq \sup_{f \in \Gamma_1(k; B_1, B)} \sup_{g \in \Gamma_1(k; B_1, B)} \inf_{\hat{M}} \max_{h \in \{f, g\}} \mathbb{E}_h |\hat{M} - M(h)| \\ & \geq \sup_{g \in \Gamma_1(k; B_1, B)} \inf_{\hat{M}} \max_{h \in \{f_1, g\}} \mathbb{E}_h |\hat{M} - M(h)| \\ & \geq a_1 \rho_m(\varepsilon; f_1) \\ & = a_1 c_{B,k} \varepsilon^{\frac{2k}{2k+1}}, \end{aligned} \quad (\text{A.3.7})$$

where  $c_{B,k} = 2^{\frac{-(k+1)}{2k+1}} B^{\frac{1}{2k+1}}$ .

Similarly, for estimating the minimizer, take  $f_1 = \frac{B_1}{2} |t - \frac{1}{2}|^k$ , we have the classical minimax rate being lower bounded by

$$\inf_{\hat{Z}} \sup_{f \in \Gamma_1(k; B_1, B)} E_f |\hat{Z} - Z(f)| \geq a_1 \left( \frac{2}{B_1^2} \right)^{\frac{1}{2k+1}} \varepsilon^{\frac{2}{2k+1}}. \quad (\text{A.3.8})$$

Note that the locally smooth convex function class  $\Gamma_1(k; B_1, B)$  is a subset of locally smooth function class, so the lower bounds for  $\Gamma_1(k; B_1, B)$  apparently hold for locally smooth function class. From here we can see that while our local minimax rates are primarily based on the properties of convex functions, it's also useful for establishing lower bounds for locally smooth function class.

To further illustrates this point, we show that this trick is also very useful for establishing lower bounds for estimating the minimum for globally smooth functions, which is also intensively investigated in the literature.

The globally smooth convex function class  $\Gamma_2(B, k)$  is defined as

$$\Gamma_2(B, k) = \{f \in \mathcal{F} : |f(t) - f(Z(f))| \leq B|t - Z(f)|^k, \forall t \in [0, 1]\}. \quad (\text{A.3.9})$$

Note that the globally smoothness differs from the locally smoothness in that we are not only interested in the local behavior around the minimizer. Globally smooth convex function class is a smaller function class when compared with locally smooth convex function class (if we can let  $B_1 = 0$  to allow the same form).

The continuity moduli can be similarly defined as

$$\tilde{\omega}_m(\varepsilon; f) = \sup \{|M(f) - M(g)| : \|f - g\|_2 \leq \varepsilon, g \in \Gamma_2(B, k)\}, \quad (\text{A.3.10})$$

for  $f \in \Gamma_2(B, k)$ .

Similarly, we can show that

$$\tilde{\omega}_m(\varepsilon; f) \geq \rho_m(\varepsilon; f), \quad (\text{A.3.11})$$

the proof of which is deferred to the last part.

With Inequality (A.3.11), using similar arguments as in Inequality (A.3.7), we have that the minimax rate for estimation of minimum for function class  $\Gamma_2(B, k)$  is lower bounded by  $a_1 c_{B,k} \varepsilon^{\frac{2k}{2k+1}}$  (where  $c_{B,k} = 2^{\frac{-(k+1)}{2k+1}} B^{\frac{1}{2k+1}}$ ), which automatically serve as a lower bound for globally smooth function class.

The lower bounds in white noise model are closely related to the non-parametric regression as shown before. Despite of the large volume of literature on non-parametric regression, the



lower bounds for various smooth classes are well known. For example, for isotropic Hölder class, the lower bound is not known until lately (Belitser et al., 2021).

Now we proceed to see the advantage of local minimax benchmarks compared with classical minimax rates. Consider a collection of functions  $f_\delta = \delta|t - \frac{1}{2}|$ , for  $\delta > 0$ . This collection of functions are convex. And we have

$$\lim_{\delta \rightarrow 0^+} \rho_z(\varepsilon; f_\delta) = \frac{1}{2},$$

$$\lim_{\delta \rightarrow +\infty} \rho_m(\varepsilon; f_\delta) = \infty,$$

which are lower bounds (up to some absolute constants) for classical minimax rates for convex functions. Any procedure will be optimal under classical minimax framework, which makes the classical minimax framework meaningless in this setting.

For transferring rates under our framework into classical minimax framework for the regression setting, we only need to change  $\varepsilon$  into  $\frac{\sigma^2}{n}$ , as the discretization error is always dominated by the noise induced error in classical minimax framework.

Also note that for the settings that CLS estimator/CLSCI are considered in Ghosal and Sen (2017) or Deng et al. (2020), it can be written as  $\cup_{B>0} \Gamma_1(k; B, B)$  for  $k \geq 2$  and being and even number. Note that our procedures do not depend on  $B$  while not only achieving the optimal minimax rate in classical sense (in terms of  $n$ ) for  $\Gamma_1(k; B, B)$  but also having a risk/length smaller than an universal constant multiple of the lower bound for each and every  $B$  and  $k$ . Our methods are adaptively optimal for the settings that CLS/CLSE are investigated in.

*Proof of Inequality (A.3.6) and Inequality (A.3.11).* To prove  $\tilde{\omega}_m(\varepsilon; f) \geq \rho_m(\varepsilon; f)$ , we only need to replace  $g_\delta(t)$  in the proof of Proposition 2.2.2 to be  $g_\delta = \max\{f(t), \min\{u_\varepsilon + \delta(|t - Z(f)|^k - |t_l - Z(f)|^k), u_\varepsilon + \delta(|t - Z(f)|^k - |t_r - Z(f)|^k)\}\}$  when  $k \geq 1$ , where  $\delta < B$ . It is easy to see that this new  $g_\delta \in \Gamma_2(B, k)$ ,  $\|g_\delta - f\| \leq \varepsilon$  and  $\lim_{\delta \rightarrow 0} |M(g_\delta) - M(f)| = \rho_m(\varepsilon; f)$ .

When  $k < 1$ , we just replace the  $k$  in newly constructed  $g_\delta$  by 1.

To prove  $\hat{\omega}_z(\varepsilon; f) \geq \rho_z(\varepsilon; f)$  and  $\hat{\omega}_m(\varepsilon; f) \geq \rho_m(\varepsilon; f)$ , without loss of generality, we assume  $t_r - Z(f) = \rho_z(\varepsilon; f)$ . Note that  $k > 1$ . We only need to replace  $g_\delta(t)$  in the proof of Proposition 2.2.2 to be  $\tilde{g}_\delta(t)$ , which is defined in the following way: let  $h_s(t) = B|t - t_r + \delta|^k + s$ , as when  $\delta$  is small enough,  $\forall t > t_r - \delta$ ,  $\frac{f(t) - f(t_r - \delta)}{t - t_r + \delta}$  is lower bounded by  $\frac{\lim_{t \rightarrow t_r^-} f(t_r) - f(t)}{2}$ , so  $\exists s$  such that  $h_s(t)$  and  $g_\delta(t)$  has an intersection  $t_1 \in (t_l, t_r - \delta)$  and an intersection  $t_2 \in (t_r - \delta, t_r)$ , which satisfy  $h_s(t) > g_\delta(t), \forall t \in (t_1, t_2)$  and  $h_s(t) < g_\delta(t)$  for a small neighborhood outside  $(t_1, t_2)$ .

Let  $\tilde{g}_\delta(t) = g_\delta(t) \forall t \in [0, 1] \setminus (t_1, t_2)$ , and  $\tilde{g}_\delta(t) = h_s(t) \forall t \in (t_1, t_2)$ . Then  $\tilde{g}_\delta \in \Gamma_1(k; B_1, B) \cap \mathcal{F}$ ,  $\|\tilde{g} - f\| \leq \varepsilon$ ,  $\lim_{\delta \rightarrow 0} |Z(\tilde{g}_\delta) - Z(f)| = \rho_z(\varepsilon; f)$ , and

$$\lim_{\delta \rightarrow 0} |M(\tilde{g}_\delta) - M(f)| \geq \lim_{\delta \rightarrow 0} |M(g_\delta) - M(f)| = \rho_m(\varepsilon; f).$$

□

### A.3.3. More on the Uncertainty Principle

In this subsection, we discuss more on the generality of the Uncertainty Principle. We start with the convex smoothness class we discussed in Section A.3.2. Uncertainty principle still holds for the function class  $\Gamma_1(k; B_1, B)$ , with  $\Gamma_1(k; B_1, B)$  defined in (A.3.5), which contains all the functions  $f \in \mathcal{F}$  satisfying

$$B_1 \leq \liminf_{t \rightarrow Z(f)} \frac{|f(t) - f(Z(f))|}{|t - Z(f)|^k} \leq \overline{\lim}_{t \rightarrow Z(f)} \frac{|f(t) - f(Z(f))|}{|t - Z(f)|^k} \leq B.$$

It follows from Inequality (A.3.6) that the moduli of continuity for the minimizer and minimum over the function class  $\Gamma_1(k; B_1, B)$  have the following relationship.

$$\hat{\omega}_z(\varepsilon; f) \hat{\omega}_m(\varepsilon; f)^2 \geq \rho_z(\varepsilon; f) \rho_m(\varepsilon; f)^2 \geq \frac{\varepsilon^2}{2}. \quad (\text{A.3.12})$$

So the Uncertainty Principle also holds for  $\Gamma_1(k; B_1, B)$ .

Further, using the smoothing technique in Equation (A.3.4) in the proof of Proposition A.3.1 on the examples used in constructing the lower bound in the proof of Inequality (A.3.6), we know that the Uncertainty Principle also holds for the  $k$ -th order differentiable convex function class for any  $k$ .

So there are many choices of subclass of the convex functions in  $\mathcal{F}$  where the Uncertainty Principle holds. Interested reader can further explore other possible choices. Further, since the tension between different quantities (e.g. minimizer and minimum in our case) also exists in other problems, we believe that similar Uncertainty Principles can be developed in other settings.

#### **A.3.4. The CLS Estimator under Local Minimax Framework**

The results on the behavior of the convex least squares estimator are mostly based on the limiting distribution, which are usually achieved by carrying out Taylor expansion of the function to second order around minimizer and analysis of the empirical process. Since the limiting distribution only holds when as sample size approaches infinity for fixed function, similar arguments are not applicable to prove results that hold *for all functions within a class for any given sample size or when sample size grows to infinity*. Also, the Taylor expansion approach won't work when the second order derivative does not exist at the minimizer. Hence the tools used in establishing the performance of convex least squares in the literature is not sufficient for investigating its behavior under our non-asymptotic local minimax framework. The behavior of the convex least squares estimator under our framework takes new tools and is of separate interest.

For functions twice differentiable around the minimizer with positive second order derivative at the minimizer, under asymptotic sense (i.e. fix function  $f$ , and let sample size  $n$  go to infinity), since the convex least squares estimator for minimizer  $\hat{Z}_{\text{cvx}}$  is bounded (i.e. in

$[0, 1]$ ), from the limiting distribution (Theorem 2.9 in Deng et al. (2020)), we know that

$$\limsup_{n \rightarrow \infty} \mathbb{E}(|\hat{Z}_{\text{cvx}} - Z(f)|)(n/\sigma^2)^{1/5} \leq \left( \frac{1}{f''(Z(f))} \right)^{2/5} \text{const}_1,$$

where  $\text{const}_1$  is an absolute constant, and that

$$\liminf_{n \rightarrow \infty} \mathbb{E}(|\hat{Z}_{\text{cvx}} - Z(f)|)(n/\sigma^2)^{1/5} \geq \left( \frac{1}{f''(Z(f))} \right)^{2/5} \text{const}_2,$$

where  $\text{const}_2$  is another absolute constant. Note that for functions twice differentiable at the minimizer with positive second order derivative, the key part of the benchmark for minimizer in our framework  $\rho_z(\frac{\sigma}{\sqrt{n}}; f)$  is of the order  $(\sigma^2/n)^{1/5} \left( \frac{1}{f''(Z(f))} \right)^{2/5}$  when  $n$  goes to infinity. Although the benchmark has a discretization part as shown in Section A.1.9, the reader can easily check the order is  $(\sigma^2/n)^{1/5} \left( \frac{1}{f''(Z(f))} \right)^{2/5}$  when  $f$  is fixed and  $n$  goes to infinity. In this asymptotic sense, convex least squares estimator matches our rate, which is also the optimal rate, for functions twice differentiable at the minimizer with positive second order derivative (the lower bound provided in Section A.3.2). However, this does not imply optimality for  $\hat{Z}_{\text{cvx}}$  under our non-asymptotic framework. It is possible that there exists a sequence of twice differentiable functions with positive second order derivative at the minimizer such that the ratio of its risk to our benchmark is an increasing function of sample size.

#### A.4. Simulation Results for Chapter 2

In this section we show simulation results comparing our algorithms to the ones based on the convex least squares (CLS) estimator. Note that known theoretically valid CLS based method only exist for estimation and inference of the minimizer, we make comparison on those tasks. For estimation of the minimizer, the theoretically valid CLS based method is taking the minimizer of the CLS estimator for the whole function. For inference of the minimizer, we adopt the latest CLS based confidence interval proposed by Deng et al. (2020), which is proved to enjoy good theoretical property in some restricted settings in a

restricted sense. Since our original method introduces data splitting procedure purely for technical reason, our attention is allocated more to the non-split version.

To sum up the comparison, first of all, our proposed methods, as being iterative and *local*, run much faster than any methods based the CLS, whose complexity, in theory, according to Simonetto (2021), scales as  $\mathcal{O}(n^3)$  for generic quadratic programming solvers or  $\mathcal{O}(n^2)$  per iteration for first-order methods. For estimating the minimizer, the proposed method and the CLS based method have comparable accuracy, with CLS being very sensitive to the smoothness while our methods are steady in terms of the benchmarks when they are computable. For inference of the minimizer, while both variants of the proposed confidence interval achieve the nominal coverage, the CLS based confidence interval behaves poor in either coverage or length, which isn't surprising due to its asymptotic nature of coverage and high dependence on second order derivative.

In addition to comparison, we also tested how our methods behave compared with our theoretical results, especially for tasks for minimum. Both of our methods achieve the nominal coverage for confidence interval and all the empirical risks/empirical lengths show clear linear relationship compared with the benchmarks when the benchmarks are computable.

#### A.4.1. Experiment Design

To generate the data, we set  $\sigma = 1$ . We carried out experiments on true functions with different smoothness, minimizer location, symmetry, etc. We tested on sample sizes 100, 500, 1000, 5000, 10000, 50000. For confidence intervals, we take 5 confidence levels, namely 0.8, 0.9, 0.05, 0.98, 0.99, which corresponds to  $\alpha = 0.2, 0.1, 0.05, 0.02, 0.01$ . For each true function, each sample size, we average on 100 replicates.

For the experiment testing our methods' behavior compared with theoretical results, we choose functions with computable benchmarks, and sample sizes easier to test the relationship, which we will discuss in detail in A.4.3. Now we focus on the general functions and

comparison.

We implement and compare three methods, as summarized in Table A.1, as mentioned earlier.

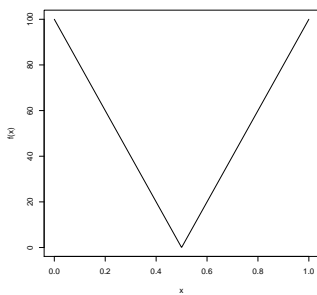
Method	Estimation		Inference	
	Minimizer	Minimum	Minimizer	Minimum
Proposed (split)	✓	✓	✓	✓
Variant (non-split & stop)	✓	✓	✓	✓
CLS based	✓		✓	

Table A.1: List of the methods to be compared and their applicable scenario.

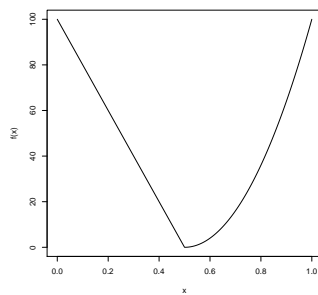
We investigate the following metrics.

- Running time of the methods.
- Empirical risk for estimating the minimizer and minimum.
- Coverage and length of confidence interval for the minimizer and the minimum. In particular, we construct confidence interval with 5 different confidence level with  $\alpha$  ranging from 0.2 to 0.01.

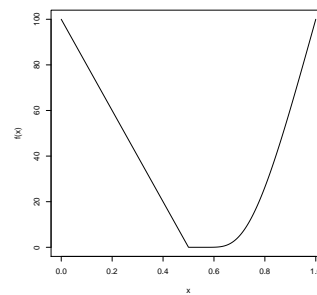
We have 10 test functions, as shown in the Equation (A.4.1). Figure A.3 shows the plots of those functions (in the order 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 from left top to right bottom), grouped based on the performances of all methods on those functions. Note that we include functions of different smoothness around the minimizer (i.e. of the types  $x$ ,  $x^2$ ,  $x^4$ ,  $\exp(-1/x)$ ), with both symmetric and asymmetric configurations. Also we include the functions with minimizer near boundary. Using similar arguments as in the proof of Proposition 2.4.1, we can convolute the true function with smooth kernel enough concentrated to the center to have a function that is smooth (i.e. differentiable to any order) and arbitrarily close to the original true function, regardless of the smoothness of the true function. So the phenomenon shown here also carries to the non-asymptotic region (i.e. small to medium sample size) of



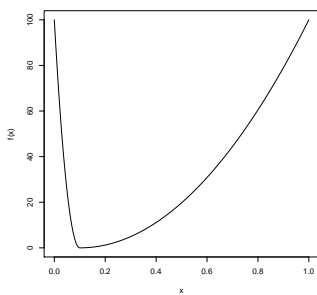
(a)  $f_1$



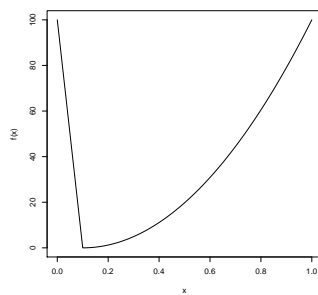
(b)  $f_2$



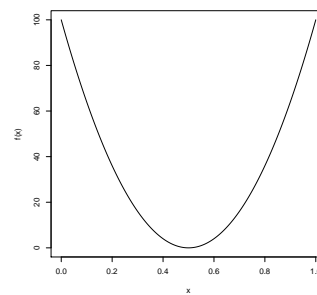
(c)  $f_3$



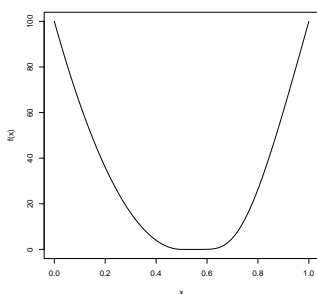
(d)  $f_4$



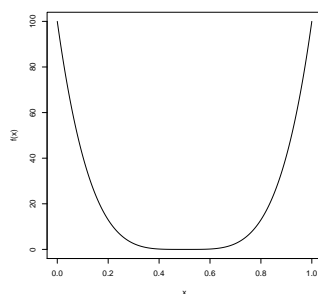
(e)  $f_5$



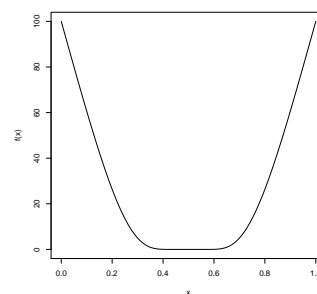
(f)  $f_6$



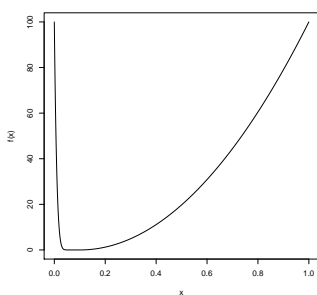
(g)  $f_7$



(h)  $f_8$



(i)  $f_9$



(j)  $f_{10}$

Figure A.3: Plot of true functions

functions of any smoothness.

$$f_1(x) = 100|2x - 1|$$

non-differentiable, symmetric, linear

$$f_2(x) = 100|2x - 1|\mathbb{1}\{x < 0.5\} + 100|2x - 1|^2\mathbb{1}\{x \geq 0.5\}$$

non-differentiable, non-symmetric,

the right side of the minimizer is positively twice differentiable

$$f_3(x) = 100|2x - 1|\mathbb{1}\{x < 0.5\} + 100 \exp\left(2 - \frac{1}{|x - 0.5|}\right)\mathbb{1}\{x \geq 0.5\}$$

non-differentiable with one side being arbitrarily differentiable (A.4.1)

with vanishing derivatives at minimizer

$$f_4(x) = 100|10x - 1|^2\mathbb{1}\{x < 0.1\} + 100|10 * x/9 - 1/9|^2\mathbb{1}\{x \geq 0.1\}$$

differentiable, with minimizer near boundary,

with both “sided” second order derivatives being positive

$$f_5(x) = 100|10x - 1|^1\mathbb{1}\{x < 0.1\} + 100|10 * x/9 - 1/9|^2\mathbb{1}\{x \geq 0.1\}$$

non-differentiable with minimizer near boundary



$$f_6(x) = 100(|2x - 1|)^2$$

twice differentiable with positive second order derivative

$$f_7(x) = 100|2x - 1|^2 \mathbb{1}\{x < 0.5\} + 100 \exp\left(2 - \frac{1}{|x - 0.5|}\right) \mathbb{1}\{x \geq 0.5\}$$

differentiable but not twice differentiable,

one side being arbitrarily differentiable with vanishing derivatives at minimizer,

non-symmetric

$$f_8(x) = 100(|2x - 1|)^4$$

fourth-order differentiable with vanishing second order derivative

$$f_9(x) = 100 \exp\left(2 - \frac{1}{|x - 0.5|}\right)$$

arbitrarily differentiable with vanishing derivatives of any order

$$f_{10}(x) = 100 \exp\left(2 - \frac{1}{|x - 0.1|}\right) \mathbb{1}\{x < 0.1\} + 100|10 * x/9 - 1/9|^2 \mathbb{1}\{x \geq 0.1\}$$

differentiable, with minimizer near boundary, one side arbitrary vanishing

derivatives, another side positive second order derivative

#### A.4.2. Numerical Results and Comparison with CLS Methods

Now we present the simulation results using the 10 test functions. In particular, we compare our methods with the CLS methods for estimation and confidence intervals for the minimizer.

**Plots and Tables** Before we give a discussion of the results, we explain how we present the results for each function. For each true function, we give the plot of the true function, the time vs log sample size plot (for all three methods), the log empirical risk vs log sample size plot for estimation of the minimizer, log empirical length vs log sample size plot for inference of the minimizer, the log empirical risk vs log sample size plot for estimation

of the minimum, and the log empirical length vs log sample size plot for inference of the minimum. For empirical lengths, we plot for  $\alpha = 0.01$ , other confidence levels are similar. We also provide tables for : CLS empirical coverage for minimizer, log risk for minimizer, and log length for minimizer for  $\alpha = 0.01$ . The plots and tables are shown in figure A.4, A.5, A.6, A.7, A.8, A.9, A.10, A.11, A.12, A.13, A.14, A.15, A.16, A.17, A.18, A.19, A.20, A.21, A.22, A.23.

**Estimation of Minimizer** In general, our methods tie with the CLS method for estimation of minimizer.

For the first five functions in Figure A.3, CLS behave better, for the functions on the third line, the behaviors are almost equal, for the last three functions, ours behave better.

We can see that when compared with CLS estimator our methods behave better at higher smoothness. CLS behaves better when at least one side is (almost) a linear function, which is to the advantage of piece-wise linear approximation. Starting at both sides being twice differentiable (not necessarily with equal second order derivative), our method becomes equal or better. Starting with both sides have vanishing third order derivatives (i.e.  $x^4$  type function), both our methods behave better. We will show in A.4.3 that our methods are stable compared to the benchmarks thus insensitive to the smoothness.

**Inference for Minimizer** For the inference of minimizer, both our methods achieve the nominal coverage. CLS confidence interval does not achieve nominal coverage consistently. For all the functions except the first and sixth function in Figure A.3, CLS confidence interval miss the nominal coverage by far. In A.4.3 we will discuss more on comparison with theoretical results for our methods.

**Estimation for Minimum** The plots show nice decreasing patterns. For the polynomial type functions, we can see a nice linear relationship between log empirical risk and log sample size, which is a good indicator of linear relationship between empirical risk and

benchmark, as benchmark is a power function (with negative power) of sample size. More on comparison with theoretical results is in Section A.4.3.

**Inference for Minimum** Both our methods achieve the nominal coverage in all settings. The plots on empirical length show a nice decreasing pattern. Comparison with theoretical is discussed in Section A.4.3.

**Computing Time** For computing time, we can see that our methods are significantly faster than CLS based methods. For our methods, we measure the total time used for producing all four results, while for CLS based methods, we only measure the time fitting an CLS takes. The time for each function is the sum of time used for 100 replicates. Although this measurement way is in favor of CLS based methods, we can still see that the our methods take much less time.

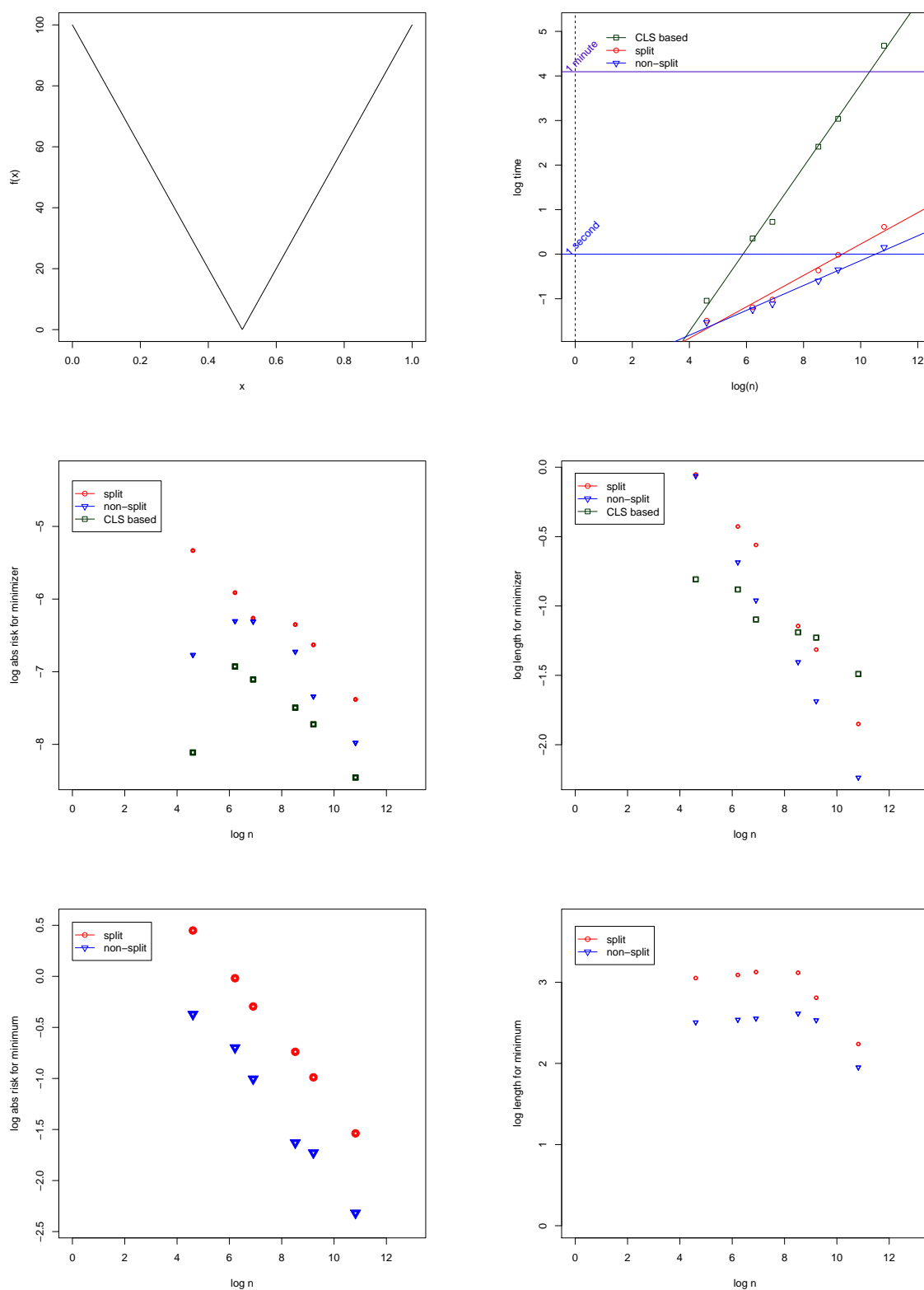


Figure A.4: Plots for  $f_1(x) = 100|2x - 1|$

	100	500	1000	5000	10000	50000
0.2	1	1	0.99	0.95	0.94	0.97
0.1	1	1	0.99	0.97	0.97	0.98
0.05	1	1	0.99	0.98	1	0.99
0.02	1	1	0.99	1	1	0.99
0.01	1	1	1	1	1	1

(a) Empirical coverage of CLS confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-0.053	-0.427	-0.56	-1.144	-1.315	-1.851
<i>non-split</i>	-0.063	-0.685	-0.959	-1.404	-1.686	-2.236
<i>CLS based</i>	-0.808	-0.881	-1.097	-1.19	-1.228	-1.49

(b) Log empirical length of confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-5.332	-5.912	-6.263	-6.351	-6.63	-7.381
<i>non-split</i>	-6.768	-6.303	-6.309	-6.724	-7.339	-7.978
<i>CLS based</i>	-8.112	-6.928	-7.106	-7.495	-7.724	-8.456

(c) Log empirical risk for minimizer

Figure A.5: Tables for  $f_1(x) = 100|2x - 1|$

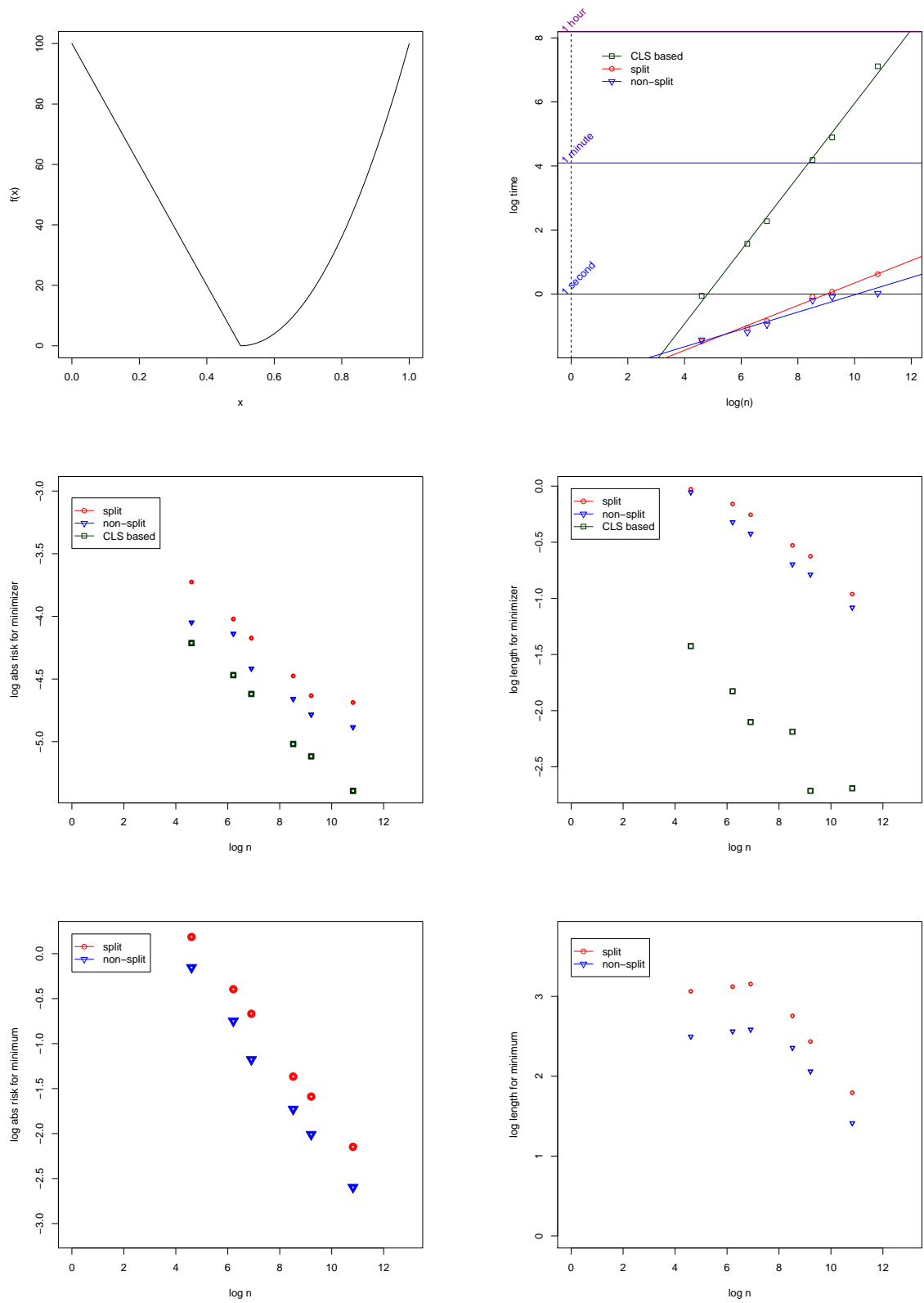


Figure A.6: Plots for  $f_2(x) = 100|2x - 1|\mathbb{1}\{x < 0.5\} + 100|2x - 1|^2\mathbb{1}\{x \geq 0.5\}$

	100	500	1000	5000	10000	50000
0.2	0.71	0.71	0.67	0.66	0.68	0.69
0.1	0.78	0.8	0.75	0.85	0.82	0.78
0.05	0.89	0.88	0.83	0.88	0.87	0.83
0.02	0.95	0.94	0.91	0.97	0.97	0.93
0.01	0.99	0.98	0.97	0.99	1	0.99

(a) Empirical coverage of CLS confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-0.027	-0.159	-0.255	-0.528	-0.625	-0.962
<i>non-split</i>	-0.055	-0.321	-0.424	-0.696	-0.787	-1.081
<i>CLS based</i>	-1.425	-1.827	-2.102	-2.187	-2.714	-2.691

(b) Log empirical length of confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-3.725	-4.021	-4.173	-4.475	-4.633	-4.687
<i>non-split</i>	-4.048	-4.139	-4.416	-4.659	-4.784	-4.884
<i>CLS based</i>	-4.213	-4.469	-4.619	-5.018	-5.117	-5.392

(c) Log empirical risk for minimizer

Figure A.7: Tables for  $f_2(x) = 100|2x - 1|\mathbb{1}\{x < 0.5\} + 100|2x - 1|^2\mathbb{1}\{x \geq 0.5\}$

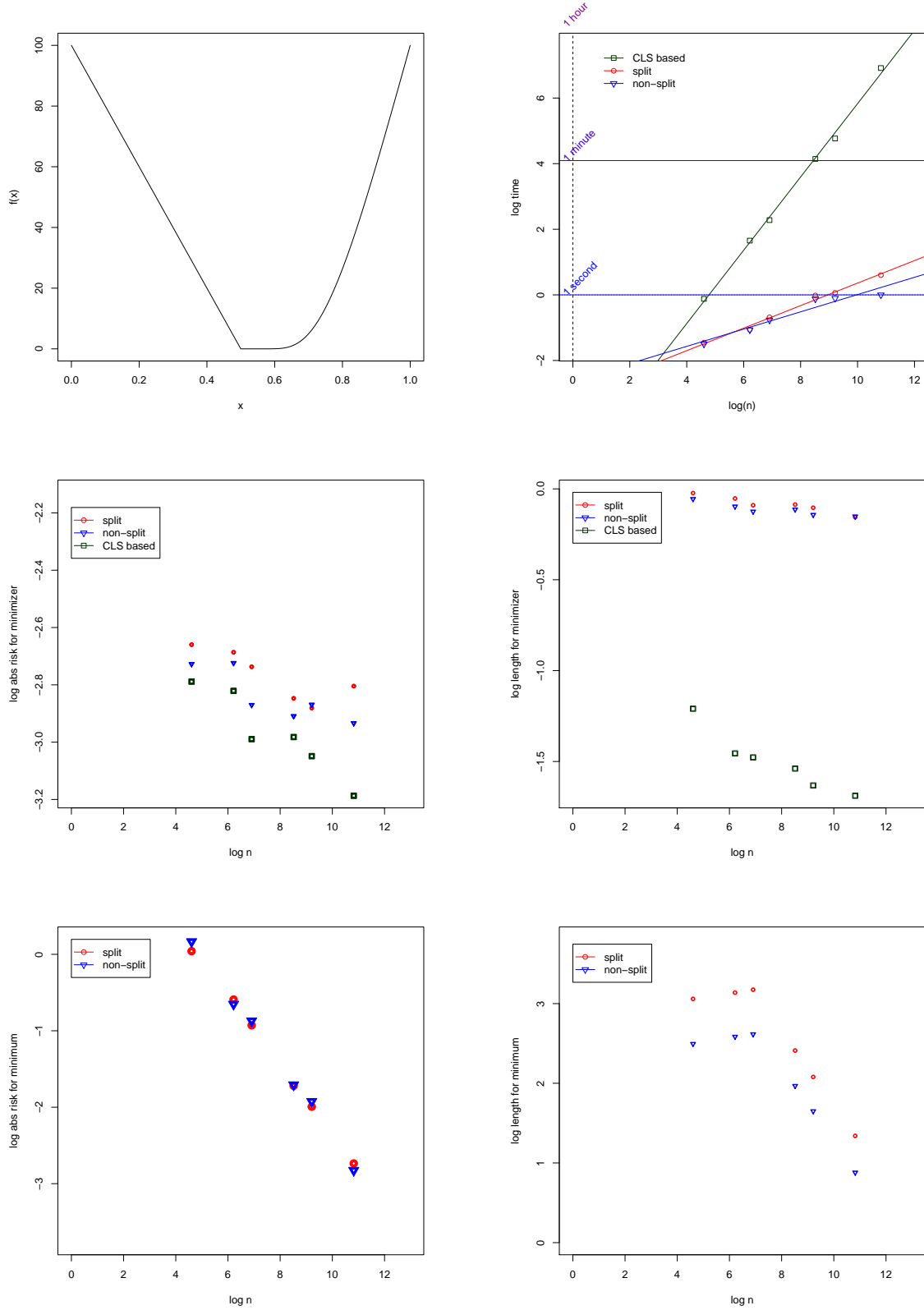


Figure A.8: Plots for  $f_3(x) = 100|2x - 1|\mathbb{1}\{x < 0.5\} + 100 \exp(2 - \frac{1}{|x-0.5|})\mathbb{1}\{x \geq 0.5\}$



	100	500	1000	5000	10000	50000
0.2	<b>0.39</b>	<b>0.34</b>	<b>0.42</b>	<b>0.37</b>	<b>0.37</b>	<b>0.38</b>
0.1	<b>0.47</b>	<b>0.39</b>	<b>0.51</b>	<b>0.43</b>	<b>0.46</b>	<b>0.45</b>
0.05	<b>0.55</b>	<b>0.48</b>	<b>0.58</b>	<b>0.55</b>	<b>0.5</b>	<b>0.54</b>
0.02	<b>0.69</b>	<b>0.68</b>	<b>0.72</b>	<b>0.75</b>	<b>0.67</b>	<b>0.79</b>
0.01	<b>0.88</b>	<b>0.89</b>	<b>0.93</b>	<b>0.91</b>	<b>0.88</b>	<b>0.94</b>

(a) Empirical coverage of CLS confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-0.023	-0.053	-0.09	-0.086	-0.104	-0.156
<i>non-split</i>	-0.056	-0.097	-0.125	-0.113	-0.142	-0.151
<i>CLS based</i>	-1.21	-1.456	-1.477	-1.539	-1.632	-1.688

(b) Log empirical length of confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-2.66	-2.686	-2.737	-2.847	-2.881	-2.805
<i>non-split</i>	-2.727	-2.724	-2.87	-2.909	-2.869	-2.934
<i>CLS based</i>	-2.789	-2.821	-2.989	-2.982	-3.049	-3.187

(c) Log empirical risk for minimizer

Figure A.9: Tables for  $f_3(x) = 100|2x - 1|\mathbb{1}\{x < 0.5\} + 100 \exp(2 - \frac{1}{|x-0.5|})\mathbb{1}\{x \geq 0.5\}$

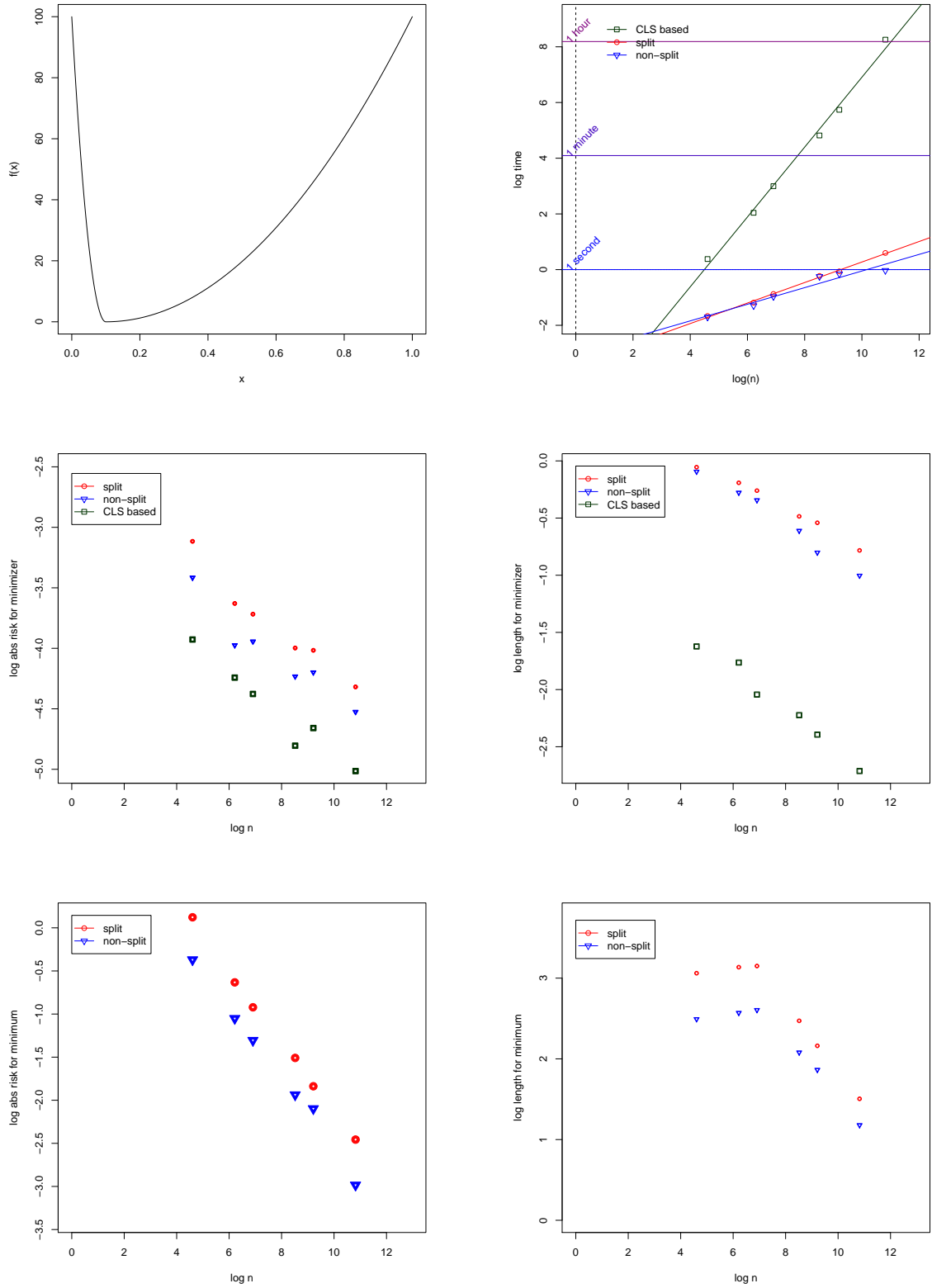


Figure A.10: Plots for  $f_4(x) = 100|10x - 1|^2 \mathbb{1}\{x < 0.1\} + 100|10 \cdot x/9 - 1/9|^2 \mathbb{1}\{x \geq 0.1\}$

	100	500	1000	5000	10000	50000
0.2	<b>0.77</b>	0.82	<b>0.68</b>	0.86	<b>0.72</b>	<b>0.73</b>
0.1	<b>0.84</b>	<b>0.9</b>	<b>0.8</b>	0.92	<b>0.81</b>	<b>0.86</b>
0.05	<b>0.93</b>	<b>0.93</b>	<b>0.88</b>	<b>0.94</b>	<b>0.83</b>	<b>0.87</b>
0.02	<b>0.97</b>	0.99	<b>0.95</b>	<b>0.97</b>	<b>0.94</b>	<b>0.94</b>
0.01	<b>0.99</b>	1	<b>0.98</b>	<b>0.99</b>	<b>0.99</b>	<b>0.97</b>

(a) Empirical coverage of CLS confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-0.055	-0.191	-0.26	-0.485	-0.541	-0.783
<i>non-split</i>	-0.094	-0.277	-0.343	-0.611	-0.801	-1.003
<i>CLS based</i>	-1.623	-1.764	-2.044	-2.224	-2.394	-2.713

(b) Log empirical length of confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-3.116	-3.629	-3.718	-3.997	-4.016	-4.319
<i>non-split</i>	-3.416	-3.974	-3.944	-4.232	-4.199	-4.525
<i>CLS based</i>	-3.927	-4.242	-4.377	-4.805	-4.659	-5.014

(c) Log empirical risk for minimizer

Figure A.11: Tables for  $f_4(x) = 100|10x - 1|^2 \mathbb{1}\{x < 0.1\} + 100|10 \cdot x/9 - 1/9|^2 \mathbb{1}\{x \geq 0.1\}$

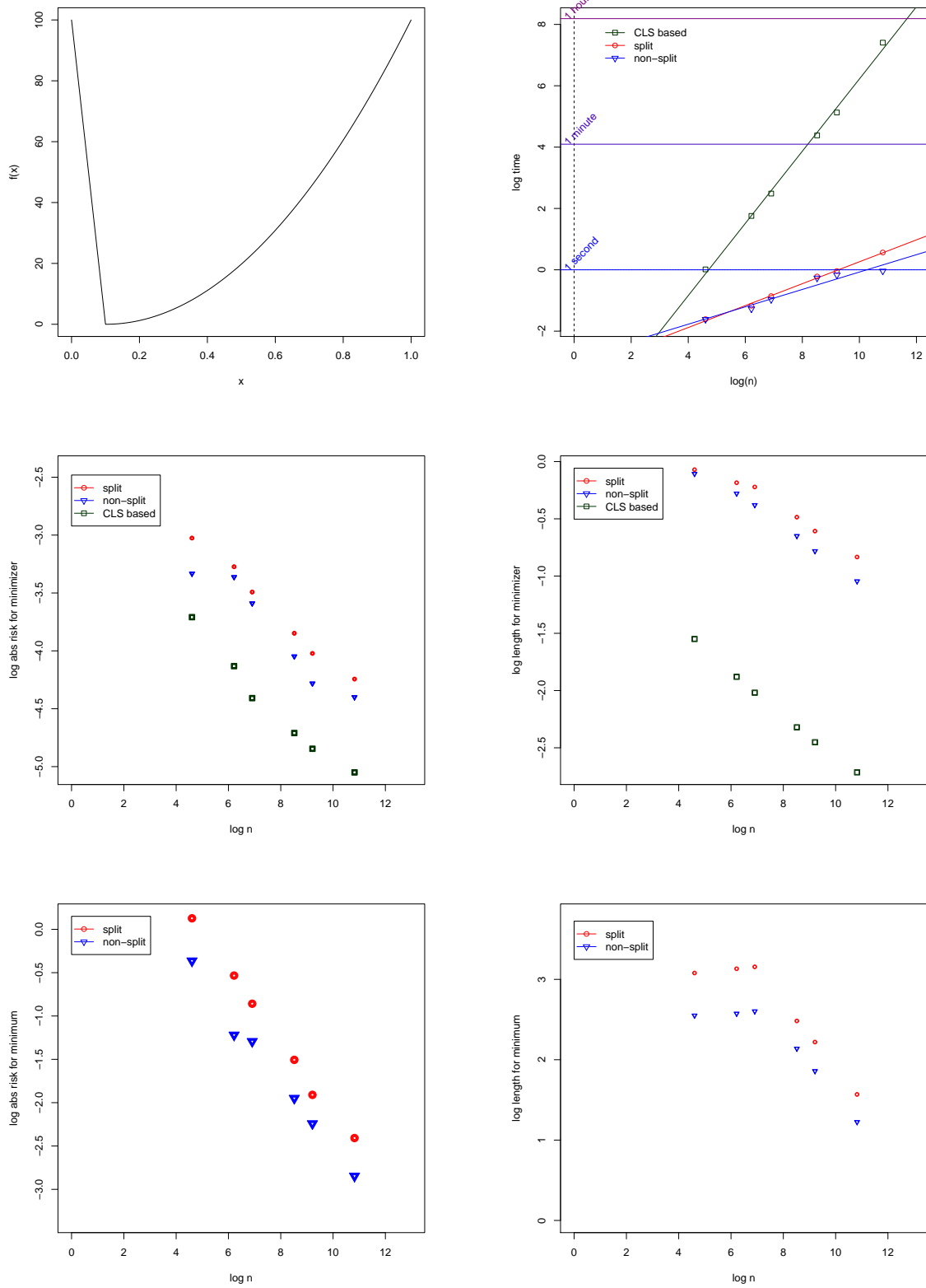


Figure A.12: Plots for  $f_5(x) = 100|10x - 1|^1 \mathbb{1}\{x < 0.1\} + 100|10 \cdot x/9 - 1/9|^2 \mathbb{1}\{x \geq 0.1\}$

	100	500	1000	5000	10000	50000
0.2	0.67	0.69	0.74	0.74	0.8	0.77
0.1	0.82	0.78	0.83	0.84	0.87	0.83
0.05	0.9	0.88	0.86	0.9	0.91	0.84
0.02	0.99	0.95	0.95	0.96	0.96	0.94
0.01	1	0.99	0.97	0.99	1	0.98

(a) Empirical coverage of CLS confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-0.07	-0.184	-0.222	-0.485	-0.607	-0.834
<i>non-split</i>	-0.107	-0.279	-0.379	-0.65	-0.782	-1.045
<i>CLS based</i>	-1.55	-1.88	-2.019	-2.321	-2.452	-2.715

(b) Log empirical length of confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-3.026	-3.274	-3.492	-3.848	-4.022	-4.244
<i>non-split</i>	-3.333	-3.363	-3.59	-4.047	-4.281	-4.4
<i>CLS based</i>	-3.709	-4.131	-4.408	-4.71	-4.846	-5.05

(c) Log empirical risk for minimizer

Figure A.13: Tables for  $f_5(x) = 100|10x - 1|^1 \mathbb{1}\{x < 0.1\} + 100|10 \cdot x/9 - 1/9|^2 \mathbb{1}\{x \geq 0.1\}$

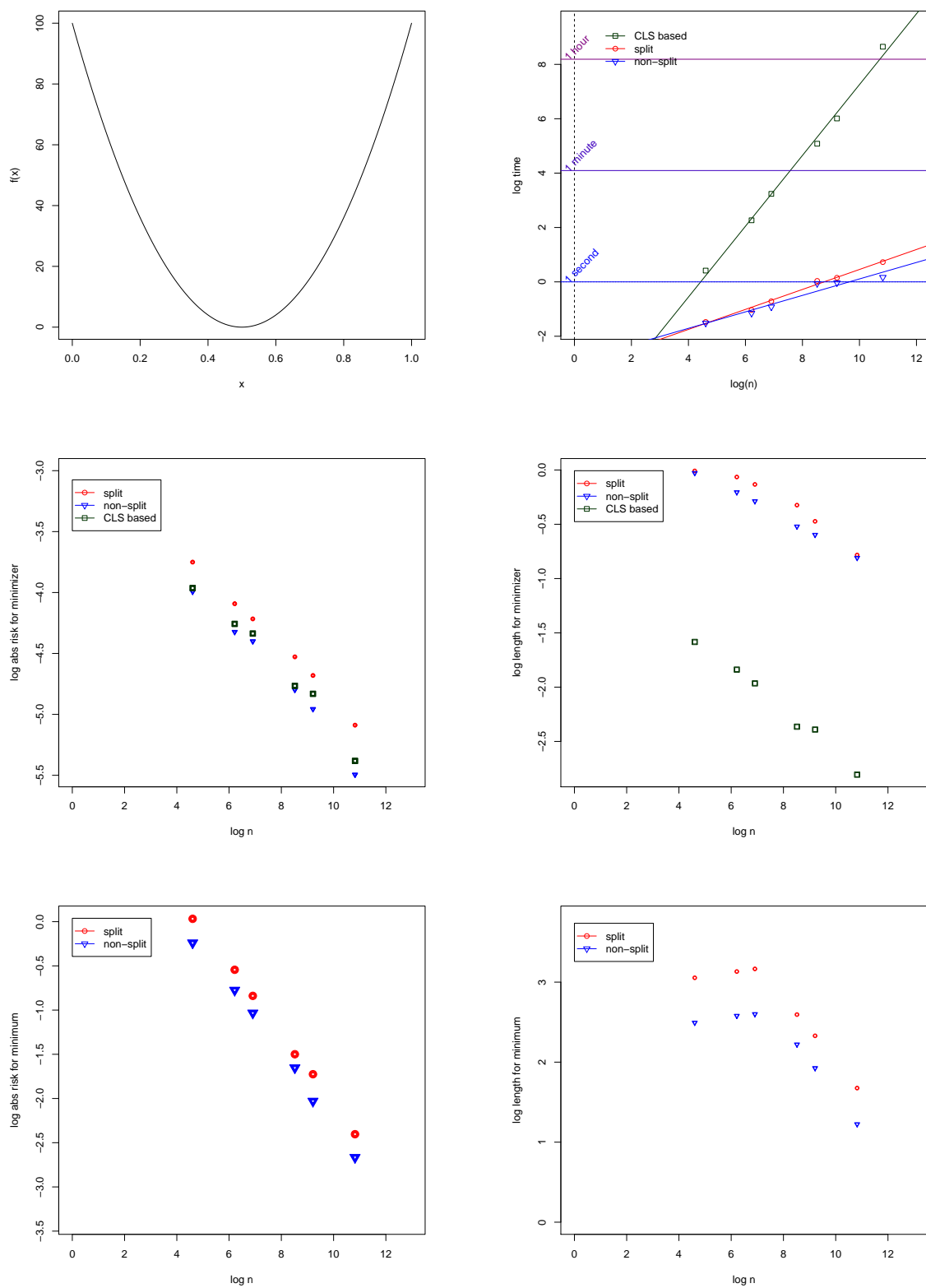


Figure A.14: Plots for  $f_6(x) = 100(|2x - 1|)^2$

	100	500	1000	5000	10000	50000
0.2	0.81	<b>0.79</b>	0.82	0.84	<b>0.76</b>	0.82
0.1	<b>0.88</b>	0.92	0.91	0.91	<b>0.89</b>	0.93
0.05	0.96	0.96	0.97	<b>0.95</b>	0.97	0.97
0.02	0.99	0.99	1	0.99	0.99	0.99
0.01	1	1	1	1	<b>0.99</b>	1

(a) Empirical coverage of CLS confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-0.009	-0.065	-0.133	-0.324	-0.473	-0.783
<i>non-split</i>	-0.028	-0.205	-0.287	-0.52	-0.597	-0.809
<i>CLS based</i>	-1.584	-1.838	-1.965	-2.364	-2.391	-2.806

(b) Log empirical length of confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-3.75	-4.092	-4.216	-4.528	-4.681	-5.088
<i>non-split</i>	-3.993	-4.323	-4.401	-4.799	-4.956	-5.495
<i>CLS based</i>	-3.963	-4.257	-4.337	-4.766	-4.831	-5.382

(c) Log empirical risk for minimizer

Figure A.15: Tables for  $f_6(x) = 100(|2x - 1|)^2$

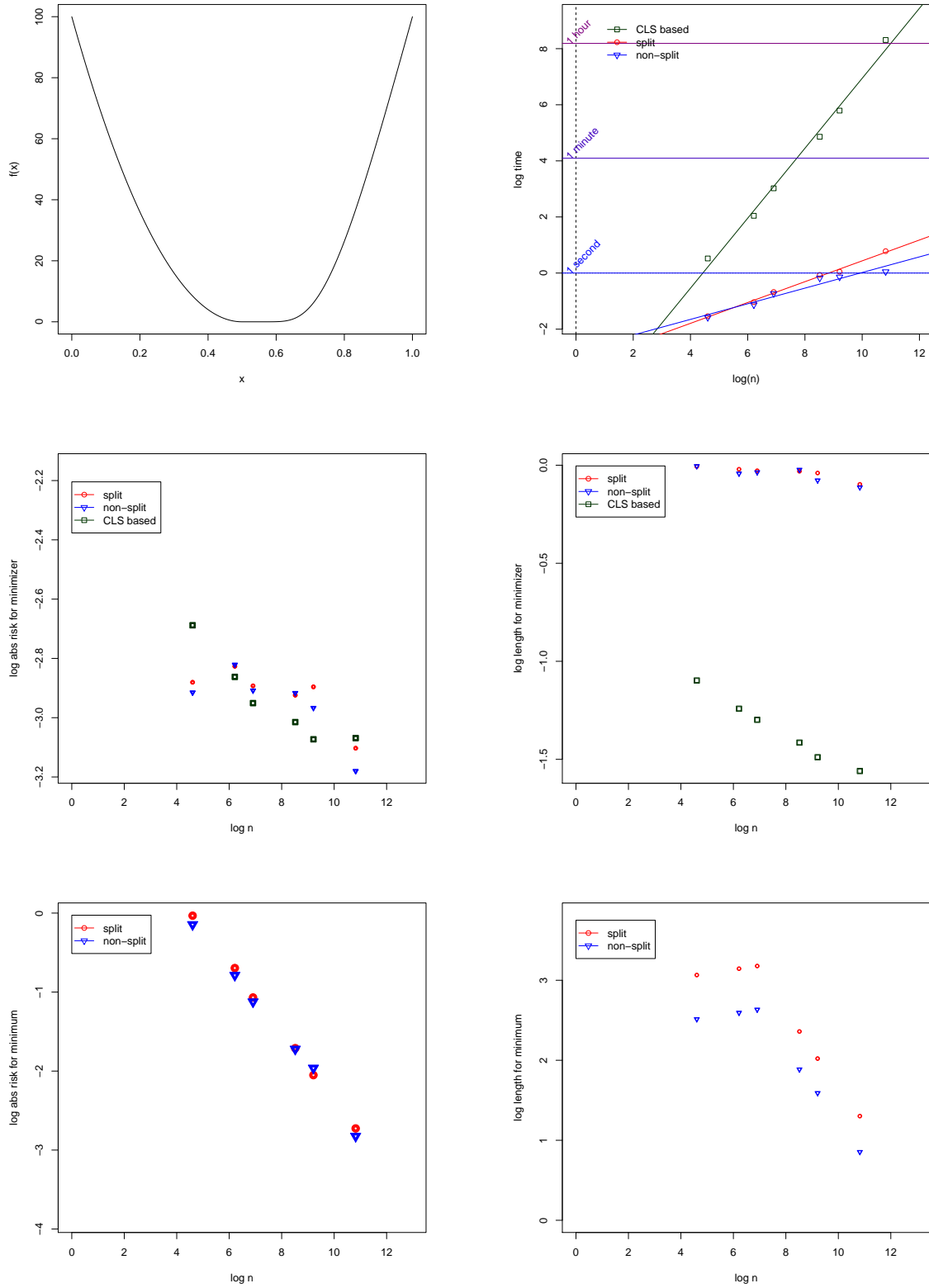


Figure A.16: Plots for  $f_7(x) = 100|2x - 1|^2 \mathbb{1}\{x < 0.5\} + 100 \exp(2 - \frac{1}{|x-0.5|}) \mathbb{1}\{x \geq 0.5\}$



	100	500	1000	5000	10000	50000
0.2	<b>0.39</b>	<b>0.44</b>	<b>0.47</b>	<b>0.43</b>	<b>0.47</b>	<b>0.35</b>
0.1	<b>0.5</b>	<b>0.55</b>	<b>0.54</b>	<b>0.49</b>	<b>0.54</b>	<b>0.44</b>
0.05	<b>0.6</b>	<b>0.62</b>	<b>0.61</b>	<b>0.66</b>	<b>0.6</b>	<b>0.52</b>
0.02	<b>0.83</b>	<b>0.88</b>	<b>0.89</b>	<b>0.81</b>	<b>0.77</b>	<b>0.83</b>
0.01	<b>0.91</b>	<b>0.93</b>	<b>0.98</b>	<b>0.93</b>	<b>0.89</b>	<b>0.96</b>

(a) Empirical coverage of CLS confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-0.007	-0.021	-0.028	-0.031	-0.04	-0.098
<i>non-split</i>	-0.005	-0.042	-0.036	-0.023	-0.077	-0.113
<i>CLS based</i>	-1.098	-1.242	-1.298	-1.415	-1.489	-1.56

(b) Log empirical length of confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-2.881	-2.827	-2.892	-2.924	-2.896	-3.103
<i>non-split</i>	-2.915	-2.821	-2.908	-2.916	-2.967	-3.18
<i>CLS based</i>	-2.688	-2.862	-2.951	-3.015	-3.073	-3.069

(c) Log empirical risk for minimizer

Figure A.17: Tables for  $f_7(x) = 100|2x - 1|^2 \mathbb{1}\{x < 0.5\} + 100 \exp(2 - \frac{1}{|x-0.5|}) \mathbb{1}\{x \geq 0.5\}$

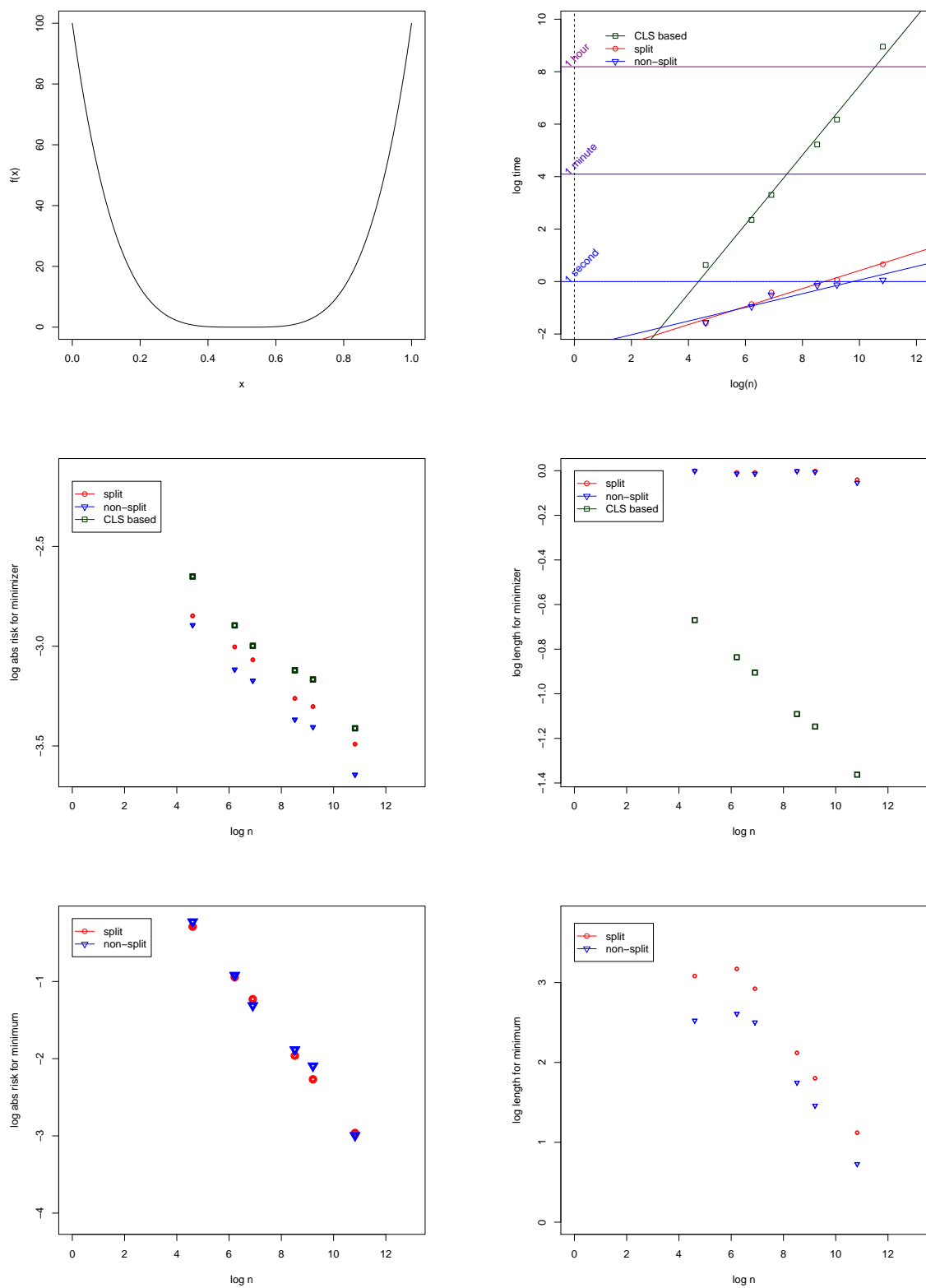


Figure A.18: Plots for  $f_8(x) = 100(|2x-1|^4)$

	100	500	1000	5000	10000	50000
0.2	0.57	0.61	0.67	0.56	0.56	0.62
0.1	0.82	0.83	0.82	0.83	0.77	0.82
0.05	0.89	0.91	0.91	0.91	0.9	0.94
0.02	0.95	0.95	0.96	0.96	0.95	0.96
0.01	0.96	0.99	0.98	0.99	0.97	0.97

(a) Empirical coverage of CLS confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	0	-0.008	-0.008	-0.002	-0.002	-0.04
<i>non-split</i>	-0.001	-0.013	-0.013	-0.002	-0.006	-0.055
<i>CLS based</i>	-0.67	-0.836	-0.905	-1.091	-1.147	-1.363

(b) Log empirical length of confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-2.848	-3.003	-3.068	-3.262	-3.302	-3.49
<i>non-split</i>	-2.894	-3.117	-3.173	-3.368	-3.405	-3.644
<i>CLS based</i>	-2.651	-2.896	-2.998	-3.121	-3.167	-3.411

(c) Log empirical risk for minimizer

Figure A.19: Tables for  $f_8(x) = 100(|2x - 1|)^4$

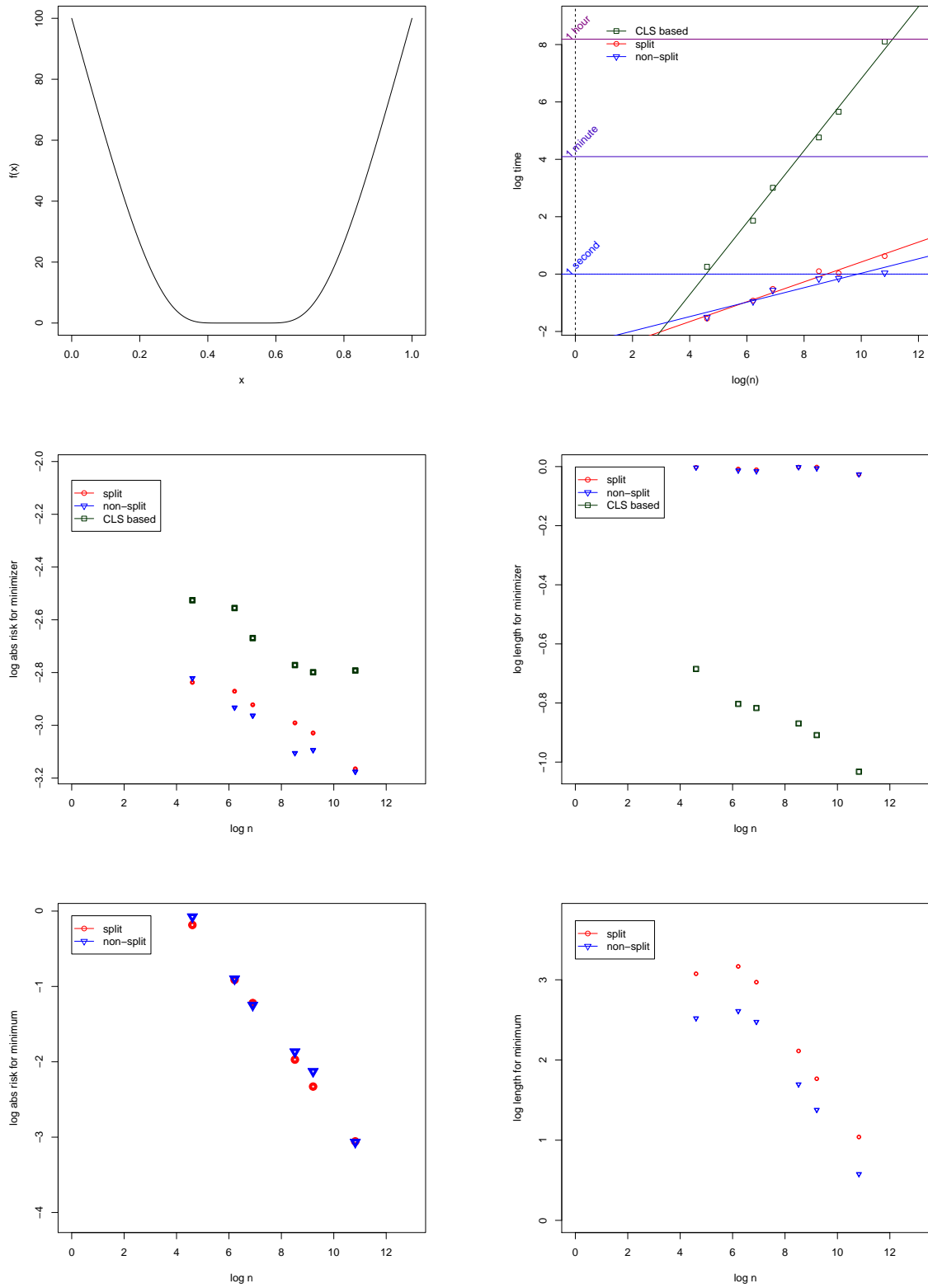


Figure A.20: Plots for  $f_9(x) = 100 \exp(2 - \frac{1}{|x-0.5|})$

	100	500	1000	5000	10000	50000
0.2	<b>0.44</b>	<b>0.39</b>	<b>0.47</b>	<b>0.46</b>	<b>0.46</b>	<b>0.33</b>
0.1	<b>0.73</b>	<b>0.66</b>	<b>0.71</b>	<b>0.76</b>	<b>0.74</b>	<b>0.69</b>
0.05	<b>0.89</b>	<b>0.84</b>	<b>0.84</b>	<b>0.93</b>	<b>0.91</b>	<b>0.86</b>
0.02	<b>0.93</b>	<b>0.9</b>	<b>0.92</b>	<b>0.97</b>	<b>0.95</b>	<b>0.93</b>
0.01	<b>0.95</b>	<b>0.93</b>	<b>0.96</b>	<b>0.97</b>	<b>0.96</b>	<b>0.95</b>

(a) Empirical coverage of CLS confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-0.003	-0.008	-0.011	-0.002	-0.002	-0.027
<i>non-split</i>	-0.003	-0.013	-0.016	-0.002	-0.006	-0.026
<i>CLS based</i>	-0.685	-0.803	-0.817	-0.869	-0.909	-1.033

(b) Log empirical length of confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-2.837	-2.871	-2.922	-2.991	-3.029	-3.165
<i>non-split</i>	-2.821	-2.933	-2.963	-3.105	-3.094	-3.176
<i>CLS based</i>	-2.526	-2.555	-2.669	-2.772	-2.799	-2.792

(c) Log empirical risk for minimizer

Figure A.21: Tables for  $f_9(x) = 100 \exp(2 - \frac{1}{|x-0.5|})$

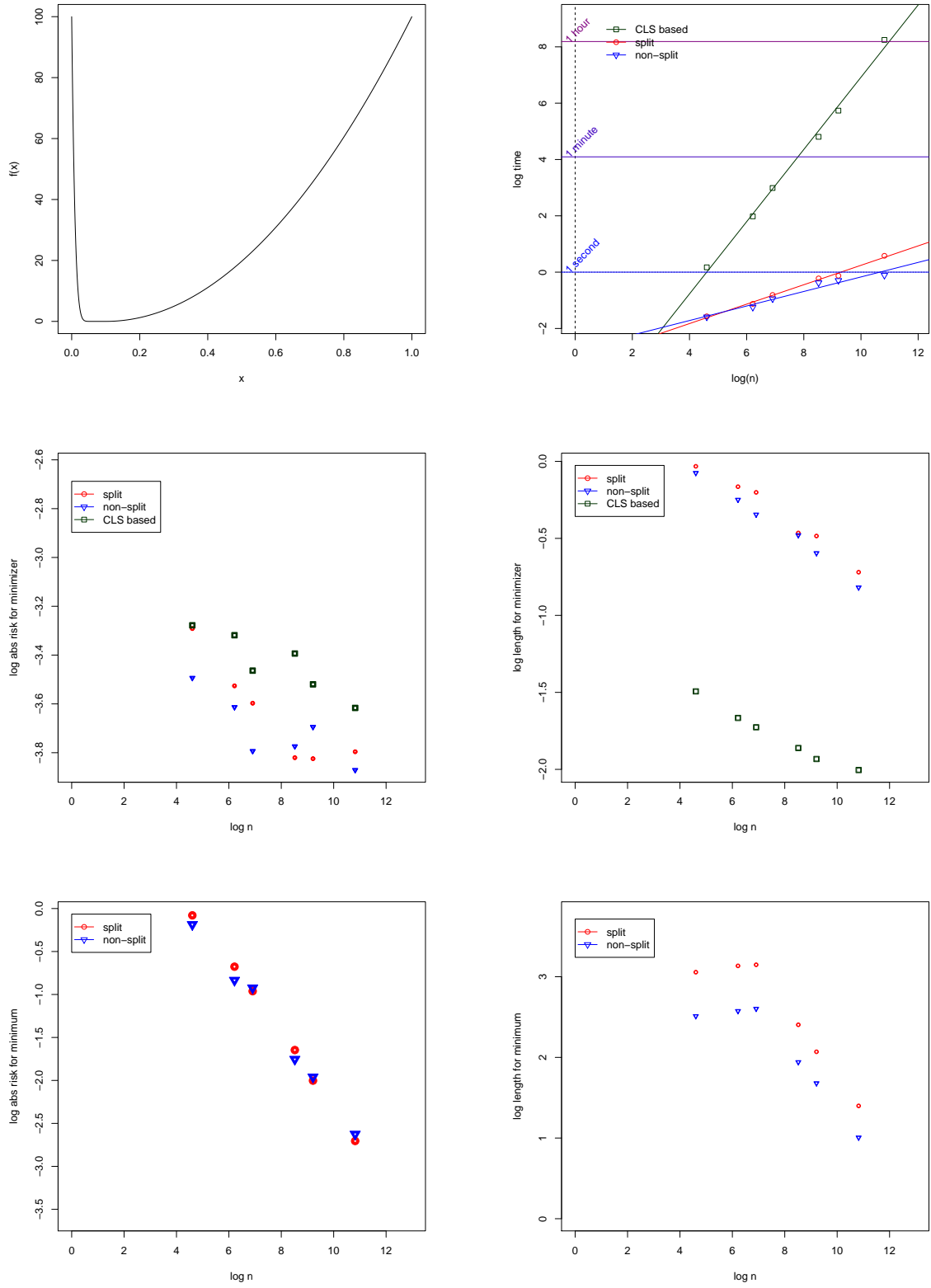


Figure A.22: Plots for  $f_{10}(x) = 100 \exp(2 - \frac{1}{|x-0.1|}) \mathbb{1}\{x < 0.1\} + 100|10 \cdot x/9 - 1/9|^2 \mathbb{1}\{x \geq 0.1\}$

	100	500	1000	5000	10000	50000
0.2	<b>0.58</b>	<b>0.49</b>	<b>0.55</b>	<b>0.39</b>	<b>0.45</b>	<b>0.44</b>
0.1	<b>0.75</b>	<b>0.68</b>	<b>0.62</b>	<b>0.53</b>	<b>0.54</b>	<b>0.53</b>
0.05	<b>0.87</b>	<b>0.83</b>	<b>0.84</b>	<b>0.66</b>	<b>0.71</b>	<b>0.65</b>
0.02	<b>0.95</b>	<b>0.92</b>	<b>0.93</b>	<b>0.85</b>	<b>0.86</b>	<b>0.88</b>
0.01	<b>0.97</b>	<b>0.97</b>	<b>0.96</b>	<b>0.95</b>	<b>0.89</b>	<b>0.97</b>

(a) Empirical coverage of CLS confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-0.032	-0.164	-0.201	-0.465	-0.485	-0.72
<i>non-split</i>	-0.075	-0.249	-0.345	-0.479	-0.596	-0.819
<i>CLS based</i>	-1.494	-1.666	-1.727	-1.861	-1.933	-2.004

(b) Log empirical length of confidence interval for minimizer

	100	500	1000	5000	10000	50000
<i>split</i>	-3.291	-3.526	-3.597	-3.82	-3.824	-3.796
<i>non-split</i>	-3.493	-3.613	-3.793	-3.774	-3.694	-3.871
<i>CLS based</i>	-3.278	-3.319	-3.464	-3.394	-3.52	-3.616

(c) Log empirical risk for minimizer

Figure A.23: Tables for  $f_{10}(x) = 100 \exp(2 - \frac{1}{|x-0.1|}) \mathbb{1}\{x < 0.1\} + 100|10 \cdot x/9 - 1/9|^2 \mathbb{1}\{x \geq 0.1\}$

### A.4.3. Comparison with Benchmarks

In this subsection, we consider the functions where the benchmarks can be explicitly calculated. The primary task is to investigate the relationship between empirical risks/lengths and the benchmarks.

We consider a different set of functions whose benchmarks can be easily calculated:

$$\begin{aligned} h_1(t) &= 100|t - 0.5|, \\ h_2(t) &= 200|2(t - 0.5)|^{\frac{3}{2}}, \\ h_3(t) &= 200|2(t - 0.5)|^2, \\ h_4(t) &= 200|2(t - 0.5)|^3, \\ h_5(t) &= 200|2(t - 0.5)|^4. \end{aligned} \tag{A.4.2}$$

All other settings are the same as before except that we take roughly exponentially equally spaced sample sizes.

We calculated the corresponding benchmarks (the discretization errors in these examples are negligible):  $\rho_z(\sqrt{1/n}; f)$  and  $\rho_m(\sqrt{1/n}; f)$ .

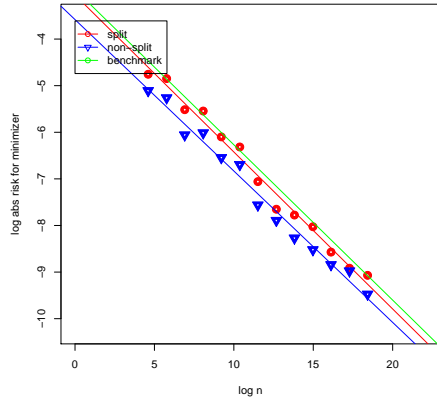
The plots of log risk/length vs log sample size for minimizer and minimum with reference line of benchmark are shown in Figures A.24, A.25, A.26, A.27. For estimation of minimizer, in addition to the almost identical slope with reference line (i.e. linear relationship between empirical risk and benchmark), the intercept difference of the reference line and the log risk of non-split version ranges between 0.6472699 and 1.036388, meaning that  $\frac{\rho_z(\sqrt{1/n}; f)}{\text{empirical risk for minimizer}}$  for non-split version ranges in  $[1.910318, 2.819016]$ , implying that the performance of non-split version is quite robust when smoothness varies.

For other three tasks, excluding the outlier points that are apparently influenced by the truncation for confidence interval, the slopes of the methods and the reference line are

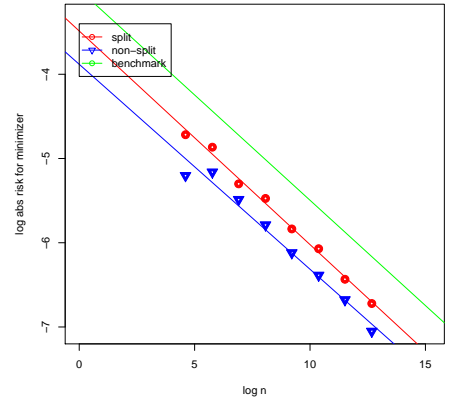


almost identical.

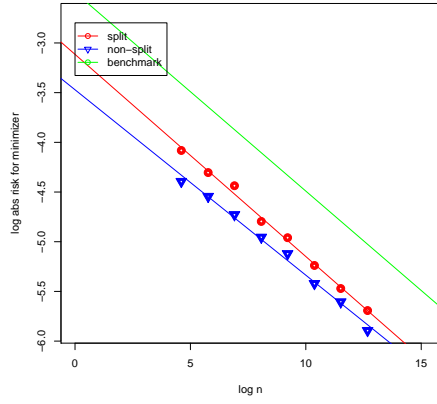
The empirical performance, therefore, agree with the theoretical results.



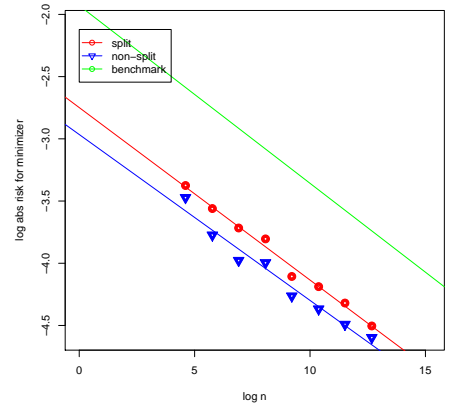
(a)  $h_1$



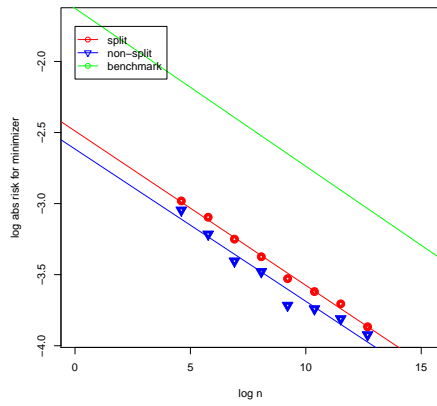
(b)  $h_2$



(c)  $h_3$

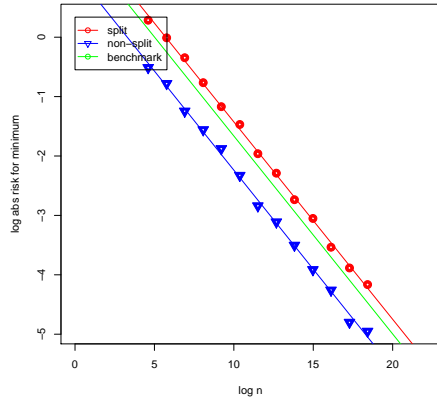


(d)  $h_4$

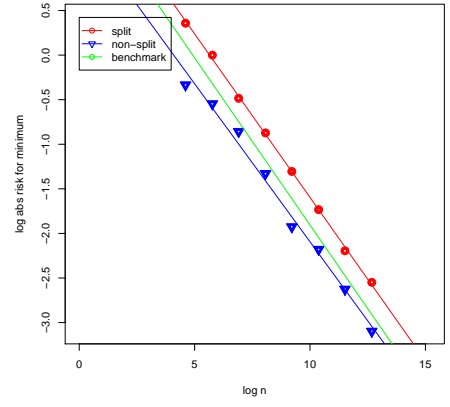


(e)  $h_5$

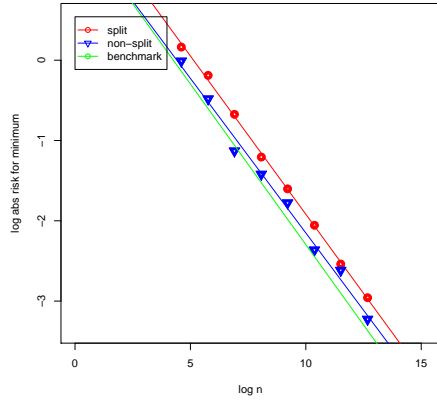
Figure A.24: Empirical risks for minimizer



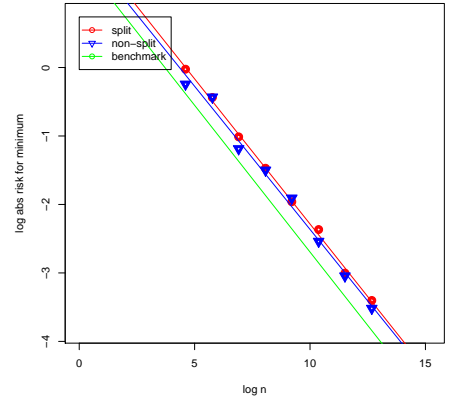
(a)  $h_1$



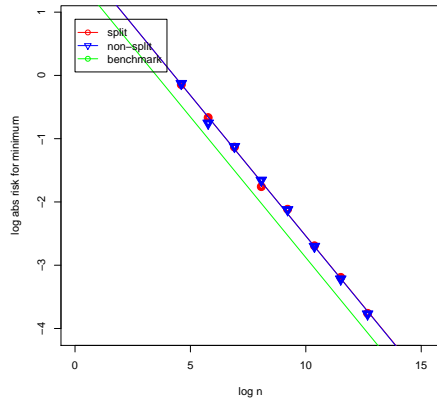
(b)  $h_2$



(c)  $h_3$

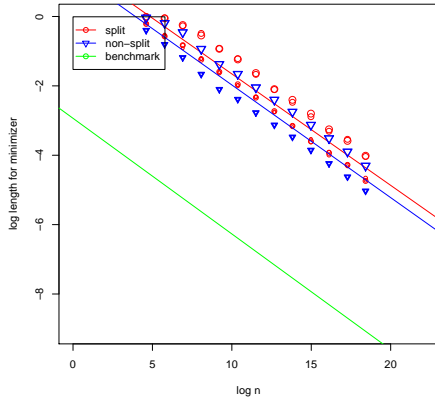


(d)  $h_4$

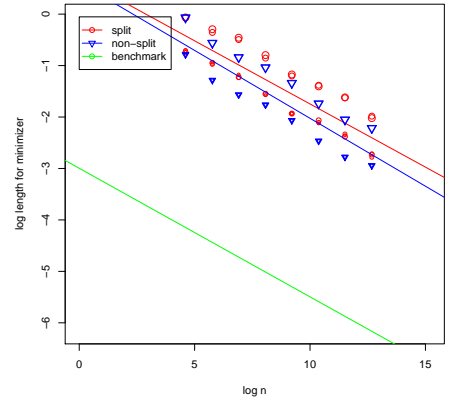


(e)  $h_5$

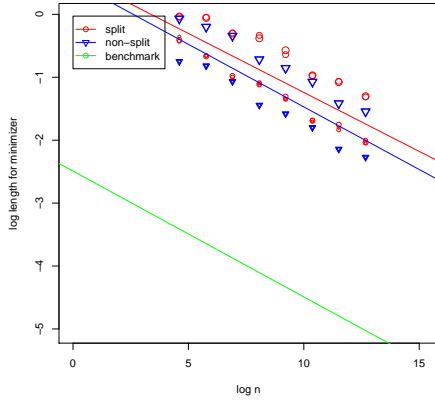
Figure A.25: Empirical risks for minimum



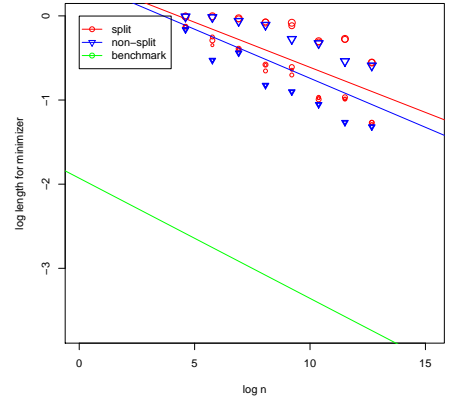
(a)  $h_1$



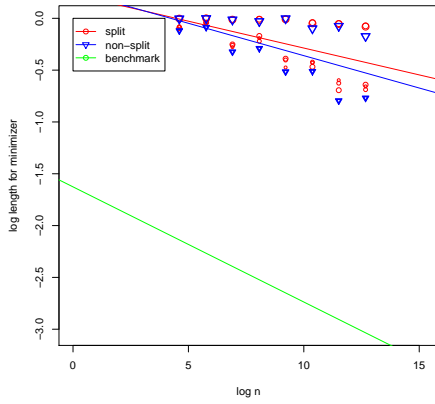
(b)  $h_2$



(c)  $h_3$

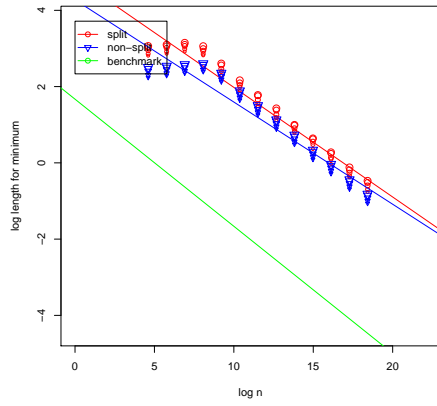


(d)  $h_4$

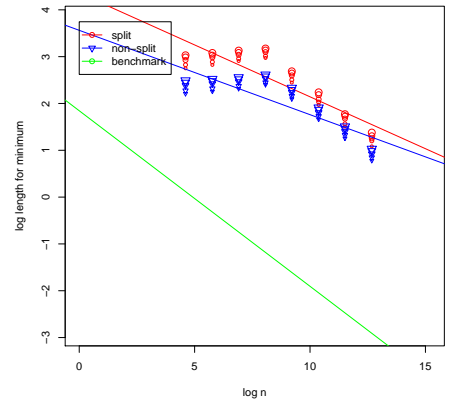


(e)  $h_5$

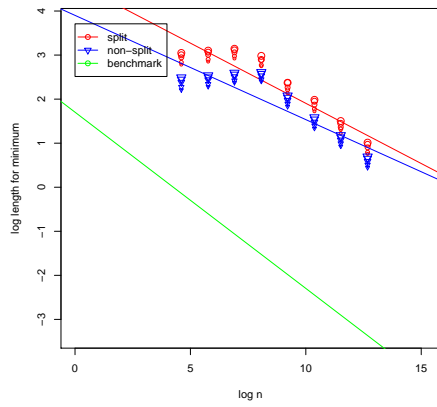
Figure A.26: Empirical lengths for minimizer



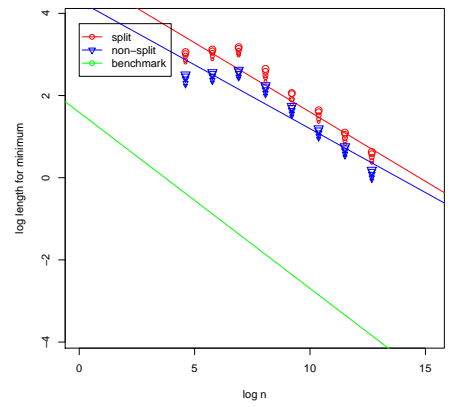
(a)  $h_1$



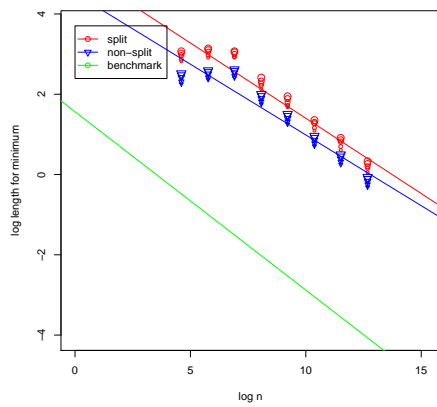
(b)  $h_2$



(c)  $h_3$



(d)  $h_4$



(e)  $h_5$

Figure A.27: Empirical lengths for minimum

## A.5. Proofs of the Results in Chapter 3

### A.5.1. Notation

Here we recollect or introduce notation that will be used later. We use  $Z(f)$ ,  $M(f)$  to denote the minimizer and minimum of function  $f$ , where  $f$  can be univariate or multivariate.

Recall that

$$\begin{aligned}\rho_m(\varepsilon; f) &= \max\{\rho : \int_0^1 (\max\{\rho, f(t)\} - f(t))^2 dt \leq \varepsilon^2\} - M(f) \\ \rho_z(\varepsilon; f) &= \max\{|t - Z(f)| : f(t) \leq \rho_m(\varepsilon; f) + M(f)\}.\end{aligned}\tag{A.5.1}$$

for  $f \in \mathcal{F}$ .

### A.5.2. Proof of Theorem 3.2.1

For the ease of notation, denote  $\mathcal{D}$  to be  $[0, 1]^s$ .

We start with minimizer. We start with lower bounds.

Let  $\mathbf{f} \in \mathcal{F}_s$ . Let  $\mathbf{g} \in \mathcal{F}_s$ , which we will specify later. Take  $\theta \in \{-1, 1\}$  as parameter to be estimated, with  $\mathbf{f}_1 = \mathbf{f}$  and  $\mathbf{f}_{-1} = \mathbf{g}$ .

For any estimator  $\hat{Z}$  for estimating the minimizer, consider the projected estimator that projects  $\hat{Z}$  to the line determined by  $Z(\mathbf{f})$  and  $Z(\mathbf{g})$  :

$$\hat{Z}_p = Z(\mathbf{f}) + \langle \hat{Z} - Z(\mathbf{f}), \frac{Z(\mathbf{g}) - Z(\mathbf{f})}{\|Z(\mathbf{f}) - Z(\mathbf{g})\|} \rangle.\tag{A.5.2}$$

It's easy to see that

$$E_{\mathbf{f}} \left( \|\hat{Z}_p - Z(\mathbf{f})\|^2 \right) \leq E_{\mathbf{f}} \left( \|\hat{Z} - Z(\mathbf{f})\|^2 \right)$$

and

$$E_{\mathbf{g}} \left( \|\hat{Z}_p - Z(\mathbf{g})\|^2 \right) \leq E_{\mathbf{g}} \left( \|\hat{Z} - Z(\mathbf{g})\|^2 \right).$$

Therefore, we only need to consider the projected estimators  $\hat{Z}_p$  for calculating  $R_z(\varepsilon; \mathbf{f})$ . Similarly, we only need to consider projected confidence hypercube  $CI_p$  is the smallest hypercube containing  $\{Z(\mathbf{f}) + \langle \mathbf{t} - Z(\mathbf{f}), \frac{Z(\mathbf{g}) - Z(\mathbf{f})}{\|Z(\mathbf{g}) - Z(\mathbf{f})\|} \rangle : \mathbf{t} \in CI\}$  for calculating  $L_{\alpha, z}(\varepsilon; \mathbf{f})$ , as projection does not weaken confidence level and projected hypercube has smaller hypercube-diameter.

Note that any projected estimator  $\hat{Z}_p$  of the minimizer  $Z(\mathbf{f}_\theta)$  gives an estimator of  $\theta$  by

$$\hat{\theta} = \left\langle \frac{\hat{Z}_p - \frac{Z_p(\mathbf{f}_1) + Z_p(\mathbf{f}_{-1})}{2}}{\left\| \frac{Z_p(\mathbf{f}_1) - Z_p(\mathbf{f}_{-1})}{2} \right\|}, \frac{Z_p(\mathbf{f}_1) - Z_p(\mathbf{f}_{-1})}{\|Z_p(\mathbf{f}_1) - Z_p(\mathbf{f}_{-1})\|} \right\rangle,$$

and therefore  $\mathbb{E}_\theta \|\hat{Z}_p - Z(\mathbf{f}_\theta)\|^2 = \|Z(\mathbf{f}_1) - Z(\mathbf{f}_{-1})\|^2 \mathbb{E}_\theta \frac{|\hat{\theta} - \theta|^2}{2}$ . Let  $\mathbb{P}_\theta$  be the probability measure associated with the white noise model corresponding to  $\mathbf{f}_\theta$ . On the other hand, through calculating the Radon-Nikodym derivative  $\frac{d\mathbb{P}_1}{d\mathbb{P}_{-1}}(Y)$  by Girsanov's theorem,

$$\frac{dP_{\mathbf{f}}}{dP_{\mathbf{g}}}(Y) = \exp \left( \int_{\mathcal{D}} \frac{\mathbf{f}(\mathbf{t}) - \mathbf{g}(\mathbf{t})}{\varepsilon^2} dY(\mathbf{t}) - \frac{1}{2} \int_{\mathcal{D}} \frac{\mathbf{f}(\mathbf{t})^2 - \mathbf{g}(\mathbf{t})^2}{\varepsilon^2} d\mathbf{t} \right), \quad (\text{A.5.3})$$

a sufficient statistic for  $\theta$  is given by

$$W = \frac{\int_{\mathcal{D}} (\mathbf{f}_1(\mathbf{t}) - \mathbf{f}_{-1}(\mathbf{t})) dY(\mathbf{t}) - \frac{1}{2} \int_{\mathcal{D}} (\mathbf{f}_1(\mathbf{t})^2 - \mathbf{f}_{-1}(\mathbf{t})^2) d\mathbf{t}}{\varepsilon \|\mathbf{f}_1 - \mathbf{f}_{-1}\|}. \quad (\text{A.5.4})$$

Then

$$W \sim N \left( \frac{\theta}{2} \cdot \frac{\|\mathbf{f}_1 - \mathbf{f}_{-1}\|}{\varepsilon}, 1 \right) \quad \text{under } \mathbb{P}_\theta.$$

Note that for any  $\omega_z(\varepsilon; \mathbf{f}) > \delta > 0$ , there exists  $\mathbf{h}_\delta \in \mathcal{F}_s$  such that  $\|\mathbf{f} - \mathbf{h}_\delta\| = \varepsilon$  and that  $\|Z(\mathbf{f}) - Z(\mathbf{h}_\delta)\|^2 \geq \omega_z(\varepsilon; \mathbf{f}) - \delta$ , we let  $\mathbf{g} = \mathbf{h}_\delta$ . Then we have  $R_z(\varepsilon; f) \geq (\omega_z(\varepsilon; \mathbf{f}) - \delta) \cdot r_2$ ,

where  $r_2$  is the minimax risk of the two-point problem based on an observation  $X \sim N(\frac{\theta}{2}, 1)$ ,

$$r_2 = \inf_{\hat{\theta}} \max_{\theta=\pm 1} \mathbb{E}_{\theta} \frac{|\hat{\theta} - \theta|^2}{4}.$$

Elementary calculation shows that  $r_2 \geq 0.1$ . Taking  $\delta \rightarrow 0^+$ , we have  $R_z(\varepsilon; \mathbf{f}) \geq 0.1\omega_z(\varepsilon; \mathbf{f})$ .

So we have  $a \geq 0.1$ .

Now we turn to the upper bounds. We start with stating a property of  $\omega_z(\varepsilon; \mathbf{f})$  in Proposition A.5.1.

**Proposition A.5.1.** *Suppose  $\mathbf{f} \in \mathcal{F}_s$ ,  $c \in (0, 1)$ , then we have*

$$\omega_z(\varepsilon; \mathbf{f}) \geq \omega_z(c\varepsilon; \mathbf{f}) \geq \frac{1}{9} \max \left\{ \left(\frac{c}{2}\right)^{\frac{2}{3}}, c \right\} \omega_z(\varepsilon; \mathbf{f}). \quad (\text{A.5.5})$$

*Proof.* The left hand side is apparent, we will prove the right hand side. Using Proposition A.5.3, we have

$$\begin{aligned} \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i c\varepsilon; f_i)^2 &\leq \omega_z(c\varepsilon; \mathbf{f}) \leq 9 \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i c\varepsilon; f_i)^2, \\ \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i \varepsilon; f_i)^2 &\leq \omega_z(\varepsilon; \mathbf{f}) \leq 9 \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i \varepsilon; f_i)^2. \end{aligned} \quad (\text{A.5.6})$$

Using Proposition 2.2.1 in Chapter 2, namely

$$\max \left\{ \left(\frac{q}{2}\right)^{\frac{2}{3}}, q \right\} \leq \frac{\rho_z(q\varepsilon; f)}{\rho_z(\varepsilon; f)} \leq 1, \text{ for } q \in [0, 1]$$

, we know  $\rho_z(\varepsilon; f)$  is a continuous function of  $\varepsilon \geq 0$  for  $f \in \mathcal{F}$ . So there exists  $(\tilde{b}_1, \dots, \tilde{b}_s)$  and  $(\bar{b}_1, \dots, \bar{b}_s)$  attaining the suprema:

$$\begin{aligned} \tilde{b}_i &\geq 0, \text{ for } 1 \leq i \leq s, \sum_{i=1}^s \tilde{b}_i^2 = 1, \sum_{i=1}^s \rho_z(\tilde{b}_i c\varepsilon; f_i)^2 = \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i c\varepsilon; f_i)^2, \\ \bar{b}_i &\geq 0, \text{ for } 1 \leq i \leq s, \sum_{i=1}^s \bar{b}_i^2 = 1, \sum_{i=1}^s \rho_z(\bar{b}_i \varepsilon; f_i)^2 = \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i \varepsilon; f_i)^2. \end{aligned} \quad (\text{A.5.7})$$



Also we have

$$\sum_{i=1}^s \rho_z(\tilde{b}_i c \varepsilon; f_i)^2 \leq \sum_{i=1}^s \rho_z(\tilde{b}_i \varepsilon; f_i)^2 \leq \sum_{i=1}^s \rho_z(\bar{b}_i \varepsilon; f_i)^2, \quad (\text{A.5.8})$$

and

$$\sum_{i=1}^s \rho_z(\tilde{b}_i c \varepsilon; f_i)^2 \geq \sum_{i=1}^s \rho_z(\bar{b}_i c \varepsilon; f_i)^2 \geq \sum_{i=1}^s \max \left\{ \left( \frac{c}{2} \right)^{\frac{2}{3}}, c \right\} \rho_z(\bar{b}_i \varepsilon; f_i)^2. \quad (\text{A.5.9})$$

Combining equations (A.5.6), (A.5.8), (A.5.9) we have

$$\omega_z(c \varepsilon; \mathbf{f}) \geq \frac{1}{9} \max \left\{ \left( \frac{c}{2} \right)^{\frac{2}{3}}, c \right\} \omega_z(\varepsilon; \mathbf{f}). \quad (\text{A.5.10})$$

□

Now we continue with the upper bounds.

Recalling  $W$  define in (A.5.4), let

$$\hat{Z} = \text{sign}(W) \cdot \frac{Z(\mathbf{f}) - Z(\mathbf{g})}{2} + \frac{Z(\mathbf{f}) + Z(\mathbf{g})}{2}. \quad (\text{A.5.11})$$

Then

$$\mathbb{E}_{\mathbf{f}}(\|\hat{Z} - Z(\mathbf{f})\|^2) = \mathbb{E}_{\mathbf{g}}(\|\hat{Z} - Z(\mathbf{g})\|^2) = \|Z(\mathbf{f}) - Z(\mathbf{g})\|^2 \Phi\left(-\frac{\|\mathbf{f} - \mathbf{g}\|}{2\varepsilon}\right). \quad (\text{A.5.12})$$

Therefore,

$$\begin{aligned} R_z(\varepsilon; f) &\leq \sup_{f \in \mathcal{F}_s} \|Z(f) - Z(g)\|^2 \Phi\left(-\frac{\|\mathbf{f} - \mathbf{g}\|}{2\varepsilon}\right) \\ &\leq \sup_{c > 0} \omega_z(c \varepsilon; \mathbf{f}) \Phi\left(-\frac{c}{2}\right) \\ &\leq \max\{0.5 \omega_z(\varepsilon; \mathbf{f}), \sup_{c \geq 1} \omega_z(c \varepsilon; \mathbf{f}) \Phi\left(-\frac{c}{2}\right)\}. \end{aligned} \quad (\text{A.5.13})$$

In addition

$$\sup_{c \geq 1} \omega_z(c\varepsilon; \mathbf{f}) \Phi\left(-\frac{c}{2}\right) \leq 9 \sup_{c \geq 1} \min\{(2c)^{\frac{2}{3}}, c\} \Phi\left(-\frac{c}{2}\right) \omega_z(\varepsilon; \mathbf{f}) \leq 3.1 \omega_z(\varepsilon; \mathbf{f}). \quad (\text{A.5.14})$$

Take  $A = 3.1$  gives the result.

Now we turn to the minimum and start with estimation. We start with the lower bound.

Recall that  $W$  defined in (A.5.4) is a sufficient statistics for  $\theta$ .

Then similarly to the proof of that for minimizer we have that

$$R_m(\varepsilon; \mathbf{f}) \geq a \omega_m(\varepsilon; \mathbf{f}). \quad (\text{A.5.15})$$

For the upper bound. We start with a proposition.

**Proposition A.5.2.** *For  $c > 1$ , we have*

$$\omega_m(c\varepsilon; \mathbf{f}) \leq c^2 \omega_m(\varepsilon; \mathbf{f}), \tilde{\omega}_m(c\varepsilon; \mathbf{f}) \leq c \tilde{\omega}_m(\varepsilon; \mathbf{f}). \quad (\text{A.5.16})$$

*Proof.* Suppose  $\mathbf{g}$  satisfies  $\|\mathbf{g} - \mathbf{f}\|_2 \leq c\varepsilon$ . Then calculation show that

$$|g_0 - f_0|^2 + \sum_{i=1}^s \|g_i - f_i\|^2 \leq c^2 \varepsilon^2, \quad (\text{A.5.17})$$

Let  $h_i(t) = \frac{1}{c} g_i(t) + \frac{c-1}{c} f_i(t)$ . Let  $\mathbf{h}(\mathbf{t}) = \frac{1}{c} g_0 + \frac{c-1}{c} f_0 + \sum_{i=1}^s h_i(t_i)$  Then we have that

$$\|\mathbf{h} - \mathbf{f}\|^2 \leq \varepsilon^2, \quad (\text{A.5.18})$$

and that

$$|M(\mathbf{h}) - M(\mathbf{f})| = \frac{1}{c} |M(\mathbf{g}) - M(\mathbf{f})|. \quad (\text{A.5.19})$$

This gives the statement of the proposition.

□

Recalling  $W$  define in (A.5.4), let

$$\hat{M} = \text{sign}(W) \cdot \frac{M(\mathbf{f}) - M(\mathbf{g})}{2} + \frac{M(\mathbf{f}) + M(\mathbf{g})}{2}. \quad (\text{A.5.20})$$

Then

$$\mathbb{E}_{\mathbf{f}}(\|\hat{M} - M(\mathbf{f})\|^2) = \mathbb{E}_{\mathbf{g}}(\|\hat{M} - M(\mathbf{g})\|^2) = \|M(\mathbf{f}) - M(\mathbf{g})\|^2 \Phi\left(-\frac{\|\mathbf{f} - \mathbf{g}\|}{2\varepsilon}\right). \quad (\text{A.5.21})$$

With Proposition A.5.2 we have that

$$\begin{aligned} R_m(\varepsilon; \mathbf{f}) &\leq \sup_{c>0} \omega_m(c\varepsilon; \mathbf{f}) \Phi\left(-\frac{c}{2}\right) \leq \max\{0.5\omega_m(\varepsilon; \mathbf{f}), \sup_{c\geq 1} \omega_m(c\varepsilon; \mathbf{f}) \Phi\left(-\frac{c}{2}\right)\} \\ &\leq \omega_m(\varepsilon; \mathbf{f}) \max\{0.5, \sup_{c\geq 1} c^2 \Phi\left(-\frac{c}{2}\right)\} \leq \omega_m(\varepsilon; \mathbf{f}). \end{aligned} \quad (\text{A.5.22})$$

For the inference of the minimum, we again start with the lower bound.

$$\begin{aligned} L_{\alpha,m}(\varepsilon; \mathbf{f}) &\geq \sup_{\mathbf{g} \in \mathcal{F}_s} \inf_{CI_{m,\alpha} \in \mathcal{I}_{m,\alpha}(\mathbf{f}, \mathbf{g})} \mathbb{P}_{\mathbf{f}}(\{M(\mathbf{g}), M(\mathbf{f})\} \in CI_{m,\alpha}) |M(\mathbf{f}) - M(\mathbf{g})| \\ &\geq \sup_{\mathbf{g} \in \mathcal{F}_s, \|\mathbf{g} - \mathbf{f}\| \leq \varepsilon} (1 - \alpha - \mathbb{P}_{\mathbf{f}}(M(\mathbf{g}) \notin \mathcal{I}_{m,\alpha}(\mathbf{f}, \mathbf{g}))) \tilde{\omega}_m(\varepsilon; \mathbf{f}) \\ &\geq (1 - \alpha - \Phi(-z_\alpha + 1)) \tilde{\omega}_m(\varepsilon; \mathbf{f}) \geq (0.6 - \alpha) \tilde{\omega}_m(\varepsilon; \mathbf{f}). \end{aligned} \quad (\text{A.5.23})$$

The second to last inequality is due to Neyman-Pearson inequality.

For the upper bound, we recollect our sufficient statistics (A.5.4) and associated notation, let

$$CI_{m,\alpha} = \begin{cases} \{M(\mathbf{g})\} & W < -z_\alpha + 0.5 \frac{\|\mathbf{f} - \mathbf{g}\|}{\varepsilon} \\ \{M(\mathbf{f})\} & W \geq (z_\alpha - \frac{\|\mathbf{f} - \mathbf{g}\|}{2\varepsilon}) \vee (-z_\alpha + \frac{\|\mathbf{f} - \mathbf{g}\|}{2\varepsilon}) \\ \{M(\mathbf{f}) + (M(\mathbf{g}) - M(\mathbf{f})) \cdot t : t \in [0, 1]\} & \text{otherwise} \end{cases}$$

Clearly, we have  $P_{\mathbf{f}}(M(\mathbf{f}) \notin CI_\alpha) \leq \alpha, P_{\mathbf{g}}(M(\mathbf{g}) \notin CI_\alpha) \leq \alpha$ . For the expected squared length, we have for  $\theta \in \{-1, 1\}$ ,

$$\mathbb{E}_{\mathbf{f}_\theta}(|CI_{m,\alpha}|) \leq \|M(\mathbf{f}) - M(\mathbf{g})\| \left( \Phi(z_\alpha - \frac{\|\mathbf{f} - \mathbf{g}\|}{\varepsilon}) - \alpha \right)_+ \quad (\text{A.5.24})$$

$$\begin{aligned} \mathbb{E}_{\mathbf{f}_\theta}(|CI_{m,\alpha}|) &\leq \max\{\tilde{\omega}_m(\varepsilon; \mathbf{f})(1 - 2\alpha), \sup_{c>1} \tilde{\omega}_m(c\varepsilon; \mathbf{f})(\Phi(z_\alpha - c) - \alpha)_+\} \\ &\leq \tilde{\omega}_m(\varepsilon; \mathbf{f}) \max\{(1 - 2\alpha), \sup_{c>1} c(\Phi(z_\alpha - c) - \alpha)_+\} \\ &\leq \tilde{\omega}_m(\varepsilon; \mathbf{f})(1 - 2\alpha) \times 2z_\alpha. \end{aligned} \quad (\text{A.5.25})$$

### A.5.3. Proof of Theorem 3.2.2

We start with stating two propositions, which are proved later.

**Proposition A.5.3.** *Let  $\rho_z(\varepsilon; f)$  be defined in (3.2.8) for  $f \in \mathcal{F}$ , and let  $\mathbf{f} \in \mathcal{F}_s$ . Then*

$$\sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i \varepsilon; f_i)^2 \leq \omega_z(\varepsilon; \mathbf{f}) \leq \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s 9\rho_z(b_i \varepsilon; f_i)^2, \quad (\text{A.5.26})$$

where  $b_i$  are non-negative.

**Proposition A.5.4.** *Suppose  $f_i \in \mathcal{F}$ , for  $i = 1, 2, \dots, s$ , then we have*

$$\frac{1}{3}s^{-\frac{2}{3}} \sum_{i=1}^s \rho_z(\varepsilon; f_i)^2 \leq \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i \varepsilon; f_i)^2 \leq \sum_{i=1}^s \rho_z(\varepsilon; f_i)^2. \quad (\text{A.5.27})$$

And for any  $\beta \leq s$ , exist  $(f_1, \dots, f_s)$  such that  $\sum_{i=1}^s \rho_z(\varepsilon; f_i)^2 = \beta$  and

$$\sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i \varepsilon; f_i)^2 = s^{-\frac{2}{3}} \sum_{i=1}^s \rho_z(\varepsilon; f_i)^2. \quad (\text{A.5.28})$$

For  $\beta \leq s$ , for any  $\delta > 0$ , there exist  $(f_1, \dots, f_s)$  such that  $\sum_{i=1}^s \rho_z(\varepsilon; f_i)^2 = \beta$  and

$$\sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i \varepsilon; f_i)^2 \geq \sum_{i=1}^s \rho_z(\varepsilon; f_i)^2 - \delta. \quad (\text{A.5.29})$$

Inequality (A.5.27) in Proposition A.5.4 and (A.5.26) in Proposition A.5.3 implies Inequality 3.2.9 of Theorem 3.2.2.

Construct  $\mathbf{f}(\mathbf{t}) = \sum_{i=1}^s \int_0^s f_i(x) dx + \sum_{i=1}^s (f(t_i) - \int_0^s f_i(x) dx)$  with  $f_i$  in Equation (A.5.28). Then together with the right hand side of Inequality (A.5.26) gives Inequality (3.2.10) of Theorem 3.2.2. Similar construct  $\mathbf{f}$  with  $f_i$  in Inequality (A.5.29) with  $\delta_0 = \delta$  gives Inequality (3.2.11) in Theorem 3.2.2.

### Proof of Proposition A.5.3

Suppose  $\mathbf{g} \in \mathcal{F}_s$ , such that  $\|\mathbf{g} - \mathbf{f}\| \leq \varepsilon$ ,  $\mathbf{g}(\mathbf{t}) = g_0 + g_1(t_1) + g_2(t_2) + \dots + g_s(t_s)$ . Using the continuity of  $\rho_z(\varepsilon; f)$  with respect to  $\varepsilon$  implied by Proposition 2.2.1 in Chapter 2, we know there exist  $(\bar{b}_1, \bar{b}_2, \dots, \bar{b}_s)$  such that

$$\bar{b}_i \geq 0, \text{ for } 1 \leq i \leq s, \sum_{i=1}^s \bar{b}_i^2 = 1, \sum_{i=1}^s \rho_z(\bar{b}_i \varepsilon; f_i)^2 = \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i \varepsilon; f_i)^2. \quad (\text{A.5.30})$$

We only need to prove

$$\sum_{i=1}^s \rho_z(\bar{b}_i \varepsilon; f_i)^2 \leq \omega_z(\varepsilon; \mathbf{f}) \leq \sum_{i=1}^s 9 \rho_z(\bar{b}_i \varepsilon; f_i)^2. \quad (\text{A.5.31})$$

We start with proving the upper bound.

Since  $\|\mathbf{g} - \mathbf{f}\| \leq \varepsilon$ , we have

$$\begin{aligned}\varepsilon^2 &\geq \|\mathbf{f} - \mathbf{g}\|^2 = \int_{\mathcal{D}} \left( f_0 - g_0 + \sum_{i=1}^2 f_i(t_i) - g_i(t_i) \right)^2 dt \\ &= (f_0 - g_0)^2 + \sum_{i=1}^s \int_0^1 (f_i(t) - g_i(t))^2 dt.\end{aligned}\tag{A.5.32}$$

Denote  $\tilde{b}_i = \sqrt{\frac{\int_0^1 (f_i(t) - g_i(t))^2 dt}{\varepsilon^2}}$  for  $1 \leq i \leq s$ , then we have  $\sum_{i=1}^s \tilde{b}_i^2 = 1$ .

Therefore, using Proposition 2.2.2 in Chapter 2, we have

$$\|Z(\mathbf{f}) - Z(\mathbf{g})\|^2 = \sum_{i=1}^s |Z(f_i) - Z(g_i)|^2 \leq \sum_{i=1}^s 9\rho_z(\tilde{b}_i\varepsilon; f_i)^2 \leq \sum_{i=1}^s 9\rho_z(\bar{b}_i\varepsilon; f_i)^2.\tag{A.5.33}$$

For the lower bound, we construct a class of function  $\mathbf{g}_\delta \in \mathcal{F}_s$ , with  $\frac{1}{2} \min_{1 \leq i \leq s} \rho_z(\bar{b}_i\varepsilon; f_i) > \delta > 0$ . We construct the constant and components:  $g_{\delta,i}$  for  $0 \leq s$ . Let  $g_{\delta,0} = f_0$ . For  $1 \leq i \leq s$ , suppose  $x_{l,i}, x_{r,i}$  are left and right end points of the interval  $\{x : f_i(x) \leq M(f_i) + \rho_m(\bar{b}_i\varepsilon; f_i)\}$ . And without loss of generality, we assume  $x_{r,i} = Z(f_i) + \rho_z(\bar{b}_i\varepsilon; f_i)$ . Define univariate convex function  $h_{\delta,i}$  as follow.

$$h_{\delta,i}(t) = \max\{f_i(t), f_i(x_{r,i} - \delta) - \frac{\rho_m(\bar{b}_i\varepsilon; f_i) + M(f_i) - f_i(x_{r,i} - \delta)}{x_{r,i} - \delta - x_{l,i}}(t - x_{r,i})\}.\tag{A.5.34}$$

Define univariate function  $g_{\delta,i}$  as

$$g_{\delta,i}(t) = h_{\delta,i}(t) - \int_0^1 h_{\delta,i}(t) dt.\tag{A.5.35}$$

Then we have  $\int_0^1 g_{\delta,i}(t) dt = 0$ , so the definition defines a valid  $\mathbf{g}_\delta \in \mathcal{F}_s$ .

Further for  $i = 1, 2, \dots, s$ , we have

$$\int_0^1 (g_{\delta,i}(t) - f_i(t))^2 dt = \int_0^1 (h_{\delta,i}(t) - f_i(t))^2 dt - \left( \int_0^1 h_{\delta,i}(t) dt \right)^2 \leq \bar{b}_i^2 \varepsilon^2, \quad (\text{A.5.36})$$

and

$$|Z(g_{\delta,i}) - Z(f_i)| \geq \rho_z(\bar{b}_i \varepsilon; f_i) - \delta. \quad (\text{A.5.37})$$

Therefore, we have

$$\|\mathbf{g}_\delta - \mathbf{f}\|^2 \leq \varepsilon^2, \|Z(\mathbf{g}_\delta) - Z(\mathbf{f})\|^2 \geq \sum_{i=1}^s (\rho_z(\bar{b}_i \varepsilon; f_i) - \delta)^2. \quad (\text{A.5.38})$$

Let  $\delta \rightarrow 0^+$ , we have

$$\omega_z(\varepsilon; \mathbf{f}) \geq \sum_{i=1}^s \rho_z(\bar{b}_i \varepsilon; f_i)^2. \quad (\text{A.5.39})$$

#### Proof of Proposition A.5.4

We start with the right hand side and its almost-attainability.

Since  $b_i \in [0, 1]$  for  $1 \leq i \leq s$ , we have  $\rho_z(b_i \varepsilon; f_i) \leq \rho_z(\varepsilon; f_i)$ . The right hand side then apparently hold.

We first assume  $\beta$  in not an integer. Let  $s_1 = \lfloor \beta - \delta \rfloor$ ,  $s_2 = \beta - \lfloor \beta \rfloor$ ,  $s_3 = s - \lceil \beta \rceil$ .

Let  $k_1, k_2, k_3 > 0$ .

Now we start defining  $f_i \in \mathcal{F}$  for  $1 \leq i \leq s$ .

If  $s_1 \geq 1$ , for  $1 \leq i \leq s_1$ , let

$$f_i(t) = k_1(t - \frac{1}{2}). \quad (\text{A.5.40})$$

If  $s_3 \geq 1$ , for  $n - s_3 + 1 \leq i \leq n$  let

$$f_i(t) = k_3(t - \frac{1}{2}). \quad (\text{A.5.41})$$

Let

$$f_{s_1+1}(t) = k_2(t - \frac{1}{2}). \quad (\text{A.5.42})$$

Suppose  $0 < \delta < \frac{1}{2}s_2$ .

If  $s_3 \geq 1$ , choose  $k_3$  such that

$$\rho_z(\varepsilon; f_n) = \sqrt{\frac{\delta}{2s_3}}, \quad (\text{A.5.43})$$

Define  $s_4 = s_2 - \frac{\delta}{2}$  if  $s_3 \geq 1$ , otherwise  $s_4 = s_2$ . Choose  $k_2$  such that

$$\rho_z(\varepsilon; f_{s_1+1}) = \sqrt{s_4}. \quad (\text{A.5.44})$$

Now suppose  $b_{s_1+1}$  is the smallest  $b \in [0, 1)$  such that

$$\rho_z(b\varepsilon; f_{s_1+1}) \geq \sqrt{s_4 - \frac{\delta}{2}}. \quad (\text{A.5.45})$$

If  $s_1 \geq 1$ , choose  $k_1$  such that

$$\rho_z(\sqrt{\frac{1-b^2}{s_1}}\varepsilon; f_1) = 1. \quad (\text{A.5.46})$$

It's easy to verify that the above construction is legitimate and satisfy equation (A.5.29).

When  $\beta = n$ , choose large enough  $k$  such that  $\rho_z(\frac{1}{\sqrt{s}}\varepsilon; k(t - 0.5)) = 1$ , and let  $f_i = k(t - 0.5)$  for  $1 \leq k \leq s$ .

When  $\beta \leq n - 1$  and is integer, for  $\delta < 0.5$ , let  $s_1 = \beta - 1$ ,  $s_3 = n - \beta$ ,  $s_4 = 1 - \frac{\delta}{2}$ . And



choose  $k_3, k_2, k_1$  as the case where  $\beta$  is not integer.

Now we proceed with the left hand side.

Recalling Proposition 2.2.1 in Chapter 2, we have

$$\sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i \varepsilon; f_i)^2 \geq \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s (b_i^2/4)^{\frac{2}{3}} \rho_z(\varepsilon; f_i)^2 \geq \frac{1}{3} \left( \sum_{i=1}^s \rho_z(\varepsilon; f_i)^6 \right)^{\frac{1}{3}}, \quad (\text{A.5.47})$$

The last inequality take  $b_i = \sqrt{\frac{\rho_z(\varepsilon; f_i)^6}{\sum_{i=1}^s \rho_z(\varepsilon; f_i)^6}}$ .

Cauchy-Schwarz inequality gives

$$\frac{1}{3} \left( \sum_{i=1}^s \rho_z(\varepsilon; f_i)^6 \right)^{\frac{1}{3}} \geq \frac{1}{3} s^{-\frac{2}{3}} \sum_{i=1}^s \rho_z(\varepsilon; f_i)^2, \quad (\text{A.5.48})$$

which concludes the left hand side.

For the attainability up to constant multiple, let  $k > 0$ , which we will pick later. Let  $f_i(t) = k(t - 0.5)$  for  $1 \leq i \leq s$ . Pick  $k > 0$  such that  $\rho_z(\varepsilon; f_i) = \sqrt{\frac{\beta}{s}}$ . Then we have that

$$\sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s \rho_z(b_i \varepsilon; f_i)^2 = \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s b_i^{\frac{4}{3}} \rho_z(\varepsilon; f_i)^2 = \sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s b_i^{\frac{4}{3}} \frac{\beta}{s}. \quad (\text{A.5.49})$$

Through basic calculation, we have  $\sup_{\sum_{i=1}^s b_i^2 \leq 1} \sum_{i=1}^s b_i^{\frac{4}{3}} = s^{\frac{1}{3}}$ , which gives inequality (A.5.28).

#### A.5.4. Proof of Theorem 3.2.3

We start with the upper bound. Suppose  $\|\mathbf{g} - \mathbf{f}\| \leq \varepsilon$ . Suppose  $\mathbf{g}(\mathbf{t}) = g_0 + \sum_{i=1}^s g_i(t_i)$ , where  $\int_0^1 g_i(t) dt = 0$ . Calculation show that  $\|\mathbf{g} - \mathbf{f}\| \leq \varepsilon$  implies

$$|g_0 - f_0|^2 + \sum_{i=1}^s \|g_i - f_i\|^2 \leq \varepsilon^2. \quad (\text{A.5.50})$$

Suppose  $\varepsilon_i = \|g_i - f_i\|$ . Then we have that

$$\begin{aligned}
|M(\mathbf{g}) - M(\mathbf{f})|^2 &\leq (|g_0 - f_0| + \sum_{i=1}^s |M(g_i) - M(f_i)|)^2 \leq (|g_0 - f_0| + \sum_{i=1}^s 3\rho_m(\varepsilon_i; f_i))^2 \\
&\leq (|g_0 - f_0| + \sum_{i=1}^s 3(\frac{\varepsilon_i}{\varepsilon})^{\frac{4}{3}} \rho_m(\varepsilon; f_i))^2 \\
&\leq \left( \varepsilon^2 + \sum_{i=1}^s \rho_m(\varepsilon; f_i)^2 \right) \left( \left( \frac{|g_0 - f_0|}{\varepsilon} \right)^2 + \sum_{i=1}^s 9(\frac{\varepsilon_i}{\varepsilon})^{\frac{8}{3}} \right) \\
&\leq \left( \sum_{i=1}^s \rho_m(\varepsilon; f_i)^2 \right) \frac{9(s+1)}{s},
\end{aligned} \tag{A.5.51}$$

where the second Inequality is due to Proposition 2.2.1.

Now that we have the upper bound, we turn to the lower bound. Let

$$\varepsilon_i = \frac{\rho_m(\varepsilon; f_i)}{\sqrt{\sum_{j=1}^s \rho_m(\varepsilon; f_j)^2}} \sqrt{\frac{1}{1 + \sum_{i=1}^s (1 \wedge 2\rho_z(\varepsilon; f_i))}} \varepsilon. \tag{A.5.52}$$

Suppose  $\delta > 0$  is small enough quantity, which will be set going to 0 later. We construct components of an alternative function. Without loss of generality we assume  $t_{i,l}, t_{i,r}$  are the left and right end points of the interval  $\{t : f_i(t) \leq M(f_i) + \rho_m(\varepsilon_i; f_i)\}$  and that  $t_{i,r} = Z(f_i) + \rho_z(\varepsilon_i; f_i)$ . Suppose  $g_{i,\delta}(t) = \max\{f_i(t), f_i(t_l) + \frac{-\delta}{t_{i,r}-t_{i,l}}(t - t_{i,l})\}$ , and let  $\mathbf{h}_\delta(\mathbf{t}) = f_0 + \sum_{i=1}^s g_i(t_i)$ . Then we have for small enough  $\delta > 0$ ,

$$\begin{aligned}
\|\mathbf{h}_\delta - \mathbf{f}\|^2 &\leq \left( \sum_{i=1}^s \int_0^1 g_i(t) dt \right)^2 + \sum_{i=1}^s \varepsilon_i^2 - \sum_{i=1}^s \left( \int_0^1 g_i(t) dt \right)^2 \\
&\leq \sum_{i=1}^s \varepsilon_i^2 (1 + \sum_{i=1}^s (1 \wedge 2\rho_z(\varepsilon_i; f_i))) \leq \varepsilon^2.
\end{aligned} \tag{A.5.53}$$

We also have

$$\begin{aligned}
\lim_{\delta \rightarrow 0^+} (M(\mathbf{h}_\delta) - M(\mathbf{f})) &\geq \sum_{i=1}^s \rho_m(\varepsilon_i; f_i) \geq \sum_{i=1}^s \rho_m(\varepsilon; f_i) \frac{\varepsilon_i}{\varepsilon} \\
&\geq \sqrt{\sum_{i=1}^s \rho_m(\varepsilon; f_i)^2} \sqrt{\frac{1}{1 + \sum_{i=1}^s (1 \wedge 2\rho_z(\varepsilon; f_i))}}.
\end{aligned} \tag{A.5.54}$$

This gives the lower bound.

#### A.5.5. Proof of Theorem 3.2.4

$$\begin{aligned}
&\inf_{CI_{z,\alpha} \in \mathcal{I}_{z,\alpha}(\mathcal{F}_s)} \mathbb{E}_{\mathbf{f}}(V(CI_{z,\alpha})) \\
&\geq \sup_{\mathbf{g} \in \mathcal{F}_s} \inf_{CI_{z,\alpha} \in \mathcal{I}_{z,\alpha}(\mathbf{f}, \mathbf{g})} \mathbb{E}_{\mathbf{f}}(V(CI_{z,\alpha})) \\
&\geq \sup_{\mathbf{g} \in \mathcal{F}_s} \inf_{CI_{z,\alpha} \in \mathcal{I}_{z,\alpha}(\mathbf{f}, \mathbf{g})} \mathbb{E}_{\mathbf{f}}(\mathbb{1}_{\{\{Z(\mathbf{f}), Z(\mathbf{g})\} \subset CI_{z,\alpha}\}}) \sup_{\mathbf{g} \in \mathcal{F}_s} \Pi_{i=1}^s |Z(g_i) - Z(f_i)| \\
&\geq \sup_{\mathbf{g} \in \mathcal{F}_s} \left( 1 - \alpha - \Phi(-z_\alpha + \frac{\|\mathbf{f} - \mathbf{g}\|}{\varepsilon}) \right) \sup_{\mathbf{g} \in \mathcal{F}_s} \Pi_{i=1}^s |Z(g_i) - Z(f_i)|
\end{aligned} \tag{A.5.55}$$

Let  $g_{i,\delta}$  be constructed as follows. Without loss of generality, we assume  $t_{i,r} = Z(f_i) + \rho_z(\varepsilon/\sqrt{s}; f_i)$  satisfies  $f_i(t_{i,r}) \leq \rho_m(\varepsilon/\sqrt{s}; f_i) + M(f_i)$  and  $t_{i,l}$  is the left end point of  $\{t : f_i(t) \leq \rho_m(\varepsilon/\sqrt{s}; f_i) + M(f_i)\}$ . Let

$$g_{i,\delta}(t) = \max\{f_i(t), M(f_i) + \rho_m(\varepsilon/\sqrt{s}; f_i) + \frac{-\delta}{t_{i,r} - t_{i,l}}(t - t_{i,l})\}. \tag{A.5.56}$$

Define

$$\mathbf{g}_\delta(\mathbf{t}) = f_0 + \sum_{i=1}^s g_{i,\delta}(t_i) - \sum_{i=1}^s \int_0^1 g_{i,\delta}(t) dt.$$

It's clear that

$$\|\mathbf{g}_\delta - \mathbf{f}\| \leq \varepsilon.$$

It is obvious that  $Z(g_{\delta,i}) = Z(g_{i,\delta})$ .

$$\lim_{\delta \rightarrow 0^+} \Pi_{i=1}^s |Z(g_{\delta,i}) - Z(f_i)| \geq \Pi_{i=1}^s \rho_z(\varepsilon/\sqrt{s}; f_i) \geq \left(\frac{1}{2\sqrt{s}}\right)^{\frac{2s}{3}} \Pi_{i=1}^s \rho_z(\varepsilon; f_i). \quad (\text{A.5.57})$$

Going back to Inequality (A.5.55) we have that

$$\inf_{CI_{z,\alpha} \in \mathcal{I}_{z,\alpha}(\mathcal{F}_s)} \mathbb{E}_{\mathbf{f}}(V(CI_{z,\alpha})) \geq (0.6 - \alpha) \left(\frac{1}{2\sqrt{s}}\right)^{\frac{2s}{3}} \Pi_{i=1}^s \rho_z(\varepsilon; f_i). \quad (\text{A.5.58})$$

### A.5.6. Proof of Theorem 3.2.5

We prove the theorem by proving the following two propositions.

**Proposition A.5.5.** *For any estimator of the minimizer,  $\hat{Z}$ , if*

$$\mathbb{E}_{\mathbf{f}} \left( \|\hat{Z} - Z(\mathbf{f})\|^2 \right) \leq \gamma R_z(\varepsilon; \mathbf{f})$$

for  $\mathbf{f} \in \mathcal{F}_s$  and  $\gamma < \gamma_0$ , where  $\gamma_0$  is a positive constant, then there exists  $\mathbf{f}_1 \in \mathcal{F}_s$  such that

$$\mathbb{E}_{\mathbf{f}_1} \left( \|\hat{Z} - Z(\mathbf{f}_1)\|^2 \right) \geq c_{z,s} \left( \log \frac{1}{\gamma} \right)^{\frac{2}{3}} R_z(\varepsilon; \mathbf{f}_1), \quad (\text{A.5.59})$$

where  $c_{z,s}$  is a constant depending on  $s$  only.

**Proposition A.5.6.** *For any estimator of the minimum,  $\hat{M}$ , if*

$$\mathbb{E}_{\mathbf{f}}(|\hat{M} - M(\mathbf{f})|^2) \leq \gamma R_m(\varepsilon; \mathbf{f})$$

for  $\mathbf{f} \in \mathcal{F}_s$  and  $\gamma < \gamma_0/s$ , where  $\gamma_0$  is a positive constant, then there exists  $\mathbf{f}_1 \in \mathcal{F}_s$  such that

$$\mathbb{E}_{\mathbf{f}_1} \left( |\hat{M} - M(\mathbf{f}_1)|^2 \right) \geq c_{m,s} \left( \log \frac{1}{\gamma} \right)^{\frac{2}{3}} R_m(\varepsilon; \mathbf{f}_1), \quad (\text{A.5.60})$$

where  $c_{m,s}$  is a constant depending on  $s$  only.

### Proof of Proposition A.5.5

Let  $\sigma = \frac{\Phi^{-1}(1-6.9 \cdot 2\gamma)\varepsilon}{\sqrt{5}}$ . Let  $F(\gamma) = (\sigma/\varepsilon)^2$ .

Then for  $\gamma \leq 0.0024558/54$ , we have  $\sigma \geq \sqrt{\frac{4}{3}}\varepsilon$ .

Suppose  $(w_1, w_2, \dots, w_s)$  achieves

$$\sup_{\sum_{j=1}^s w_j^2 \leq 1, w_j \geq 0} \sum_i^s \rho_z(w_i \varepsilon; f_i)^2. \quad (\text{A.5.61})$$

The compactness of  $\{(w_1, w_2, \dots, w_s) : \sum_{j=1}^s w_j^2 \leq 1, w_j \geq 0\}$  and the continuity of  $\sum_i^s \rho_z(w_i \varepsilon; f_i)^2$  implies that supremum is attainable. So  $(w_1, w_2, \dots, w_s)$  is well defined.

Also, it's easy to see that  $\sum_{j=1}^s w_j^2 = 1$ .

Denote set  $B$  as

$$B = \{(b_1, b_2, \dots, b_s) : \sum_{i=1}^s b_i \leq 1, b_i \geq \max\{\frac{w_i}{\sqrt{F(\gamma)}}, \sqrt{1/4s}\}\}. \quad (\text{A.5.62})$$

It's clear that  $B$  is not null set, and

$$(\sqrt{\frac{w_1^2}{F(\gamma)} + \frac{1}{4s}}, \sqrt{\frac{w_2^2}{F(\gamma)} + \frac{1}{4s}}, \dots, \sqrt{\frac{w_s^2}{F(\gamma)} + \frac{1}{4s}}) \in B. \quad (\text{A.5.63})$$

Let  $(b_1, b_2, \dots, b_s)$  achieves

$$\sup_{(b_1, b_2, \dots, b_s) \in B} \left( \sum_{k=1}^s \rho_z(b_k \sqrt{F(\gamma)} \varepsilon; f_k)^2 \right)^3 / \left( \sum_{i=1}^s \frac{\rho_z(b_i \sqrt{F(\gamma)} \varepsilon; f_i)^4}{\rho_m(b_i \sqrt{F(\gamma)} \varepsilon; f_i)^4} \right). \quad (\text{A.5.64})$$

Then it is clear that

$$\begin{aligned}
& \left( \sum_{k=1}^s \rho_z(b_k \sqrt{F(\gamma)} \varepsilon; f_k)^2 \right)^3 / \left( \sum_{i=1}^s \frac{\rho_z(b_i \sqrt{F(\gamma)} \varepsilon; f_i)^4}{\rho_m(b_i \sqrt{F(\gamma)} \varepsilon; f_i)^4} \right) \\
& \geq \min_{1 \leq k \leq s} \left( \rho_z(b_k \sqrt{F(\gamma)} \varepsilon; f_k)^2 \right)^3 / \left( \frac{\rho_z(b_k \sqrt{F(\gamma)} \varepsilon; f_k)^4}{\rho_m(b_k \sqrt{F(\gamma)} \varepsilon; f_k)^4} \right) \\
& \geq \min_{1 \leq k \leq s} \left( \frac{1}{2} b_k^2 F(\gamma) \varepsilon^2 \right)^2 \geq \left( \frac{F(\gamma)}{8s} \varepsilon^2 \right)^2,
\end{aligned} \tag{A.5.65}$$

and that

$$\sum_{k=1}^s \rho_z(b_k \sqrt{F(\gamma)} \varepsilon; f_k)^2 \geq \sum_{k=1}^s \rho_z(w_k \varepsilon; f_k)^2 \geq \frac{1}{9} \omega_z(\varepsilon; \mathbf{f}), \tag{A.5.66}$$

where the very last inequality comes from Proposition A.5.3.

For each  $1 \leq k \leq s$ , we construct  $\tilde{f}_k$ .

Let  $x_l, x_r$  be the left and right end points of the interval  $\{x : f_k(x) \leq M(f_k) + \rho_m(b_k \sigma; f_k)\}$ .

Without loss of generality, suppose  $f_k(Z(f_k) + \rho_z(b_k \sigma; f_k)) \leq M(f_k) + \rho_m(b_k \sigma; f_k)$ .

Let  $g_{2,k}(t) = \max\{f_k(t), f_k(x_r) + \frac{M(f_k) + 2\rho_m(b_k \sigma; f_k) - f_k(x_r)}{x_l - x_r}(t - x_r)\}$ .

Calculation similar to that in Lemma A.1.5 shows that

$$\begin{aligned}
\|g_{2,k} - f_k\| & \leq \sqrt{5} b_k \sqrt{F(\gamma)} \varepsilon \\
\rho_z(\eta; g_{2,k}) & \leq \left(\frac{16}{3}\right)^{\frac{1}{3}} \left(\frac{\eta}{\sqrt{b_k^2 \sigma^2 / 3}}\right)^{\frac{2}{3}} \rho_z(b_k \sigma; f_k).
\end{aligned} \tag{A.5.67}$$

Let

$$\mathbf{g}(\mathbf{t}) = f_0 + \sum_{k=1}^s \left( g_{2,k}(t_k) - \int_0^1 g_{2,k}(t) dt \right). \tag{A.5.68}$$

Then we know that

$$\|\mathbf{g} - \mathbf{f}\| \leq \Phi^{-1}(1 - 6 \cdot 9 \cdot 2\gamma) \varepsilon, \tag{A.5.69}$$

that

$$\|Z(\mathbf{g}) - Z(\mathbf{f})\|^2 = \sum_{k=1}^s \rho_z(b_k \sigma; f_k)^2 \geq \frac{1}{9} \omega_z(\varepsilon; \mathbf{f}), \quad (\text{A.5.70})$$

and that

$$\begin{aligned} \omega_z(\varepsilon; \mathbf{g}) &\leq 9 \sup_{\sum_{j=1}^s d_j^2 \leq 1, d_j \geq 0} \sum_{k=1}^s \rho_z(d_j \varepsilon; g_{2,k})^2 \\ &\leq 9 \sup_{\sum_{j=1}^s d_j^2 \leq 1, d_j \geq 0} \sum_{k=1}^s \left(\frac{16}{3}\right)^{\frac{2}{3}} \left(\frac{d_k \varepsilon}{\sqrt{b_k^2 \sigma^2 / 3}}\right)^{\frac{4}{3}} \rho_z(b_k \sigma; f_k)^2. \end{aligned} \quad (\text{A.5.71})$$

Taking derivative of

$$\sum_{k=1}^s \left(\frac{d_k}{\sqrt{b_k^2}}\right)^{\frac{4}{3}} \rho_z(b_k \sigma; f_k)^2 \quad (\text{A.5.72})$$

with respect to

$$(d_1^2, d_2^2, \dots, d_s^2), \quad (\text{A.5.73})$$

we have

$$\left( \frac{2}{3} (d_1^2)^{-\frac{1}{3}} b_1^{-\frac{4}{3}} \rho_z(b_1 \sigma; f_1)^2, \dots, \frac{2}{3} (d_s^2)^{-\frac{1}{3}} b_s^{-\frac{4}{3}} \rho_z(b_s \sigma; f_s)^2 \right). \quad (\text{A.5.74})$$

Note that the constraint for  $d_1^2, d_2^2, \dots, d_s^2$  is

$$\sum_{k=1}^s d_k^2 = 1, d_j^2 \geq 0 \text{ for } 1 \leq j \leq s. \quad (\text{A.5.75})$$

Therefore, we have that

$$\begin{aligned}
\sum_{k=1}^s \left( \frac{d_k}{\sqrt{b_k^2}} \right)^{\frac{4}{3}} \rho_z(b_k \sigma; f_k)^2 &\leq \sum_{k=1}^s \left( \frac{\rho_z(b_k \sigma; f_k)^6 / b_k^4 \sum_{j=1}^s \left( \rho_z(b_j \sigma; f_j)^6 / b_j^4 \right)}{b_k^2} \right)^{\frac{2}{3}} \rho_z(b_k \sigma; f_k)^2 \\
&\leq \left( \sum_{j=1}^s \rho_z(b_j \sigma; f_j)^6 / b_j^4 \right)^{\frac{1}{3}} \\
&\leq \left( \sum_{j=1}^s \sigma^4 4 \cdot \frac{\rho_z(b_j \sigma; f_j)^4}{\rho_m(b_j \sigma; f_j)^4} \right)^{\frac{1}{3}}.
\end{aligned} \tag{A.5.76}$$

Using Inequality (A.5.65) and going back to Inequality (A.5.71), we have that

$$\begin{aligned}
\omega_z(\varepsilon; \mathbf{g}) &\leq 9 \cdot (16 \cdot 3)^{\frac{2}{3}} \left( \frac{\varepsilon}{\sigma} \right)^{\frac{4}{3}} \cdot \sigma^{\frac{4}{3}} \cdot 4^{\frac{1}{3}} \left( \frac{8s}{F(\gamma)\varepsilon^2} \right)^{\frac{2}{3}} \sum_{k=1}^s \rho_z(b_k \sigma; f_k)^2 \\
&= 9 \cdot (16 \cdot 3)^{\frac{2}{3}} \cdot 2^{\frac{8}{3}} \left( \frac{s}{F(\gamma)} \right)^{\frac{2}{3}} \|Z(\mathbf{f}) - Z(\mathbf{g})\|^2.
\end{aligned} \tag{A.5.77}$$

Recall that when we let  $\mathbf{f}_\theta = \mathbf{f}$  for  $\theta = 1$  and  $\mathbf{f}_\theta = \mathbf{g}$  for  $\theta = -1$ , a sufficient statistic would be  $W$  defined in (A.5.4).

Note that we have

$$\mathbb{E}_{\mathbf{f}} \left( \|\hat{Z} - Z(\mathbf{f})\|^2 \right) \leq \gamma R_z(\varepsilon; \mathbf{f}) \leq 6\gamma \omega_z(\varepsilon; \mathbf{f}), \tag{A.5.78}$$

where the last Inequality comes from Theorem 3.2.1.

Denote event  $D = \{\|\hat{Z} - Z(\mathbf{f})\| \geq \frac{1}{18} \omega_z(\varepsilon; \mathbf{f})\}$ . Then

$$P_{\mathbf{f}}(D) \leq \frac{6\gamma \omega_z(\varepsilon; \mathbf{f})}{\frac{1}{18} \omega_z(\varepsilon; \mathbf{f})} = 108\gamma \leq 0.00491163. \tag{A.5.79}$$

So we have that

$$P_{\mathbf{g}}(D) \leq \frac{1}{2}. \tag{A.5.80}$$



Hence we have that

$$\begin{aligned}
\mathbb{E}_{\mathbf{g}} \left( \|\hat{Z} - Z(\mathbf{g})\|^2 \right) &\geq \mathbb{E}_{\mathbf{g}} \left( \left( \|Z(\mathbf{f}) - Z(\mathbf{g})\| - \frac{1}{18} \omega_z(\varepsilon; \mathbf{f}) \right)_+^2 \mathbb{1}\{D^c\} \right) \\
&\geq \mathbb{E}_{\mathbf{g}} \left( \frac{1}{4} \|Z(\mathbf{f}) - Z(\mathbf{g})\|^2 \mathbb{1}\{D^c\} \right) \geq \frac{1}{8} \|Z(\mathbf{f}) - Z(\mathbf{g})\|^2 \\
&\geq \frac{1}{8} \frac{1}{9} (16 \cdot 3)^{-\frac{2}{3}} \cdot 2^{-\frac{8}{3}} \frac{F(\gamma)^{\frac{2}{3}}}{s^{\frac{2}{3}}} \omega_z(\varepsilon; \mathbf{g}) \\
&\geq \frac{1}{8} \frac{1}{9} (16 \cdot 3)^{-\frac{2}{3}} \cdot 2^{-\frac{8}{3}} \frac{1}{6} R_z(\varepsilon; \mathbf{g}) \frac{F(\gamma)^{\frac{2}{3}}}{s^{\frac{2}{3}}}.
\end{aligned} \tag{A.5.81}$$

Note that  $F(\gamma) = z_{108\gamma}^2/5$ , so  $F(\gamma) \sim \log(\frac{1}{\gamma})$ , so we have

$$\mathbb{E}_{\mathbf{g}} \left( \|\hat{Z} - Z(\mathbf{g})\|^2 \right) \geq c_z \cdot s^{-\frac{2}{3}} \log\left(\frac{1}{\gamma}\right)^{\frac{2}{3}} R_z(\varepsilon; \mathbf{g}). \tag{A.5.82}$$

for some constant  $c_z > 0$ .

Letting  $c_{z,s} = c_z \cdot s^{-\frac{2}{3}}$  and  $\mathbf{f}_1 = \mathbf{g}$  gives the statement of the Proposition.

### Proof of Proposition A.5.6

Take  $\sigma = \Phi^{-1}(1 - 108(s+1)^2\gamma/s)\varepsilon$ .

Suppose  $\gamma \leq 0.158655s/108(s+1)^2$ . Then we know that  $\sigma > 1$

Take the construction of  $\mathbf{h}_\delta$  in the Proof of Theorem 3.2.3 with the noise level being  $\sigma$ . Then we know that

$$\begin{aligned}
\|\mathbf{h}_\delta - \mathbf{f}\| &\leq \sigma, \\
\lim_{\delta \rightarrow 0^+} \|M(\mathbf{f}) - M(\mathbf{h}_\delta)\|^2 &\geq \frac{\sum_{k=1}^s \rho_m(\sigma; f_k)^2}{1+s} \geq \left(\frac{\sigma}{\varepsilon}\right)^{\frac{4}{3}} \frac{\sum_{k=1}^s \rho_m(\varepsilon; f_k)^2}{1+s} \\
&\geq \Phi^{-1}(1 - 2(s+1)\gamma)^{\frac{4}{3}} \frac{\sum_{k=1}^s \rho_m(\varepsilon; h_{\delta,k})^2}{1+s} \\
&\geq \Phi^{-1}(1 - 2(s+1)\gamma)^{\frac{4}{3}} \frac{s}{9(s+1)^2} \omega_m(\varepsilon; \mathbf{h}_\delta) \\
&\geq \Phi^{-1}(1 - 2(s+1)\gamma)^{\frac{4}{3}} \frac{s}{9(s+1)^2} \frac{1}{6} R_m(\varepsilon; \mathbf{h}_\delta).
\end{aligned} \tag{A.5.83}$$

Note that  $\frac{\sigma}{\varepsilon} > 1$ . Hence, there exists  $\delta_0 > 0$ , such that for  $\delta_0 > \delta > 0$ , we have

$$\|M(\mathbf{f}) - M(\mathbf{h}_\delta)\|^2 \geq \frac{s}{9(s+1)^2} \omega_m(\varepsilon; \mathbf{f}) \geq \frac{s}{54(s+1)^2} R_m(\varepsilon; \mathbf{f}). \quad (\text{A.5.84})$$

Denote event

$$D = \{\|\hat{M} - M(\mathbf{f})\|^2 \geq \frac{s}{108(s+1)^2} R_m(\varepsilon; \mathbf{f})\}. \quad (\text{A.5.85})$$

Then we know that

$$P_{\mathbf{f}}(D) \leq \gamma \cdot \frac{108(s+1)^2}{s}. \quad (\text{A.5.86})$$

So

$$P_{\mathbf{h}_\delta}(D) \leq \frac{1}{2}. \quad (\text{A.5.87})$$

Therefore, we have that

$$\begin{aligned} \mathbb{E}_{\mathbf{h}_\delta} \left( \|\hat{M} - M(\mathbf{h}_\delta)\|^2 \right) &\geq \mathbb{E}_{\mathbf{h}_\delta} \left( \left(1 - \frac{1}{\sqrt{2}}\right)^2 \|M(\mathbf{f}) - M(\mathbf{h}_\delta)\|^2 \mathbb{1}\{D^c\} \right) \\ &\geq \frac{3 - 2\sqrt{2}}{4} \|M(\mathbf{f}) - M(\mathbf{h}_\delta)\|^2. \end{aligned} \quad (\text{A.5.88})$$

From Inequality (A.5.83), we know that there exists  $0 < \delta_1 < \delta_0$ , such that for  $\delta < \delta_1$ , we have

$$\|M(\mathbf{f}) - M(\mathbf{h}_\delta)\|^2 \geq \Phi^{-1}(1 - 2(s+1)\gamma)^{\frac{4}{3}} \frac{s}{55(s+1)^2} R_m(\varepsilon; \mathbf{h}_\delta). \quad (\text{A.5.89})$$

Hence,

$$\mathbb{E}_{\mathbf{h}_\delta} \left( \|\hat{M} - M(\mathbf{h}_\delta)\|^2 \right) \geq \frac{3 - 2\sqrt{2}}{4} \Phi^{-1}(1 - 2(s+1)\gamma)^{\frac{4}{3}} \frac{s}{55(s+1)^2} R_m(\varepsilon; \mathbf{h}_\delta). \quad (\text{A.5.90})$$

Note that  $\Phi^{-1}(1 - 2(s+1)\gamma)^{\frac{4}{3}} \sim \log(\frac{1}{s\gamma})^{\frac{2}{3}}$  as  $\gamma \rightarrow 0^+$  and that  $\log(\frac{1}{s\gamma})^{\frac{2}{3}} \geq (\log(\frac{1}{\gamma})/\log(s))^{\frac{2}{3}}$  for  $\gamma < \frac{1}{3s}$ , so we have the statement by taking  $\mathbf{f}_1 = \mathbf{h}_\delta$ .

### A.5.7. Proof of Proposition 3.3.1

We start with the first item.

Suppose  $\mathfrak{P}(Y^1) = \mathfrak{P}(Y^2)$  for  $Y^1, Y^2 \in \mathfrak{Y}$ . Then for  $\mathcal{A} = [a_1, A_1] \times [a_2, A_2] \times \cdots \times [a_s, A_s] \subset [0, 1]^s$ , we have

$$\begin{aligned} \int_{\mathcal{A}} dY^1 &= \int_{\mathcal{A}} d\mathbf{er}(Y^1) + \sum_{i=1}^s \Pi_{j \neq i}(A_j - a_j) \int_{a_i}^{A_i} d\boldsymbol{\pi}_i(Y^1) \\ &= \int_{\mathcal{A}} d\mathbf{er}(Y^2) + \sum_{i=1}^s \Pi_{j \neq i}(A_j - a_j) \int_{a_i}^{A_i} d\boldsymbol{\pi}_i(Y^2) \\ &= \int_{\mathcal{A}} dY^2. \end{aligned} \tag{A.5.91}$$

Therefore, using Dynkin's  $\pi - \lambda$  theorem,  $Y^1 = Y^2$ .

Now we continue with the second item.

Again, from Dynkin's  $\pi - \lambda$  theorem, we only need to prove that for any

$$[a_1, A_1], [a_2, A_2], \dots, [a_s, A_s] \subset [0, 1] \text{ and } \mathfrak{B} = [b_1, B_1] \times [b_2, B_2] \times \cdots \times [b_s, B_s],$$

the following variables are independent:

$$\int_{[a_1, A_1]} d\boldsymbol{\pi}_1(Y), \int_{[a_2, A_2]} d\boldsymbol{\pi}_2(Y), \dots, \int_{[a_s, A_s]} d\boldsymbol{\pi}_s(Y), \int_{[b_1, B_1] \times [b_2, B_2] \times \cdots \times [b_s, B_s]} d\mathbf{er}(Y).$$

Note that  $\boldsymbol{\pi}_i(Y)[A_i] - \boldsymbol{\pi}_i(Y)[a_i] = \int_{[a_i, A_i]} d\boldsymbol{\pi}_i(Y)$ , but we use integral form whenever possible to ease understanding as we have stochastic processes of different dimensions.

From the definition 3.3.1 of  $\boldsymbol{\pi}_i(Y)$  and  $\mathbf{er}(Y)$ , we know that

$$\left( \int_{[a_1, A_1]} d\boldsymbol{\pi}_1(Y), \int_{[a_2, A_2]} d\boldsymbol{\pi}_2(Y), \dots, \int_{[a_s, A_s]} d\boldsymbol{\pi}_s(Y), \int_{\mathfrak{B}} d\mathbf{er}(Y) \right)$$

is joint normal random vector. To prove independence we only need to prove the correlations are zero.

For  $1 \leq i < j \leq s$ , we have

$$\begin{aligned}
& COV\left(\int_{[a_i, A_i]} d\boldsymbol{\pi}_i(Y), \int_{[a_j, A_j]} d\boldsymbol{\pi}_j(Y)\right) \\
&= \mathbb{E}\left(\left(\int_{t_i \in [a_i, A_i], \mathbf{t}_{-i} \in [0, 1]^{s-1}} dW - (A_i - a_i) \int_{[0, 1]^s} dW\right) \cdot \right. \\
&\quad \left. \left(\int_{t_j \in [a_j, A_j], \mathbf{t}_{-j} \in [0, 1]^{s-1}} dW - (A_j - a_j) \int_{[0, 1]^s} dW\right)\right) \\
&= 0.
\end{aligned} \tag{A.5.92}$$

For  $1 \leq i \leq s$ , suppose  $\mathcal{A}_i = \{\mathbf{t} : t_i \in [a_i, A_i], \mathbf{t}_{-i} \in [0, 1]^{s-1}\}$ , and  $V(\cdot)$  denotes the volume (length when one dimensional, area when two dimensional, etc.), we have

$$\begin{aligned}
& COV\left(\int_{[a_i, A_i]} d\boldsymbol{\pi}_i(Y), \int_{\mathfrak{B}} dY\right) \\
&= \mathbb{E}\left(\left(\int_{t_i \in [a_i, A_i], \mathbf{t}_{-i} \in [0, 1]^{s-1}} dW - (A_i - a_i) \int_{[0, 1]^s} dW\right) \cdot \right. \\
&\quad \left. \left(\int_{\mathfrak{B}} dW - \sum_{j=1}^s \Pi_{k \neq j}(B_k - b_k) \int_{t_j \in [b_j, B_j], \mathbf{t}_{-j} \in [0, 1]^{s-1}} dW + s \Pi_{k=1}^s(B_k - b_k) \int_{[0, 1]^s} dW\right)\right) \\
&= V(\mathcal{A}_i \cap \mathfrak{B}) - (A_i - a_i)V(\mathfrak{B}) - \sum_{j \neq i} \Pi_{k \neq j}(B_k - b_k)(B_j - b_j)(A_i - a_i) \\
&\quad - V([a_i, A_i] \cap [b_i, B_i])\Pi_{j \neq i}(B_j - b_j) + s(A_i - a_i)\Pi_{i=1}^s(B_i - b_i) + 0 \\
&= 0.
\end{aligned} \tag{A.5.93}$$

Therefore, we prove the independence.

Now we continue with the sufficiency property. Recalling the Radon-Nikodym derivative

calculated in (A.5.3), we have that for  $\mathbf{f}, \mathbf{g} \in \mathcal{F}_s$

$$\begin{aligned} \frac{dP_{\mathbf{f}}}{dP_{\mathbf{g}}}(Y) &= \exp \left( \int_{[0,1]^s} \frac{\mathbf{f}(\mathbf{t}) - \mathbf{g}(\mathbf{t})}{\varepsilon^2} dY(\mathbf{t}) - \frac{1}{2} \int_{[0,1]^s} \frac{\mathbf{f}(\mathbf{t})^2 - \mathbf{g}(\mathbf{t})^2}{\varepsilon^2} d\mathbf{t} \right) \\ &= \exp \left( \frac{1}{\varepsilon^2} \sum_{i=1}^s \int_0^1 (f_i(t) - g_i(t)) d\boldsymbol{\pi}_i(Y) - \frac{1}{2\varepsilon^2} \int_{[0,1]^s} (\mathbf{f}(\mathbf{t})^2 - \mathbf{g}(\mathbf{t})^2) d\mathbf{t} \right). \end{aligned} \quad (\text{A.5.94})$$

Hence we concludes the proof.

### A.5.8. Proof of Theorem 3.3.1

Recalling Theorem 3.2.1 and Theorem 3.2.2, we know that it suffices to prove that

$$\mathbb{E}_{\mathbf{f}} \left( \|\hat{Z} - Z(\mathbf{f})\|^2 \right) \leq C_2 \sum_{k=1}^s \rho_z(\varepsilon; \mathbf{f})^2, \quad (\text{A.5.95})$$

for an absolute constant  $C_2 > 0$ .

Since we have

$$\mathbb{E}_{\mathbf{f}} \left( \|\hat{Z} - Z(\mathbf{f})\|^2 \right) = \sum_{k=1}^s \mathbb{E}_{\mathbf{f}} \left( |\hat{Z}_k - Z(f_k)|^2 \right), \quad (\text{A.5.96})$$

we only need to prove that there is an absolute constant  $C_2 > 0$  such that for  $1 \leq k \leq s$ ,

$$\mathbb{E}_{\mathbf{f}} \left( |\hat{Z}_k - Z(f_k)|^2 \right) \leq C_2 \rho_z(\varepsilon; f_k)^2. \quad (\text{A.5.97})$$

Now we focus on any given  $k \in \{1, \dots, s\}$ .

Note that for each level  $j \geq 1$ , the localization and stopping rule only based on the following random variables  $\{\tilde{X}_{j,i,k} - \tilde{X}_{j,i-1,k} : i = 2, \dots, 2^j\} \cup \{X_{j,i,k} - X_{j,i-1,k} : i = 2, \dots, 2^j\}$ .

If we construct two stochastic process  $\tilde{\mathbf{v}}^l$  and  $\tilde{\mathbf{v}}^r$  in the following way

$$\begin{aligned} d\tilde{\mathbf{v}}^l(t) &= f_k(t)dt + \sqrt{3}\varepsilon dW^l, \\ d\tilde{\mathbf{v}}^r(t) &= f_k(t)dt + \sqrt{3}\varepsilon dW^r, \end{aligned} \quad (\text{A.5.98})$$

where  $W^l$  and  $W^r$  are independent Brownian Motion, and also define  $O_{j,i,k}, \tilde{O}_{j,i,k}$  in the same way as  $X_{j,i,k}, \tilde{X}_{j,i,k}$  with  $\mathbf{v}^l$  and  $\mathbf{v}^r$  replaced by  $\tilde{\mathbf{v}}^l$  and  $\tilde{\mathbf{v}}^r$ , then we know that the distribution under  $\mathbf{f}$  of the infinite dimension object  $Ds(X, k)$  that concatenate the following vectors with  $j = 1, 2, \dots$ :

$$\begin{aligned} &(\tilde{X}_{j,2,k} - \tilde{X}_{j,1,k}, \tilde{X}_{j,3,k} - \tilde{X}_{j,2,k}, \dots, \tilde{X}_{j,2^j,k} - \tilde{X}_{j,2^{j-1},k}, \\ &X_{j,2,k} - X_{j,1,k}, X_{j,3,k} - X_{j,2,k}, \dots, X_{j,2^j,k} - X_{j,2^{j-1},k}) \end{aligned} \quad (\text{A.5.99})$$

is the same with that having  $O_{j,i,k}, \tilde{O}_{j,i,k}$  in the place of  $X_{j,i,k}, \tilde{X}_{j,i,k}$ , which we call  $Ds(O, k)$ .

Also note that the localization procedure, stopping procedure and construction of each axis of the estimator goes in parallel with the univariate estimator in Chapter 2, and that the distribution of random variables playing a role in the entire estimation procedure (i.e.  $Ds(X, k)$ ) is the same with that of  $Ds(O, k)$ .

Hence bounding  $E_{\mathbf{f}}(|\hat{Z}_k - Z(f_k)|^2)$  here is the same with bounding  $\mathbb{E}_{f_k}(|\tilde{Z} - Z(f_k)|^2)$  with  $\tilde{Z}$  being the estimator of the minimizer of the univariate function in the setting of univariate case in Chapter 2.

Resort to the proof of that of Theorem 2.3.1 in Chapter 2 with the quantities bounding  $|\tilde{Z} - Z(f_k)|$  there being replaced by the square of it, we have

$$\mathbb{E}_{\mathbf{f}}(|\hat{Z}_k - Z(f_k)|^2) \leq \mathbb{E}_{f_k}(|\tilde{Z} - Z(f_k)|^2) \leq C_2 \rho_z(\varepsilon; f_k)^2, \quad (\text{A.5.100})$$

for an absolute constant  $C_2$ .

### A.5.9. Proof of Theorem 3.3.2

Recalling the lower bound of  $L_{\alpha,z}(\varepsilon; \mathbf{f})$  established in Theorem 3.2.4 and Proposition 2.2.1 in Chapter 2, it suffices to prove the following two two propositions.

**Proposition A.5.7** (Coverage). *The confidence hyper cube  $CI_{z,\alpha}$  defined by (3.3.12) is an*

$1 - \alpha$  level confidence cube for minimizer.

**Proposition A.5.8** (Expected Volume). *For  $\alpha \leq 0.3$ , and confidence hyper cube  $CI_{z,\alpha}$  defined by (3.3.12), we have*

$$\mathbb{E}_{\mathbf{f}}(V(CI_{z,\alpha})) \leq C_3^{\frac{s}{2}} \sum_{k=1}^s \rho_z(z_{\alpha/s}\varepsilon; f_k), \quad (\text{A.5.101})$$

where  $C_3$  is an absolute positive constant.

Note that  $\rho_z(z_{\alpha/s}\varepsilon; f_k) \leq (2z_{\alpha/s})^{\frac{2}{3}} \rho_z(\varepsilon; f_k)$ , so these two propositions lead to the theorem.

### Proof of Proposition A.5.7

By the definition of confidence hyper cube  $CI_{z,\alpha}$  in (3.3.12), its  $k$ -th coordinate  $CI_k$  only depend  $Y$  through  $\boldsymbol{\pi}_k(Y)$ . So it has mutually independent coordinates. Hence we have

$$P_{\mathbf{f}}(Z(\mathbf{f}) \in CI_{z,\alpha}) = \prod_{k=1}^s P_{\mathbf{f}}(Z(f_k) \in CI_k) \geq \prod_{k=1}^s \inf_{\mathbf{f} \in \mathcal{F}_s} P_{\mathbf{f}}(Z(f_k) \in CI_k). \quad (\text{A.5.102})$$

So it suffices to prove that  $\inf_{\mathbf{f} \in \mathcal{F}_s} P_{\mathbf{f}}(Z(f_k) \in CI_k) \geq 1 - \frac{\alpha}{s}$ .

Denote  $\dot{j}_k = \min\{j : |\hat{i}_{j,k} - i_{j,k}^*| \geq 7\}$ . Then we have for any  $\mathbf{f} \in \mathcal{F}_s$ ,

$$\begin{aligned} P_{\mathbf{f}}(Z(f_k) \notin CI_k) &= P_{\mathbf{f}}(\dot{j}_k < \hat{j}(\alpha/s, k)) = \sum_{j=3}^{\infty} \mathbb{E}_{\mathbf{f}}(\mathbb{E}_{\mathbf{f}}(\mathbb{1}\{j < \hat{j}(\alpha/s, k)\} | \mathbf{v}_k^l) \mathbb{1}\{\dot{j}_k = j\}) \\ &\leq \sum_{j=3}^{\infty} \mathbb{E}_{\mathbf{f}}(\alpha/s \mathbb{1}\{\dot{j}_k = j\}) \leq \alpha/s. \end{aligned} \quad (\text{A.5.103})$$

The first inequality is due to the distribution in (3.3.7) and that for the  $\frac{\tilde{X}_{j, \hat{i}_{j,k}-6,k} - \tilde{X}_{j, \hat{i}_{j,k}-5,k}}{\sigma_j}$ , as well as the facts that  $\hat{i}_{j,k}$  only depends on  $\mathbf{v}_k^l$ , that  $\mathbf{v}_k^l$  and  $\mathbf{v}_k^r$  are independent, and that  $j = \dot{j}_k$  implies  $S_p(j, k) \leq 0$  or that for the left side is non-positive.

This concludes the proof.

### Proof of Proposition A.5.8

Note that the coordinates of the confidence hyper cube are independent, so we have

$$\mathbb{E}_{\mathbf{f}}(V(CI_{z,\alpha})) = \Pi_{k=1}^s \mathbb{E}_{\mathbf{f}}(\|CI_k\|), \quad (\text{A.5.104})$$

it suffice to prove that there exists an absolute constant  $C_3 > 0$  such that for any  $k \in \{1, 2, \dots, s\}$ , the following holds

$$\mathbb{E}_{\mathbf{f}}(\|CI_k\|^2) \leq C_3 \rho_z(z_{\alpha/s} \varepsilon; f_k)^2. \quad (\text{A.5.105})$$

Now we recollect and introduce some notation that indicate the levels at which the localization procedure picks a interval far away from the right one.

$$\begin{aligned} \tilde{j}_k &= \min\{j : |\hat{i}_{j,k} - i_{j,k}^*| \geq 2\}, \\ \acute{j}_k &= \min\{j : |\hat{i}_{j,k} - i_{j,k}^*| \geq 5\}, \\ \grave{j}_k &= \min\{j : |\hat{i}_{j,k} - i_{j,k}^*| \geq 7\}. \end{aligned} \quad (\text{A.5.106})$$

It's clear that for any  $j \geq \tilde{j}_k$  we have

$$|\hat{i}_{j,k} - i_{j,k}^*| \geq 2. \quad (\text{A.5.107})$$

We also introduce a quantity as follow.

$$j_k^* = \min\{j : m_j \leq \frac{\rho_z(\varepsilon; f_k)}{4}\}. \quad (\text{A.5.108})$$



We have

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}}(\|CI_k\|^2) \\
& \leq 169 \sum_{j=3}^{\infty} \mathbb{E}_{\mathbf{f}}(2^{-2j} \mathbb{1}\{\hat{j}(\alpha/s, k) = j\}) \\
& \leq 169 \sum_{j=3}^{\infty} \mathbb{E}_{\mathbf{f}}(2^{-2j} \mathbb{1}\{\hat{j}(\alpha/s, k) = j, \acute{j}_k \leq j\}) + 169 \sum_{j=3}^{\infty} \mathbb{E}_{\mathbf{f}}(2^{-2j} \mathbb{1}\{\hat{j}(\alpha/s, k) = j, \acute{j}_k > j\}) \\
& \leq 169 \sum_{j=3}^{\infty} \mathbb{E}_{\mathbf{f}}(2^{-2\acute{j}_k} \mathbb{1}\{\hat{j}(\alpha/s, k) = j, \acute{j}_k \leq j\}) + 169 \sum_{j=3}^{\infty} \mathbb{E}_{\mathbf{f}}(2^{-2j} \mathbb{1}\{\hat{j}(\alpha/s, k) = j, \acute{j}_k > j\}) \\
& \leq 169 \mathbb{E}_{\mathbf{f}}(2^{-2\acute{j}_k}) + 169 \sum_{j=3}^{\infty} \mathbb{E}_{\mathbf{f}}(2^{-2j} \mathbb{1}\{\hat{j}(\alpha/s, k) = j, \acute{j}_k > j\}).
\end{aligned} \tag{A.5.109}$$

We will bound the two terms separately, now we start with the first term.

Note that we have  $\grave{j}_k \geq \acute{j}_k \geq \tilde{j}_k$  and that  $\tilde{j}_k = j$  implies one of the following happens:

$$\begin{aligned}
& \{X_{j, i_{j,k}^*+1, k} \geq X_{j, i_{j,k}^*+2, k}\}, \{X_{j, i_{j,k}^*+1, k} \geq X_{j, i_{j,k}^*+3, k}\}, \{X_{j, i_{j,k}^*+1, k} \geq X_{j, i_{j,k}^*+4, k}\}, \\
& \{X_{j, i_{j,k}^*-1, k} \geq X_{j, i_{j,k}^*-2, k}\}, \{X_{j, i_{j,k}^*-1, k} \geq X_{j, i_{j,k}^*-3, k}\}, \{X_{j, i_{j,k}^*-1, k} \geq X_{j, i_{j,k}^*-4, k}\}.
\end{aligned} \tag{A.5.110}$$

Also we have for  $j \geq j_k^* + 3$ ,  $m_j > \rho_z(\varepsilon; f_k)$ .

So we have

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}}(2^{-2\tilde{j}_k}) \\
& \leq \mathbb{E}_{\mathbf{f}}(2^{-2\tilde{j}_k}) \leq 2^{-2j_k^*+6} + \sum_{j=3}^{j_k^*-4} 2^{-2j} \mathbb{E}_{\mathbf{f}}(\mathbb{1}\{\tilde{j}_k = j\}) \\
& \leq 4\rho_z(\varepsilon; f_k)^2 + \sum_{j=3}^{j_k^*-4} 2^{-2j} \times 2 \times \left( \Phi\left(-\frac{\rho_m(\varepsilon; f_k)}{\rho_z(\varepsilon; f_k)} \frac{(2^{j_k^*-3-j}\rho_z(\varepsilon; f_k))^{\frac{3}{2}}}{\sqrt{3}\varepsilon}\right) + \right. \\
& \quad \left. \Phi\left(-2\frac{\rho_m(\varepsilon; f_k)}{\rho_z(\varepsilon; f_k)} \frac{(2^{j_k^*-3-j}\rho_z(\varepsilon; f_k))^{\frac{3}{2}}}{\sqrt{3}\varepsilon}\right) + \Phi\left(-3\frac{\rho_m(\varepsilon; f_k)}{\rho_z(\varepsilon; f_k)} \frac{(2^{j_k^*-3-j}\rho_z(\varepsilon; f_k))^{\frac{3}{2}}}{\sqrt{3}\varepsilon}\right) \right) \\
& \leq 4\rho_z(\varepsilon; f_k)^2 + \sum_{j=3}^{j_k^*-4} 2^{-2j} \times 2 \times \left( \Phi\left(-2^{\frac{3(j_k^*-3-j)}{2}} \frac{1}{\sqrt{2}} \frac{1}{\sqrt{3}}\right) + \right. \\
& \quad \left. \Phi\left(-2 \times 2^{\frac{3(j_k^*-3-j)}{2}} \frac{1}{\sqrt{2}} \frac{1}{\sqrt{3}}\right) + \Phi\left(-3 \times 2^{\frac{3(j_k^*-3-j)}{2}} \frac{1}{\sqrt{2}} \frac{1}{\sqrt{3}}\right) \right) \\
& \leq 4\rho_z(\varepsilon; f_k)^2 + 32\rho_z(\varepsilon; f_k)^2 \left( \frac{\Phi(-\frac{2}{\sqrt{3}})}{1 - 8\sqrt{2}\exp(-\frac{7}{2} \cdot \frac{4}{3})} + \frac{\Phi(-\frac{4}{\sqrt{3}})}{1 - 8\sqrt{2}\exp(-\frac{7}{2} \cdot \frac{16}{3})} \right. \\
& \quad \left. + \frac{\Phi(-2\sqrt{3})}{1 - 8\sqrt{2}\exp(-\frac{7}{2} \cdot 12)} \right) \\
& \leq 4\rho_z(\varepsilon; f_k)^2 + 4.5\rho_z(\varepsilon; f_k)^2 = 8.5\rho_z(\varepsilon; f_k)^2.
\end{aligned} \tag{A.5.111}$$

Now we turn to the second term in Inequality (A.5.109). We first define three quantities.

Let the average of  $f_k$  over  $[t_{j,i-1}, t_{j,i}]$  to be

$$\bar{f}_{j,i,k} = 2^j \int_{2^{-j} \times (i-1)}^{2^{-j} \times i} f_k(t) dt.$$

For  $i > 2^j$  or  $i \leq 0$ , define  $\bar{f}_{j,i,k} = +\infty$ . And suppose  $\infty - a = \infty$  for  $a \in [-\infty, \infty]$ , and  $\min\{\infty, a\} = a$  for  $a \in [-\infty, \infty]$ .

Let the minimum of the difference of the two neighboring intervals be

$$\Xi_{j,k} = \min\{\bar{f}_{j,i_k^*+2,k} - \bar{f}_{j,i_k^*+1,k}, \bar{f}_{j,i_k^*-2,k} - \bar{f}_{j,i_k^*-1,k}\}. \tag{A.5.112}$$

Let  $j(\zeta, k)$  be the level  $j$  such that the signal part in  $T_{j,k}$  is relatively small, specifically defined as follow.

$$j(\zeta, k) = \min\{j : \Xi_{j,k} \cdot 2^{-\frac{j}{2}} \frac{1}{\sqrt{6}\varepsilon} \leq z_\zeta + 1\}. \quad (\text{A.5.113})$$

Note that  $j(\zeta, k)$  is a determined quantity depending only on  $\zeta$  and  $f_k$ . Recall that  $\hat{j}(\alpha/s, k)$  is the stopping level, which is a random variable.

Also note that for  $j \leq j(\alpha/s, k) - 1$  we have

$$\Xi_{j,k} \cdot 2^{-\frac{j}{2}} \frac{1}{\sqrt{6}\varepsilon} \geq 2^{\frac{3(j(\alpha/s, k) - 1 - j)}{2}} (z_{\alpha/s} + 1) \quad (\text{A.5.114})$$

With these quantities, we have

$$\begin{aligned} & \sum_{j=3}^{\infty} 2^{-2j} \mathbb{E}_{\mathbf{f}}(\mathbb{1}\{\hat{j}(\alpha/s, k) = j, \acute{j}_k > j\}) \\ & \leq 2^{-2j(\alpha/s, k) + 1} + \sum_{j=3}^{j(\alpha/s, k) - 1} 2^{-2j} \Phi(-(z_{\alpha/s} + 1) \times 2^{\frac{3}{2}(j - j(\alpha/s, k) + 1)} + z_{\alpha/s}) \\ & \leq 2^{-2j(\alpha/s, k) + 1} + 2^{-2j(\alpha/s, k) + 2} \Phi(-1) \frac{1}{1 - \Phi(-2\sqrt{2})/\Phi(-1)} \\ & < 3 \cdot 2^{-2j(\alpha/s, k)}. \end{aligned} \quad (\text{A.5.115})$$

Now we introduce a lemma.

**Lemma A.5.1.** *For  $j(\zeta, k)$  defined in (A.5.113), with  $\zeta \leq 0.3$  we have*

$$\left(\frac{6\sqrt{2}(z_\zeta + 1)}{z_\zeta}\right)^{\frac{2}{3}} \rho_z(z_\zeta \varepsilon; f_k) \geq 2^{-j(\zeta, k)}. \quad (\text{A.5.116})$$

*Proof.* Without loss of generality, we assume

$$\bar{f}_{j, i_{j(\zeta, k), k} + 2, k} - \bar{f}_{j, i_{j(\zeta, k), k} + 1, k} = \Xi_{j(\zeta, k)}.$$

Let  $\mu_k = \min\{f_k(\max\{t_{j(\zeta,k),i_{j(\zeta,k),k}^*-2,0\}}, f_k(t_{j(\zeta,k),i_{j(\zeta,k),k}^*+1}))\}$ . Let the  $g_{lo} \in \mathcal{F}$  be defined as  $g_{lo}(t) = \max\{f_k(t), \mu_k\}$ .

For simplicity of notation, let  $j_0 = j(\zeta, k)$ ,  $i^* = i_{j(\zeta,k),k}^*$ .

Therefore,

$$\begin{aligned}
\|g_{lo} - f_k\|^2 &\leq (\mu_k - M(f_k))^2 \cdot 3 \cdot 2^{-j_0} \\
&\leq (f_k(t_{j_0,i^*+1}) - f_k(t_{j_0,i^*}) + f_k(t_{j_0,i^*}) - M(f_k))^2 \cdot 3 \cdot 2^{-j_0} \\
&\leq (\bar{f}_{j,i^*+2} - \bar{f}_{j,i^*+1})^2 \cdot 3 \cdot 2^{-j_0} \\
&\leq ((z_\zeta + 1) \cdot 2^{\frac{j_0}{2}} \sqrt{6\varepsilon})^2 \cdot 3 \cdot 2^{-j_0} \\
&= 6(z_\zeta + 1)^2 \times 3\varepsilon^2.
\end{aligned} \tag{A.5.117}$$

Therefore,

$$2^{-j_0} \leq \rho_z(3\sqrt{2}(z_\zeta + 1)\varepsilon; f_k) \leq \left(\frac{6\sqrt{2}(z_\zeta + 1)}{z_\zeta}\right)^{\frac{2}{3}} \rho_z(z_\zeta\varepsilon; f_k). \tag{A.5.118}$$

The last inequality is due to Proposition 2.2.1 in Chapter 2 and that  $z_\zeta \geq z_{0.3} = 0.524$

□

Lemma A.5.1 combined with Inequality (A.5.115), and note that  $\alpha/s \leq 0.3$  we have

$$\sum_{j=3}^{\infty} 2^{-2j} \mathbb{E}_{\mathbf{f}}(\mathbb{1}\{\hat{j}(\alpha/s, k) = j, \acute{j}_k > j\}) < 136\rho_z(z_{\alpha/s}\varepsilon; f_k)^2. \tag{A.5.119}$$

Also note that for  $\alpha \leq 0.3$ , we have  $\rho_z(\varepsilon; f_k) < 2.6\rho_z(z_{\alpha/s}\varepsilon; f_k)$ .

Therefore both terms in Inequality A.5.109 are bounded by multiple times  $\rho_z(z_{\alpha/s}\varepsilon; f_k)^2$ .

We conclude the proof.

### A.5.10. Proof of Theorem 3.3.3

Recalling Theorem 3.2.1 and Theorem 3.2.2, it suffice to prove

$$E \left( (\hat{M} - M(\mathbf{f}))^2 \right) \leq C_m \left( \sum_{k=1}^s \rho_m(\varepsilon; f_k) \right)^2, \quad (\text{A.5.120})$$

for an absolute positive constant  $C_m$ .

We proceed to prove this.

Recall that  $\zeta = \Phi(-2)$ .

Note that  $Y(1, 1, \dots, 1) - Y(0, 0, \dots, 0)$ ,  $2^{\hat{j}(\zeta, k)} \bar{X}_{\hat{j}(\zeta, k), i_{F, k}, k}$  for  $k = 1, 2, \dots, s$  are independent. Therefore,

$$\begin{aligned} \mathbb{E} \left( (\hat{M} - M(\mathbf{f}))^2 \right) &\leq \\ &\left( \sqrt{\mathbb{E}(Y(1, 1, \dots, 1) - Y(0, 0, \dots, 0) - f_0)^2} + \sum_{k=1}^s \sqrt{\mathbb{E} \left( 2^{\hat{j}(\zeta, k)} \bar{X}_{\hat{j}(\zeta, k), i_{F, k}, k} - M(f_k) \right)^2} \right)^2. \end{aligned} \quad (\text{A.5.121})$$

Recollect the notation

$$\bar{f}_{j, i, k} = 2^j \int_{2^{-j}(i-1)}^{2^{-j} \cdot i} f_k(t) dt. \quad (\text{A.5.122})$$

Recall that the location procedure, the stopping rule and the definition of  $i_{F, k}$  parallel those of univariate case introduced in Chapter 2, so we have that  $\bar{f}_{\hat{j}(\zeta, k), i_{F, k}, k}$  has the same distribution with that of  $\hat{f}$  in the proof of Theorem 2.3.3 with  $f_k$  being the true function.

Hence we have that

$$\mathbb{E} \left( \bar{f}_{\hat{j}(\zeta, k), i_{F, k}, k} - M(f_k) \right)^2 \leq \tilde{C}_m \rho_m(\varepsilon; f_k)^2 \quad (\text{A.5.123})$$

for all  $k \in \{1, 2, \dots, s\}$ , where  $\tilde{C}_m$  is a positive absolute constant.

Also note that

$$\bar{X}_{\hat{j}(\zeta,k),i_{F,k,k}}|\hat{j}(\zeta,k),i_{F,k}) \sim N(\bar{f}_{\hat{j}(\zeta,k),i_{F,k,k}}, (1 - 2^{-\hat{j}(\zeta,k)})2^{-\hat{j}(\zeta,k)} \times 3\varepsilon^2).$$

So we have that

$$\begin{aligned} & \mathbb{E} \left( 2^{\hat{j}(\zeta,k)} \bar{X}_{\hat{j}(\zeta,k),i_{F,k,k}} - M(f_k) \right)^2 \\ &= \mathbb{E} \left( 2^{\hat{j}(\zeta,k)} \bar{X}_{\hat{j}(\zeta,k),i_{F,k,k}} - \bar{f}_{\hat{j}(\zeta,k),i_{F,k,k}} \right)^2 + \mathbb{E} \left( \bar{f}_{\hat{j}(\zeta,k),i_{F,k,k}} - M(f_k) \right)^2 \\ &\leq \mathbb{E}((1 - 2^{-\hat{j}(\zeta,k)})2^{\hat{j}(\zeta,k)} \times 3\varepsilon^2) + \tilde{C}_m \rho_m(\varepsilon; f_k)^2. \end{aligned} \quad (\text{A.5.124})$$

Now we will bound  $\mathbb{E}(2^{\hat{j}(\zeta,k)} \times 3\varepsilon^2)$ . Note that  $\zeta = \Phi(-2) < 0.3$ , so we have that

$$\begin{aligned} \mathbb{E}_{\mathbf{f}}(2^{\hat{j}(\zeta,k)}) &\leq \sum_{j=1}^{j(\zeta,k)+3} \mathbb{E}_{\mathbf{f}}(\hat{j}(\zeta,k) = j) \times 2^j + \sum_{j=j(\zeta,k)+4}^{\infty} \mathbb{E}_{\mathbf{f}}(\hat{j}(\zeta,k) = j) \times 2^j \\ &\leq 2^{j(\zeta,k)+4} + \sum_{j=j(\zeta,k)+4}^{\infty} 2^j \Phi(-z_{\zeta} + \frac{z_{\zeta} + 1}{64})^{j-j(\zeta,k)-4} \\ &\leq 2^{j(\zeta,k)+4} + 2^{j(\zeta,k)+4} \cdot \frac{1}{1 - 0.03} \leq \frac{4}{\rho_z(z_{\zeta}\varepsilon; f_k)} \times 33. \end{aligned} \quad (\text{A.5.125})$$

The last inequality is due to Lemma A.5.3.

Going back to Inequality (A.5.121) we have that

$$\begin{aligned} \mathbb{E} \left( (\hat{M} - M(\mathbf{f}))^2 \right) &\leq \left( \varepsilon + \sum_{k=1}^s \sqrt{132 \frac{3\varepsilon^2}{\rho_z(z_{\zeta}\varepsilon; f_k)} + \tilde{C}_m \rho_m(\varepsilon; f_k)^2} \right)^2 \\ &\leq \left( \varepsilon + \sum_{k=1}^s \sqrt{800 + \tilde{C}_m} \times \rho_m(z_{\zeta}\varepsilon; f_k) \right)^2 \\ &\leq C_m \left( \sum_{k=1}^s \rho_m(\varepsilon; f_k) \right)^2. \end{aligned} \quad (\text{A.5.126})$$

### A.5.11. Proof of Theorem 3.3.4

Recalling the lower bound for  $L_{\alpha,m}(\varepsilon; \mathbf{f})$  established in Theorem 3.2.1 and Theorem 3.2.2, it suffices to prove the following propositions.

**Proposition A.5.9** (Coverage). *The confidence interval  $CI_{m,\alpha}$  defined by (3.3.18) is an  $1 - \alpha$  level confidence cube for minimum.*

**Proposition A.5.10** (Expected Length). *For  $\alpha \leq 0.3$ , and confidence interval  $CI_{m,\alpha}$  defined by (3.3.18), we have*

$$\mathbb{E}_{\mathbf{f}}(|CI_{m,\alpha}|) \leq \tilde{C}_{m,s,\alpha} \sum_{k=1}^s \rho_m(\varepsilon; f_k), \quad (\text{A.5.127})$$

where  $\tilde{C}_{m,s,\alpha}$  is an absolute positive constant depending on  $s$  and  $\alpha$ .

### Proof of Proposition A.5.9

Recall that  $\zeta = \alpha/4s$ . Let the event  $A_1$  be

$$A_1 = \left\{ Z(f_k) \in [2^{-\hat{j}(\zeta,k)+1} \times (\hat{i}_{\hat{j}(\zeta,k)-1,k} - 7), 2^{-\hat{j}(\zeta,k)+1} \times (\hat{i}_{\hat{j}(\zeta,k)-1,k} + 6)] \right. \\ \left. \text{for all } k \in \{1, 2, \dots, s\} \right\}. \quad (\text{A.5.128})$$

Then from Theorem 3.3.2 we know that  $P(A_1) \geq 1 - \alpha/4$ . Easy calculation shows that  $A_1$  can also be written as

$$A_1 = \{Z(f_k) \in [2^{-\hat{j}(\zeta,k)-3} \cdot 16(\hat{i}_{\hat{j}(\zeta,k)-1,k} - 7), 2^{-\hat{j}(\zeta,k)-3} \cdot 16(\hat{i}_{\hat{j}(\zeta,k)-1,k} + 6)]\} \quad (\text{A.5.129})$$

Let the event  $D_{2,k}$  be

$$D_{2,k} = \{\hat{j}(\alpha/4s, k) \leq j(\alpha/4s, k) - 2\},$$

where  $j(\zeta, k)$  is defined in (A.5.113). By definition of  $j(\zeta, k)$  we know that for  $j \leq j(\zeta, k) - 1$

$$\Xi_{j,k} \cdot 2^{-\frac{j}{2}} \frac{1}{\sqrt{6\varepsilon}} > 2^{\frac{3}{2}(j(\zeta,k)-1-j)} (z_{\alpha/4s} + 1). \quad (\text{A.5.130})$$

Therefore, we have

$$\begin{aligned} & P(D_{2,k} \cap \{|\hat{i}_{\hat{j}(\zeta,k),k} - i_{j(\zeta,k),k}^*| \leq 4\}) \\ & \leq \sum_{j=1}^{j(\alpha/4s,k)-1} P\left(\hat{j}(\zeta, k) = j, |\hat{i}_{j,k} - i_{j,k}^*| \leq 4\right) \\ & \leq \Phi(-z_{\alpha/4s} - 1) \sum_{j=1}^{j(\alpha/4s,k)-1} P(|\hat{i}_{j,k} - i_{j,k}^*| \leq 4). \end{aligned} \quad (\text{A.5.131})$$

Additionally, recall  $\tilde{j}_k$  defined in (A.5.106), we have

$$\begin{aligned} & P\left(\{|\hat{i}_{\hat{j}(\zeta,k),k} - i_{j,k}^*| \geq 5, \hat{j}(\zeta, k) \leq j(\alpha/4s, k) - 1\}\right) \\ & \leq P(\tilde{j}_k \leq j(\alpha/4s, k) - 2) \leq 6 \sum_{j=1}^{j(\alpha/4s,k)-2} \Phi(-2^{3 \cdot (j(\alpha/4s,k)-1-j)/2} (z_{\alpha/4s} + 1) + z_{\alpha/4s}) \\ & \leq 6 \times \Phi(-z_{\alpha/4s} - 2\sqrt{2}) \times 1.000001. \end{aligned} \quad (\text{A.5.132})$$

Therefore, for  $\alpha \leq 0.3$ ,

$$\begin{aligned} P(D_{2,k}) & \leq \Phi(-z_{\alpha/4s} - 1) + 6.000006 \times \Phi(-z_{\alpha/4s} - 2\sqrt{2}) \\ & \leq (\alpha/4s) \times \left(\frac{4}{3} \cdot \exp(-1.5) + 6.000006 \times \frac{4}{3} \exp(-4)\right) \leq \alpha/8s. \end{aligned} \quad (\text{A.5.133})$$

Note that for each  $k$

$$\begin{aligned} & 2^{\hat{j}(\zeta,k)+3} \times \bar{X}_{\hat{j}(\zeta,k)+3,i,k} - \int_{t_{\hat{j}(\zeta,k)+3,i-1,k}}^{t_{\hat{j}(\zeta,k)+3,i,k}} f_k(t) \cdot 2^{\hat{j}(\zeta,k)+3} dt \\ & + Y(1, 1, \dots, 1) - Y(0, 0, \dots, 0) - f_0 - \sqrt{2\varepsilon} \int_0^1 B_k^1(x) dx \Big| \hat{j}(\zeta, k) \end{aligned} \quad (\text{A.5.134})$$



for  $i = 1, 2, \dots, s$  are i.i.d  $N(0, 2^{\hat{j}(\zeta, k)+3} \times 3\varepsilon^2)$ . And

$$Y(1, 1, \dots, 1) - Y(0, 0, \dots, 0) - f_0 - \sum_{k=1}^s \left( Y(1, 1, \dots, 1) - Y(0, 0, \dots, 0) - f_0 - \sqrt{2}\varepsilon \int_0^1 B_k^1(x) dx \right) \sim N(0, \varepsilon^2((s-1)^2 + 2s)). \quad (\text{A.5.135})$$

Hence we have that

$$P\left(\mathbf{f}_{hi} \leq M(\mathbf{f}) \middle| A_1\right) \leq \frac{\alpha}{4}. \quad (\text{A.5.136})$$

Also note that on the event  $A_1 \cap D_{2,k}^c$ , there is a random variable such that

$$\begin{aligned} v_k | \hat{j}(\zeta, k) &\sim N(0, 3(1 - 2^{-\hat{j}(\zeta, k)-3}) 2^{\hat{j}(\zeta, k)+3} \varepsilon^2), \\ 2^{\hat{j}(\zeta, k)+3} \min_{16 \cdot (\hat{i}_{\hat{j}(\zeta, k)-1, k} - 7) < i \leq 16 \cdot (\hat{i}_{\hat{j}(\zeta, k)-1, k} + 6)} \bar{X}_{\hat{j}(\zeta, k)+3, i, k} \\ &\leq M(f_k) + \rho_m(z_\zeta \varepsilon; f_k) + v_k \\ &\leq M(f_k) + \sqrt{3}\varepsilon z_\zeta \frac{1}{\sqrt{\rho_z(z_\zeta \varepsilon; f_k)}} + v_k, \end{aligned} \quad (\text{A.5.137})$$

and  $v_1, v_2, \dots, v_k$  are independent.

Recall Lemma A.5.1 and the definition of  $D_{2,k}^c$ , we have on the event  $A_1 \cap D_{2,k}^c$

$$2^{\hat{j}(\zeta, k)+3} \min_{16 \cdot (\hat{i}_{\hat{j}(\zeta, k)-1, k} - 7) < i \leq 16 \cdot (\hat{i}_{\hat{j}(\zeta, k)-1, k} + 6)} \bar{X}_{\hat{j}(\zeta, k)+3, i, k} \leq M(f_k) + \sqrt{3}\varepsilon z_\zeta \frac{1}{\sqrt{\rho_z(z_\zeta \varepsilon; f_k)}} + v_k. \quad (\text{A.5.138})$$

So we have that

$$P\left(\mathbf{f}_{lo} \geq M(\mathbf{f}) \middle| A_1 \cap \left(\bigcap_{k=1}^s D_{2,k}^c\right)\right) \leq \frac{\alpha}{4}. \quad (\text{A.5.139})$$

Adding the components, we have

$$\begin{aligned}
& P(M(\mathbf{f}) \notin [\mathbf{f}_{lo}, \mathbf{f}_{hi}]) \leq \\
& P(A_1^c) + \sum_{k=1}^s P(D_{2,k}) + P(\mathbf{f}_{lo} \geq M(\mathbf{f}) \mid A_1 \cap (\cap_{k=1}^s D_{2,k}^c)) + P(\mathbf{f}_{hi} \leq M(\mathbf{f}) \mid A_1) \leq \alpha.
\end{aligned} \tag{A.5.140}$$

### Proof of Proposition A.5.10

As  $\hat{j}(\zeta, 1), \hat{j}(\zeta, 2), \dots, \hat{j}(\zeta, s)$  based on independent random variables, they are independent.

Hence we have

$$\mathbb{E}(|\mathbf{f}_{hi} - \mathbf{f}_{lo}|^2) \leq \left( 2\sqrt{6}\varepsilon (S_{208, \alpha/8s} + z_{\alpha/4} + 2z_{\alpha/4s} + 2z_{\alpha/8}) \sum_{k=1}^s \mathbb{E}(2^{\frac{\hat{j}(\zeta, k)}{2}}) \right)^2. \tag{A.5.141}$$

Now we will prove the following lemma.

**Lemma A.5.2.** For  $k = 1, 2, \dots, s$ , for  $\zeta \leq 0.3$ ,

$$\mathbb{E}(2^{\frac{\hat{j}(\zeta, k)}{2}}) \leq 12.7 \times 2^{\frac{j(\zeta, k)}{2}}, \tag{A.5.142}$$

where  $j(\zeta, k)$  is defined in (A.5.113).

*Proof.*

$$\begin{aligned}
\mathbb{E}_{\mathbf{f}}(2^{\frac{\hat{j}(\zeta, k)}{2}}) & \leq \sum_{j=1}^{j(\zeta, k)+3} \mathbb{E}_{\mathbf{f}}(\hat{j}(\zeta, k) = j) \times 2^{\frac{j}{2}} + \sum_{j=j(\zeta, k)+4}^{\infty} \mathbb{E}_{\mathbf{f}}(\hat{j}(\zeta, k) = j) \times 2^{\frac{j}{2}} \\
& \leq 2^{\frac{j(\zeta, k)+5}{2}} + \sum_{j=j(\zeta, k)+4}^{\infty} 2^{\frac{j}{2}} \Phi(-z_{\zeta} + \frac{z_{\zeta} + 1}{64})^{j-j(\zeta, k)-4} \\
& \leq 2^{\frac{j(\zeta, k)+5}{2}} + 2^{\frac{j(\zeta, k)+4}{2}} \times 1.74803 \leq 12.7 \times 2^{\frac{j(\zeta, k)}{2}}
\end{aligned} \tag{A.5.143}$$

□

To bound  $2^{\frac{j(\zeta, k)}{2}}$ , we continue with another lemma

**Lemma A.5.3.** *For  $\zeta \leq 0.3$ , and  $k = 1, 2, \dots, s$  we have*

$$2^{-j(\zeta, k)} \geq \frac{1}{4} \rho_z(z_\zeta \varepsilon; f_k). \quad (\text{A.5.144})$$

*Proof.* Without loss of generality, assume  $f_k(Z(f_k) + \rho_z(z_\zeta \varepsilon; f_k)) \leq \rho_m(z_\zeta \varepsilon; f_k)$ . Suppose  $2^{-j} \leq \frac{1}{4} \rho_z(z_\zeta \varepsilon; f_k)$ , then we have that

$$\begin{aligned} & (\bar{f}_{j, i_{j,k}^*+2, k} - \bar{f}_{j, i_{j,k}^*+1, k}) \cdot 2^{-\frac{j}{2}} \frac{1}{\sqrt{6}\varepsilon} \\ & \leq \rho_m(z_\zeta \varepsilon; f_k) \cdot \frac{1}{2} \sqrt{\rho_z(z_\zeta \varepsilon; f_k)} \frac{1}{\sqrt{6}\varepsilon} \leq \frac{1}{2\sqrt{2}} z_\zeta \leq z_\zeta + 1. \end{aligned} \quad (\text{A.5.145})$$

Therefore,  $j \geq j(\zeta, k)$ , thus  $2^{-j(\zeta, k)} \geq \frac{1}{4} \rho_z(z_\zeta \varepsilon; f_k)$ .

□

Combing Lemma A.5.2 with Lemma A.5.3 and getting back to Inequality (A.5.141), we have

$$\begin{aligned} & \mathbb{E}(|\mathbf{f}_{hi} - \mathbf{f}_{lo}|^2) \\ & \leq \left( 2\sqrt{6}\varepsilon (S_{208, \alpha/8s} + z_{\alpha/4} + 2z_{\alpha/4s} + 2z_{\alpha/8}) \sum_{k=1}^s 12.7 \times 2 \frac{1}{\sqrt{\rho_z(z_{\alpha/4s}\varepsilon; f_k)}} \right)^2 \\ & \leq \left( 8\sqrt{3} \times 12.7 \times (S_{208, \alpha/8s} + z_{\alpha/4} + 2z_{\alpha/4s} + 2z_{\alpha/8}) \frac{1}{z_{\alpha/4s}} \sum_{k=1}^s \rho_m(z_{\alpha/4s}\varepsilon; f_k) \right)^2. \end{aligned} \quad (\text{A.5.146})$$

Note that

$$\rho_m(z_{\alpha/4s}\varepsilon; f_k) \leq z_{\alpha/4s} \rho_m(\varepsilon; f_k), \quad (\text{A.5.147})$$

and

$$\mathbb{E}(|\mathbf{f}_{hi} - \mathbf{f}_{lo}|) \leq \sqrt{\mathbb{E}(|\mathbf{f}_{hi} - \mathbf{f}_{lo}|^2)}. \quad (\text{A.5.148})$$

Therefore, we have the statement.

#### A.5.12. Analysis of Local Minimax Rates for Nonparametric Regression

In this section, we give lower bounds for the benchmarks defined in (3.4.2) and (3.4.3).

An additional complexity for the nonparametric regression is that two functions  $\mathbf{f}$  and  $\mathbf{g}$  can have same values on all grid points  $\frac{\mathbf{i}}{n}$  while have different minimizers or minimums. We call this error caused by discretization *discretization error*:

$$\mathfrak{D}_z(\mathbf{f}; n) = \sup_{\mathbf{g} \in \mathcal{F}_s} \{ \|Z(\mathbf{f}) - Z(\mathbf{g})\|^2 : \mathbf{f}(\frac{\mathbf{i}}{n}) = \mathbf{g}(\frac{\mathbf{i}}{n}) \text{ for all } \mathbf{i} \in \{0, 1, \dots, n\}^s \}, \quad (\text{A.5.149})$$

$$\mathfrak{D}_m(\mathbf{f}; n) = \sup_{\mathbf{g} \in \mathcal{F}_s} \{ |M(\mathbf{f}) - M(\mathbf{g})| : \mathbf{f}(\frac{\mathbf{i}}{n}) = \mathbf{g}(\frac{\mathbf{i}}{n}) \text{ for all } \mathbf{i} \in \{0, 1, \dots, n\}^s \}. \quad (\text{A.5.150})$$

$$(\text{A.5.151})$$

Note that while the discretization errors are defined for  $\mathbf{f} \in \mathcal{F}_s$ , they are also well defined for univariate convex functions by setting  $s = 1$ . With a bit abuse of notation, we use them directly for univariate convex functions as well by plugging in univariate convex function  $f$  in the place of the multivariate convex function  $\mathbf{f}$ .

It's apparent that

$$\tilde{\mathbf{R}}_{z,n}(\sigma; \mathbf{f}) \geq \frac{1}{4} \mathfrak{D}_z(\mathbf{f}; n), \tilde{\mathbf{R}}_{m,n}(\sigma; \mathbf{f}) \geq \frac{1}{4} \mathfrak{D}_m(\mathbf{f}; n)^2, \tilde{\mathbf{L}}_{m,\alpha,n}(\sigma; \mathbf{f}) \geq (1 - 2\alpha) \mathfrak{D}_m(\mathbf{f}; n). \quad (\text{A.5.152})$$

For simplicity of notation, for  $\varepsilon > 0$ , we define

$$\varphi_z(\varepsilon; f) = \rho_z(\varepsilon; f) \left( 1 \wedge \sqrt{n\rho_z(\varepsilon; f)} \right), \text{ for } f \in \mathcal{F}, \quad (\text{A.5.153})$$

$$\varphi_m(\varepsilon; f) = \rho_m(\varepsilon; f) \left( 1 \wedge \sqrt{n\rho_z(\varepsilon; f)} \right), \text{ for } f \in \mathcal{F}. \quad (\text{A.5.154})$$

Now we state the lower bounds for the benchmarks, whose proof will be given later.

$$\tilde{\mathbf{R}}_{z,n}(\sigma; \mathbf{f}) \geq \left( 0.1 \times \frac{1}{12s} \sum_{k=1}^s \varphi_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2 \right) \vee \frac{\mathfrak{D}_z(\mathbf{f}; n)}{4}, \quad (\text{A.5.155})$$

$$\tilde{\mathbf{L}}_{z,\alpha,n}(\sigma; \mathbf{f}) \geq \frac{1-\alpha-\Phi(-z_\alpha+1)}{(12s)^{s/2}} \Pi_{k=1}^s \left( \sqrt{\mathfrak{D}_z(f_k; n)} \vee \varphi_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \right), \quad (\text{A.5.156})$$

$$\tilde{\mathbf{R}}_{m,n}(\sigma; \mathbf{f}) \geq \left( \frac{1}{180} \sum_{k=1}^s \varphi_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2 \frac{1}{1 + \frac{s}{n} + \sum_{k=1}^s 2\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)} \right) \vee \frac{1}{2} \mathfrak{D}_m(\mathbf{f}; n)^2, \quad (\text{A.5.157})$$

$$\begin{aligned} \tilde{\mathbf{L}}_{m,\alpha,n}(\sigma; \mathbf{f}) &\geq (1 - \alpha - \Phi(-z_\alpha + 1)) \cdot \\ &\left( \frac{1}{3\sqrt{2}} \sqrt{\sum_{k=1}^s \varphi_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2} \sqrt{\frac{1}{1 + \frac{s}{n} + \sum_{k=1}^s 2\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}} \vee \mathfrak{D}_m(\mathbf{f}; n) \right). \end{aligned} \quad (\text{A.5.158})$$

Before continue with the proofs of the lower bounds (A.5.155), (A.5.156), (A.5.157), and (A.5.158) separately, we introduce some quantities and lemmas that will be frequently used.

We introduce a function  $l_n(\cdot, \cdot)$ . For  $f, g \in \mathcal{F}$

$$l_n(f, g) = \sqrt{\frac{\sum_{j=1}^n (f(\frac{j}{n}) - g(\frac{j}{n}))^2}{n+1}}. \quad (\text{A.5.159})$$

$l_n$  can be considered as a discrete  $L_2$  norm of the difference of function  $f$  and  $g$ .

We also have the following lemma.

**Lemma A.5.4.** *For  $f \in \mathcal{F}$ ,  $\varepsilon > 0$ , and  $\delta > 0$ , there exist  $g \in \mathcal{F}$  such that*

$$l_n(f, g) \leq \sqrt{6\varepsilon}, \quad (\text{A.5.160})$$

and that

$$\begin{aligned}
|Z(f) - Z(g)| &\geq \rho_z(\varepsilon; f)(1 \wedge \sqrt{2n\rho_z(\varepsilon; f)}) - \delta, \\
M(g) - M(f) &\geq \rho_m(\varepsilon; f)(1 \wedge \sqrt{2n\rho_z(\varepsilon; f)}) - \delta, \\
g(t) &\geq f(t) \text{ for } 0 \leq t \leq 1, \\
\frac{1}{n+1} \sum_{i=0}^n (g(\frac{i}{n}) - f(\frac{i}{n})) &\leq l_n(f, g) \sqrt{\frac{1}{n} + 2\rho_z(\varepsilon; f)}.
\end{aligned} \tag{A.5.161}$$

*Proof.* Suppose  $\eta > 0$  is a small number. For  $\mu > 0$ , we next define convex function  $g_{\eta, \mu}$ . Suppose  $t_{l, \mu}, t_{r, \mu}$  are left and right end points of  $\{t : f(t) \leq \mu + M(f)\}$ . When  $t_{l, \mu} + t_{r, \mu} \geq 2Z(f)$ .

$$g_{\eta, \mu}(t) = \max\{f(t), \mu + M(f) + \frac{-\eta}{t_{r, \mu} - t_{l, \mu}}(t - t_{l, \mu})\}. \tag{A.5.162}$$

When  $t_{l, \mu} + t_{r, \mu} \leq 2Z(f)$ .

$$g_{\eta, \mu}(t) = \max\{f(t), \mu + M(f) + \frac{\eta}{t_{r, \mu} - t_{l, \mu}}(t - t_{r, \mu})\}. \tag{A.5.163}$$

For  $\rho_z(\varepsilon; f) \geq \frac{1}{2n}$ , we have

$$l_n(f, g_{\eta, \rho_m(\varepsilon; f)}) \leq \sqrt{6}\|f - g\| \leq \sqrt{6}\varepsilon, \tag{A.5.164}$$

for any  $\eta > 0$ . And we also have that

$$\lim_{\eta \rightarrow 0^+} |Z(g_{\eta, \varepsilon}) - Z(f)| \geq \rho_z(\varepsilon; f). \tag{A.5.165}$$

For  $\rho_z(\varepsilon; f) \leq \frac{1}{2n}$ , we have that

$$l_n(f, g_{\eta, \rho_m(\varepsilon; f)\sqrt{2n\rho_z(\varepsilon; f)}}) \leq \sqrt{6}\|f - g\| \leq \sqrt{6}\varepsilon, \tag{A.5.166}$$

for any  $\eta > 0$ .

$$\lim_{\eta \rightarrow 0^+} |Z(g_{\eta, \varepsilon \sqrt{2n\rho_z(\varepsilon; f)}}) - Z(f)| \geq \rho_z(\varepsilon; f) \sqrt{2n\rho_z(\varepsilon; f)}. \quad (\text{A.5.167})$$

Let  $\mu = \rho_m(\varepsilon; f)(1 \wedge \sqrt{2n\rho_z(\varepsilon; f)})$ .

Then we have that

$$\begin{aligned} l_n(f, g_{\eta, \mu}) &\leq 6\varepsilon^2, \\ \lim_{\eta \rightarrow 0^+} M(g_{\eta, \mu}) - M(f) &\geq \rho_m(\varepsilon; f)(1 \wedge \sqrt{2n\rho_z(\varepsilon; f)}), \\ \lim_{\eta \rightarrow 0^+} |Z(g_{\eta, \mu}) - Z(f)| &\geq \rho_z(\varepsilon; f)(1 \wedge \sqrt{2n\rho_z(\varepsilon; f)}), \\ g_{\eta, \mu}(t) &\geq f(t) \text{ for all } 0 \leq t \leq 1, \\ \left( \frac{1}{n+1} \sum_{i=1}^n (g_{\eta, \mu}(\frac{i}{n}) - f(\frac{i}{n})) \right) \\ &\leq l_n(f, g_{\eta, \mu})^2 \frac{|\{i : g_{\eta, \mu}(\frac{i}{n}) > f(\frac{i}{n})\}|}{n+1} \leq l_n(f, g_{\eta, \mu})^2 \frac{2n\rho_z(\varepsilon; f) + 1}{n+1}. \end{aligned} \quad (\text{A.5.168})$$

Take  $\eta$  small enough gives the statement.

□

Now we continue with analyzing the probability structure of the nonparametric regression setting.

For  $\mathbf{f}, \mathbf{g} \in \mathcal{F}_s$ , denote the probability distribution under  $\mathbf{f}$  as  $P_{\mathbf{f}}$  and that under  $\mathbf{g}$  as  $P_{\mathbf{g}}$ .

Then for observation  $\{y_{\mathbf{i}}\}$ , we have

$$\log \left( \frac{P_{\mathbf{f}}(\{y_{\mathbf{i}}\})}{P_{\mathbf{g}}(\{y_{\mathbf{i}}\})} \right) = \sum_{\mathbf{i} \in \{0, 1, \dots, n\}^s} \left( \frac{y_{\mathbf{i}}(\mathbf{f}(\mathbf{i}) - \mathbf{g}(\mathbf{i}))}{\sigma^2} + \frac{-\mathbf{f}(\mathbf{i})^2 + \mathbf{g}(\mathbf{i})^2}{2\sigma^2} \right). \quad (\text{A.5.169})$$

If we set  $\mathbf{f}_\theta = \mathbf{f}\mathbb{1}\{\theta = 1\} + \mathbf{g}\mathbb{1}\{\theta = -1\}$ , then we know that

$$W = \sum_{\mathbf{i} \in \{0,1,\dots,n\}^s} \frac{y_{\mathbf{i}}(\mathbf{f}(\mathbf{i})) - \mathbf{g}(\mathbf{i}))}{\sigma \sqrt{\sum_{\mathbf{i} \in \{0,1,\dots,n\}^s} (\mathbf{f}(\mathbf{i})) - \mathbf{g}(\mathbf{i}))^2}} + \frac{-\mathbf{f}(\mathbf{i})^2 + \mathbf{g}(\mathbf{i})^2}{2\sigma \sqrt{\sum_{\mathbf{i} \in \{0,1,\dots,n\}^s} (\mathbf{f}(\mathbf{i})) - \mathbf{g}(\mathbf{i}))^2}} \quad (\text{A.5.170})$$

is a sufficient statistic for  $\theta$ , and

$$W \sim N\left(\theta \frac{1}{2} \frac{\sqrt{\sum_{\mathbf{i} \in \{0,1,\dots,n\}^s} (\mathbf{f}(\mathbf{i})) - \mathbf{g}(\mathbf{i}))^2 / (n+1)^s}}{\sigma / (n+1)^{\frac{s}{2}}}, 1\right). \quad (\text{A.5.171})$$

### Proof of Inequality (A.5.155)

Recall Lemma A.5.4, take  $\varepsilon^2 = \frac{\sigma^2}{6(n+1)^s} \frac{1}{s}$ . Take

$$\delta < 0.001 \min_{1 \leq k \leq s} \rho_z(\varepsilon; f_k) \left(1 \wedge \sqrt{n\rho_z(\varepsilon; f_k)}\right).$$

Take  $g_{k,\delta}$  to be the function satisfying (A.5.161) in Lemma A.5.4 for  $f = f_k$ . Let

$$h_{k,\delta}(t) = g_{k,\delta}(t) - \frac{1}{n+1} \sum_{i=0}^n (g_{k,\delta}(\frac{i}{n}) - f_k(\frac{i}{n})). \quad (\text{A.5.172})$$

Let

$$\mathbf{h}_\delta(\mathbf{t}) = f_0 + \sum_{k=1}^s h_{k,\delta}(t_k). \quad (\text{A.5.173})$$

It's easy to check  $\mathbf{h}_\delta \in \mathcal{F}_s$ .

Then Lemma A.5.4 together with elementary calculation show that

$$\frac{\sqrt{\sum_{\mathbf{i} \in \{0,1,\dots,n\}^s} (\mathbf{f}(\mathbf{i})) - \mathbf{g}(\mathbf{i}))^2 / (n+1)^s}}{\sigma / (n+1)^{\frac{s}{2}}} \leq 1, \quad (\text{A.5.174})$$



and that

$$\|Z(\mathbf{h}_\delta) - Z(\mathbf{f})\|^2 \geq \sum_{k=1}^s \left( \rho_z(\varepsilon; f_k) \left( 1 \wedge \sqrt{2n\rho_z(\varepsilon; f_k)} \right) - \delta \right)^2. \quad (\text{A.5.175})$$

Recall that  $W$  defined in (A.5.171) is sufficient statistic for  $\theta$ , we have

$$\begin{aligned} \tilde{\mathbf{R}}_{z,n}(\sigma; \mathbf{f}) &\geq \inf_{\hat{Z}} \max\{\mathbb{E}_{\mathbf{f}}(\|\hat{Z} - Z(\mathbf{f})\|^2), \mathbb{E}_{\mathbf{h}_\delta}(\|\hat{Z} - Z(\mathbf{h}_\delta)\|^2)\} \geq r_2 \|Z(\mathbf{f}) - Z(\mathbf{h}_\delta)\|^2, \\ &\geq r_2 \sum_{k=1}^s \left( \rho_z(\varepsilon; f_k) \left( 1 \wedge \sqrt{2n\rho_z(\varepsilon; f_k)} \right) - \delta \right)^2, \end{aligned} \quad (\text{A.5.176})$$

where

$$r_2 = \inf_{\hat{\theta}} \max_{\theta=\pm 1} \mathbb{E}_{\theta} \frac{|\hat{\theta} - \theta|^2}{4},$$

for  $W \sim N(\frac{\theta}{2}, 1)$ . Elementary calculation shows that  $r_2 > 0.1$ .

Now we take  $\delta \rightarrow 0^+$ , we have that

$$\begin{aligned} \tilde{\mathbf{R}}_{z,n}(\sigma; \mathbf{f}) &\geq 0.1 \sum_{k=1}^s \rho_z(\varepsilon; f_k)^2 (1 \wedge 2n\rho_z(\varepsilon; f_k)) \\ &\geq 0.1 \times \frac{1}{12s} \sum_{k=1}^s \varphi_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2, \end{aligned} \quad (\text{A.5.177})$$

where the last inequality comes from Proposition 2.2.1.

Note that  $\tilde{\mathbf{R}}_{z,n}(\sigma; \mathbf{f}) \geq \frac{\mathfrak{D}_z(\mathbf{f}; n)}{4}$  apparently. We concludes the proof.

**Proof of Inequality (A.5.156)**

Take  $h_{k,\delta}$  constructed in (A.5.172).

Let  $\tilde{\delta} < 0.01$  be a small positive number.

Take  $f_{k,alt,\tilde{\delta}} \in \mathcal{F}$  satisfying

$$\begin{aligned} f_{k,alt,\tilde{\delta}}\left(\frac{i}{n}\right) &= f_k\left(\frac{i}{n}\right) \text{ for } 0 \leq i \leq n, \\ |Z(f_{k,alt,\tilde{\delta}}) - Z(f_k)| &\geq \frac{1}{2}\sqrt{(1-\tilde{\delta})\mathfrak{D}_z(f_k;n)}. \end{aligned} \quad (\text{A.5.178})$$

Take

$$\begin{aligned} \mathbf{h}_{\delta,\tilde{\delta}}(\mathbf{t}) &= f_0 + \sum_k^s \left( h_{k,\delta}(t_k) \mathbb{1}\{|Z(h_{k,\delta}) - Z(f_k)| \geq |Z(f_{k,alt,\tilde{\delta}}) - Z(f_k)|\} \right. \\ &\quad \left. + f_{k,alt,\tilde{\delta}}(t_k) \mathbb{1}\{|Z(h_{k,\delta}) - Z(f_k)| < |Z(f_{k,alt,\tilde{\delta}}) - Z(f_k)|\} \right). \end{aligned} \quad (\text{A.5.179})$$

It's easy to check that  $\mathbf{h}_{\delta,\tilde{\delta}} \mathcal{F}_s$ .

Then we have that

$$\frac{\sqrt{\sum_{\mathbf{i} \in \{0,1,\dots,n\}^s} (\mathbf{f}(\mathbf{i}) - \mathbf{g}(\mathbf{i}))^2 / (n+1)^s}}{\sigma / (n+1)^{\frac{s}{2}}} \leq 1, \quad (\text{A.5.180})$$

and that

$$\|Z(\mathbf{h}_{\delta,\tilde{\delta}})_k - Z(\mathbf{f})_k\| \geq \left( \frac{1}{2}\sqrt{(1-\tilde{\delta})\mathfrak{D}_z(f_k;n)} \vee \sum_{k=1}^s \left( \rho_z(\varepsilon; f_k) \left( 1 \wedge \sqrt{2n\rho_z(\varepsilon; f_k)} \right) - \delta \right) \right), \quad (\text{A.5.181})$$

for  $k \in \{1, 2, \dots, s\}$ .

Therefore, we have for  $CI_{m,\alpha} \in \mathcal{I}_{m,\alpha,n}(\mathcal{F}_s)$ ,

$$\begin{aligned} \mathbb{E}_{\mathbf{f}}(V(CI_{m,\alpha})) &\geq (1 - \alpha - \Phi(-z_\alpha + 1)) \times \\ &\quad \Pi_{k=1}^s \left( \frac{1}{2}\sqrt{(1-\tilde{\delta})\mathfrak{D}_z(f_k;n)} \vee \sum_{k=1}^s \left( \rho_z(\varepsilon; f_k) \left( 1 \wedge \sqrt{2n\rho_z(\varepsilon; f_k)} \right) - \delta \right) \right). \end{aligned} \quad (\text{A.5.182})$$

Note that  $\alpha \leq 0.3$  gives  $1 - \alpha - \Phi(-z_\alpha + 1) > 0$ .

Take  $\delta, \tilde{\delta} \rightarrow 0^+$ , we have

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}}(V(CI_{m,\alpha})) \\
& \geq (1 - \alpha - \Phi(-z_\alpha + 1)) \Pi_{k=1}^s \left( \frac{1}{2} \sqrt{\mathfrak{D}_z(f_k; n)} \vee \left( \rho_z(\varepsilon; f_k) \left( 1 \wedge \sqrt{2n\rho_z(\varepsilon; f_k)} \right) \right) \right) \\
& \geq (1 - \alpha - \Phi(-z_\alpha + 1)) \Pi_{k=1}^s \left( \frac{1}{2} \sqrt{\mathfrak{D}_z(f_k; n)} \vee \frac{1}{\sqrt{12s}} \varphi_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \right) \\
& \leq (1 - \alpha - \Phi(-z_\alpha + 1)) (12s)^{-\frac{s}{2}} \Pi_{k=1}^s \left( \sqrt{\mathfrak{D}_z(f_k; n)} \vee \varphi_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \right).
\end{aligned} \tag{A.5.183}$$

**Proof of Inequality (A.5.157) and Inequality (A.5.158)**

Let

$$\varepsilon_k = \frac{\varphi_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}{\sqrt{\sum_{i=1}^s \varphi_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_i\right)^2}} \frac{1}{\sqrt{6}} \frac{\sigma}{(n+1)^{\frac{s}{2}}} \frac{1}{1 + \frac{s}{n} + \sum_{i=1}^s 2\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_i\right)}. \tag{A.5.184}$$

Recall Lemma A.5.4. Let  $\delta = \frac{0.1}{s} \cdot \min_{1 \leq k \leq s} \varphi_m(\varepsilon_k; f_k)$ . For each  $k \in \{1, 2, \dots, s\}$ , take  $\varepsilon = \varepsilon_k$ , and take let  $g_{k,\delta}$  be the function  $g$  in Lemma A.5.4.

Let  $\tilde{\delta} < 0.01$  be a small positive number.

Take  $f_{k,alt,\tilde{\delta}} \in \mathcal{F}$  satisfying

$$\begin{aligned}
f_{k,alt,\tilde{\delta}}\left(\frac{i}{n}\right) &= f_k\left(\frac{i}{n}\right) \text{ for } 0 \leq i \leq n, \\
|M(f_{k,alt,\tilde{\delta}}) - M(f_k)| &\geq \frac{1}{2}(1 - \tilde{\delta})\mathfrak{D}_m(f_k; n).
\end{aligned} \tag{A.5.185}$$

Let

$$\mathbf{g}_\delta(\mathbf{t}) = f_0 + \sum_{k=1}^s g_{k,\delta}(t_k). \tag{A.5.186}$$

Clearly  $\mathbf{g}_\delta \in \mathcal{F}_s$ .

With a bit abuse of notation, in this proof let

$$\Delta_k = \frac{1}{n+1} \sum_{i=0}^n g_{k,\delta}\left(\frac{i}{n}\right) - f_k\left(\frac{i}{n}\right) \quad (\text{A.5.187})$$

Then we have that

$$\begin{aligned} & \frac{\sqrt{\sum_{\mathbf{i} \in \{0,1,\dots,n\}^s} (\mathbf{f}(\mathbf{i}) - \mathbf{g}_\delta(\mathbf{i}))^2 / (n+1)^s}}{\sigma / (n+1)^{\frac{s}{2}}} \\ &= \frac{\sqrt{(\sum_{k=1}^s \Delta_k)^2 + \sum_{k=1}^s l_n(f_k, g_{k,\delta}) - \Delta_k)^2}}{\sigma / (n+1)^{\frac{s}{2}}} \\ &\leq \frac{\sqrt{\sum_{k=1}^s l_n(f_k, g_{k,\delta}) - \Delta_k)^2} \sqrt{1 + \frac{s}{n} + \sum_{k=1}^s 2\rho_z(\varepsilon_i; f_k)}}{\sigma / (n+1)^{\frac{s}{2}}} \quad (\text{A.5.188}) \\ &\leq \frac{\sqrt{\sum_{k=1}^s 6\varepsilon_k^2} \sqrt{1 + \frac{s}{n} + \sum_{k=1}^s 2\rho_z(\varepsilon; f_k)}}{\sigma / (n+1)^{\frac{s}{2}}} \\ &\leq 1 \end{aligned}$$

Also, by Lemma A.5.4, we have that

$$\begin{aligned} M(\mathbf{g}_\delta) - M(\mathbf{f}) &= \sum_{k=1}^s M(g_{k,\delta}) - M(f_k) \geq \sum_{k=1}^s \rho_m(\varepsilon_k; f)(1 \wedge \sqrt{2n\rho_z(\varepsilon_k; f)}) - \delta \\ &\geq \sum_{k=1}^s \sqrt{\frac{1}{3}} \frac{\varepsilon_k}{\sigma / (n+1)^{\frac{s}{2}}} \varphi_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) - \delta \\ &\geq \frac{1}{3\sqrt{2}} \sqrt{\sum_{k=1}^s \varphi_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2} \sqrt{\frac{1}{1 + \frac{s}{n} + \sum_{k=1}^s 2\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}} - s\delta. \end{aligned} \quad (\text{A.5.189})$$

Recall the sufficient statistic  $W$  given in (A.5.171).

So we have

$$\begin{aligned}\tilde{\mathbf{R}}_{m,n}(\sigma; \mathbf{f}) &\geq \inf_{\hat{M}} \max\{\mathbb{E}_{\mathbf{f}}(|\hat{M} - M(\mathbf{f})|^2), \mathbb{E}_{\mathbf{g}_\delta}(|\hat{M} - M(\mathbf{g}_\delta)|^2)\} \\ &\geq r_2 |M(\mathbf{f}) - M(\mathbf{g}_\delta)|^2,\end{aligned}\tag{A.5.190}$$

where

$$r_2 = \inf_{\hat{\theta}} \max_{\theta=\pm 1} \mathbb{E}_{\theta} \frac{|\hat{\theta} - \theta|^2}{4},$$

for  $W \sim N(\frac{\theta}{2}, 1)$ . Elementary calculation shows that  $r_2 > 0.1$ .

Let  $\delta \rightarrow 0^+$ , so we have

$$\tilde{\mathbf{R}}_{m,n}(\sigma; \mathbf{f}) \geq \frac{1}{180} \sum_{k=1}^s \varphi_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2 \frac{1}{1 + \frac{s}{n} + \sum_{k=1}^s 2\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}.\tag{A.5.191}$$

It's apparent that  $\tilde{\mathbf{R}}_{m,n}(\sigma; \mathbf{f}) \geq \frac{1}{4} \mathfrak{D}_m(\mathbf{f}; n)^2$ . This concludes the proof of Inequality (A.5.157).

We now turn to the proof of Inequality (A.5.158) .

Let  $\tilde{\delta} < 0.01$  be a small positive number. Then there exist  $\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2 \in \mathcal{F}_s$  such that

$$\tilde{\mathbf{f}}_1\left(\frac{\mathbf{i}}{n}\right) = \mathbf{f}\left(\frac{\mathbf{i}}{n}\right) = \tilde{\mathbf{f}}_2\left(\frac{\mathbf{i}}{n}\right) \text{ for } \mathbf{i} \in \{0, 1, \dots, n\}^s, \quad |M(\tilde{\mathbf{f}}_1) - M(\tilde{\mathbf{f}}_2)| \geq (1 - \tilde{\delta}) \mathfrak{D}_m(\mathbf{f}; n),\tag{A.5.192}$$

Suppose  $CI_{m,\alpha} \in \mathcal{I}_{m,\alpha,n}(\mathcal{F}_s)$ .

It's clear that  $CI_{m,\alpha} \in \mathcal{I}_{m,\alpha,n}(\{\mathbf{f}, \mathbf{g}_\delta\})$ ,  $CI_{m,\alpha} \in \mathcal{I}_{m,\alpha,n}(\{\tilde{\mathbf{f}}_2, \tilde{\mathbf{f}}_1\})$ . Therefore, we have that

$$\tilde{\mathbf{L}}_{m,\alpha,n}(\sigma; \mathbf{f}) \geq (1 - 2\alpha) \cdot (1 - \tilde{\delta}) \mathfrak{D}_m(\mathbf{f}; n),\tag{A.5.193}$$

and that

$$\begin{aligned}
& \tilde{\mathbf{L}}_{m,\alpha,n}(\sigma; \mathbf{f}) \\
& \geq (1 - \alpha - \Phi(-z_\alpha + 1)) \cdot |M(\mathbf{f}) - M(\mathbf{g}_\delta)| \\
& \geq (1 - \alpha - \Phi(-z_\alpha + 1)) \cdot \frac{1}{3\sqrt{2}} \sqrt{\sum_{k=1}^s \varphi_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2} \sqrt{\frac{1}{1 + \frac{s}{n} + \sum_{k=1}^s 2\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}} \\
& \quad - s\delta.
\end{aligned} \tag{A.5.194}$$

Letting  $\delta, \tilde{\delta} \rightarrow 0^+$  gives Inequality (A.5.158).

#### A.5.13. Proof of Proposition 3.4.1

The idea of the proof is very similar to that for white noise model.

Invertibility follows from definition. Independence follows from the observation that the concatenation of the elements is this  $s+1$  tuple  $\mathfrak{P}(\{y_{\mathbf{i}}\})$  follows a joint normal distribution and that covariance of elements from different places of the tuple is 0. The sufficiency rises from factorization of the probability.

#### A.5.14. Proof of Theorem 3.4.1

We have

$$\mathbb{E}_{\mathbf{f}} \left( \|\hat{Z} - Z(\mathbf{f})\|^2 \right) \leq \sum_{k=1}^s \mathbb{E}_{\mathbf{f}} \left( \|\hat{Z}_k - Z(f_k)\|^2 \right). \tag{A.5.195}$$

Note that Proposition 2.2.1 gives

$$\rho_z\left((z_\zeta + 1) \frac{\sqrt{6}\sigma}{\sqrt{n}(n+1)^{\frac{s-1}{2}}}; f_k\right) \leq \left(3 \times 4\sqrt{3}\right)^{\frac{2}{3}} \rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \tag{A.5.196}$$

for  $\zeta \leq \Phi(-2)$ . Also note that  $\mathfrak{D}_z(\mathbf{f}; n) = \sum_{k=1}^s \mathfrak{D}_z(f_k; n)$ .

Recall the lower bound for  $\tilde{\mathbf{R}}_{z,n}(\sigma; \mathbf{f})$  given in Inequality (A.5.155).

So it is sufficient to prove that for  $\zeta \leq 0.15$  the following holds

$$\begin{aligned} \mathbb{E}_{\mathbf{f}} \left( \|\hat{Z}_k - Z(f_k)\|^2 \right) \leq \\ \check{C}_2 \rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{\sqrt{n}(n+1)^{\frac{s-1}{2}}}; f_k)^2 \sqrt{n \rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{\sqrt{n}(n+1)^{\frac{s-1}{2}}}; f_k) \vee 1 + 2\mathfrak{D}_z(f_k; n)}, \end{aligned} \quad (\text{A.5.197})$$

for an absolute constant  $\check{C}_2 > 0$ .

Now we proceed with proving it.

First we introduce a quantity for a general  $\zeta > 0$ :

$$\begin{aligned} \xi_k(\zeta) = \sup \left\{ \xi : \min \left\{ \sqrt{\xi} [f_k(Z(f_k) + \xi) - M(f_k)], \right. \right. \\ \left. \left. \sqrt{\xi} [f_k(Z(f_k) - \xi) - M(f_k)] \right\} \times \frac{\sqrt{n}}{\sqrt{6}\sigma/(n+1)^{\frac{s-1}{2}}} \leq z_\zeta + 1 \right\}. \end{aligned} \quad (\text{A.5.198})$$

Then let

$$\mathbf{j}_k(\zeta) = \max \left\{ j : \frac{2^{J-j}}{n} > \xi_k(\zeta) \right\}. \quad (\text{A.5.199})$$

We further introduce the following quantities.

$$\begin{aligned} \mathbf{i}_{k,j}^* &= \max \left\{ i : Z(f_k) \in \left[ \frac{2^{J-j} \cdot (i-1)}{n} - \frac{1}{2n}, \frac{2^{J-j} \cdot i}{n} - \frac{1}{2n} \right] \right\} \\ \tilde{\mathbf{j}}_k &= \min \left( \{j : |\hat{\mathbf{i}}_{k,j} - \mathbf{i}_{k,j}^*| \geq 2\} \cup \infty \right), \\ \acute{\mathbf{j}}_k &= \min \left( \{j : |\hat{\mathbf{i}}_{k,j} - \mathbf{i}_{k,j}^*| \geq 5\} \cup \infty \right), \\ \grave{\mathbf{j}}_k &= \min \left( \{j : |\hat{\mathbf{i}}_{k,j} - \mathbf{i}_{k,j}^*| \geq 7\} \cup \infty \right). \end{aligned} \quad (\text{A.5.200})$$

Then we immediately have the following facts that we summarize into a lemma.

**Lemma A.5.5.** For  $j \leq \min\{J, \mathfrak{j}_k(\zeta)\}$ , we have

$$\frac{1}{\tilde{\sigma}_{k,j}} \sum_{h=(\mathfrak{i}_{k,j}^*+1)2^{J-j}}^{(\mathfrak{i}_{k,j}^*+2)2^{J-j}-1} \left( f_k\left(\frac{h}{n}\right) - f_k\left(\frac{h-2^{J-j}}{n}\right) \right) \geq 2^{\frac{3}{2}(\mathfrak{j}_k(\zeta)-j)} (z_\zeta + 1), \quad (\text{A.5.201})$$

and

$$\frac{1}{\tilde{\sigma}_{k,j}} \sum_{h=(\mathfrak{i}_{k,j}^*-2)2^{J-j}}^{(\mathfrak{i}_{k,j}^*-1)2^{J-j}-1} \left( f_k\left(\frac{h-2^{J-j}}{n}\right) - f_k\left(\frac{h}{n}\right) \right) \geq 2^{\frac{3}{2}(\mathfrak{j}_k(\zeta)-j)} (z_\zeta + 1). \quad (\text{A.5.202})$$

When  $\tilde{\mathfrak{j}}_k = j$ , then one of the following happens

$$\begin{aligned} Y_{k,j,\mathfrak{i}_{k,j}^*+2}^l &\leq Y_{k,j,\mathfrak{i}_{k,j}^*+1}^l, Y_{k,j,\mathfrak{i}_{k,j}^*+3}^l \leq Y_{k,j,\mathfrak{i}_{k,j}^*+1}^l, Y_{k,j,\mathfrak{i}_{k,j}^*+4}^l \leq Y_{k,j,\mathfrak{i}_{k,j}^*+1}^l, \\ Y_{k,j,\mathfrak{i}_{k,j}^*-2}^l &\leq Y_{k,j,\mathfrak{i}_{k,j}^*-1}^l, Y_{k,j,\mathfrak{i}_{k,j}^*-3}^l \leq Y_{k,j,\mathfrak{i}_{k,j}^*-1}^l, Y_{k,j,\mathfrak{i}_{k,j}^*-4}^l \leq Y_{k,j,\mathfrak{i}_{k,j}^*-1}^l. \end{aligned} \quad (\text{A.5.203})$$

Now we will state three lemmas, the proofs of which are left to latter parts.

**Lemma A.5.6.** Suppose  $\zeta \leq 0.5$ .

$$\mathbb{E}_{\mathbf{f}} \left( 2^{-2\tilde{\mathfrak{j}}_k} \mathbb{1}_{\{\tilde{\mathfrak{j}}_k \leq J\}} \right) \leq \check{C}_0 2^{-2\mathfrak{j}_k(\zeta)} \left( 1 \wedge 2^{J-\mathfrak{j}_k(\zeta)} \right), \quad (\text{A.5.204})$$

where  $\check{C}_0 = \max\{\sup_{x \geq 1} 2x^2 \Phi(-x), 2\}$ .

*Remark A.5.1.* Note that the left hand side of Inequality (A.5.204) does not depend on  $\zeta$ , but we state this more general lemma.

**Lemma A.5.7.** Suppose  $\zeta \leq 0.5$ .

$$\mathbb{E}_{\mathbf{f}} \left( 2^{-2\check{\mathfrak{j}}_k(\zeta)} \mathbb{1}_{\{\check{\mathfrak{j}}_k(\zeta) < \infty\}} \mathbb{1}_{\{\tilde{\mathfrak{j}}_k > \check{\mathfrak{j}}_k(\zeta)\}} \right) \leq \check{C}_0 2^{-2\mathfrak{j}_k(\zeta)} \left( 1 \wedge 2^{J-\mathfrak{j}_k(\zeta)} \right), \quad (\text{A.5.205})$$

where  $\check{C}_0 = \max\{\sup_{x \geq 1} 2x^2 \Phi(-x), 2\}$ .



**Lemma A.5.8.** *Suppose  $\zeta \leq 0.5$ .*

$$\begin{aligned} & \mathbb{E}_{\mathbf{f}} \left( |\hat{Z}_k - Z(f_k)|^2 \mathbb{1}\{\check{\mathbf{j}}_k(\zeta) = \infty, \tilde{\mathbf{j}}_k > J\} \right) \\ & \leq 64 \cdot 2^{-2\mathbf{j}_k(\zeta)} \left( 1 \wedge 2^{J-\mathbf{j}_k(\zeta)} \right) + 2\mathfrak{D}_z(f_k; n). \end{aligned} \quad (\text{A.5.206})$$

With these lemmas, we have that

$$\mathbb{E}_{\mathbf{f}} \left( |\hat{Z}_k - Z(f_k)|^2 \right) \leq \check{C}_1 \cdot 2^{-2\mathbf{j}_k(\zeta)} \left( 1 \wedge 2^{J-\mathbf{j}_k(\zeta)} \right) + 2\mathfrak{D}_z(f_k; n), \quad (\text{A.5.207})$$

where  $\check{C}_1 = 64 + 2\check{C}_0$ .

Now we introduce the following lemma about  $\xi_k(\zeta)$  and  $\mathbf{j}_k(\zeta)$ , which immediately concludes the proof of Theorem 3.4.1.

**Lemma A.5.9.** *For  $\zeta > 0$ , we have*

$$2\rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k) \geq \xi_k(\zeta) \geq \frac{1}{2} \rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k). \quad (\text{A.5.208})$$

$$\frac{n+2}{2} \leq 2^J \leq n+1. \quad (\text{A.5.209})$$

$$2^{-\mathbf{j}_k(\zeta)} \leq \frac{2n}{2^J} \xi_k(\zeta) \leq 8\rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k). \quad (\text{A.5.210})$$

### Proof of Lemma A.5.6

A basic property of normal tail bound is that  $\frac{\Phi(-2\sqrt{2}x)}{\Phi(-x)}$  decreases with  $x > 0$  increasing.

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}} \left( 2^{-2\check{\mathbf{j}}_k} \mathbb{1}\{\check{\mathbf{j}}_k \leq J\} \right) \\
& \leq \sum_{j=1}^J 2^{-2\mathbf{j}_k(\zeta)} \cdot 2^{-2j+2\mathbf{j}_k(\zeta)} \left( \Phi(-2^{\frac{3}{2}(\mathbf{j}_k(\zeta)-j)}(z_\zeta + 1)) \mathbb{1}\{j \leq \mathbf{j}_k(\zeta)\} + \mathbb{1}\{j > \mathbf{j}_k(\zeta)\} \right) \\
& \leq \mathbb{1}\{J \leq \mathbf{j}_k(\zeta)\} 2^{-2\mathbf{j}_k(\zeta)} \cdot 2^{-2J+2\mathbf{j}_k(\zeta)} \Phi(-2^{\frac{3}{2}(\mathbf{j}_k(\zeta)-J)}(z_\zeta + 1)) \frac{1}{1 - 4^{\frac{\Phi(-2\sqrt{2})}{\Phi(-1)}}} \\
& \quad + \mathbb{1}\{J > \mathbf{j}_k(\zeta)\} 2^{-2\mathbf{j}_k(\zeta)} \left( \frac{1}{1 - 4^{\frac{\Phi(-2\sqrt{2})}{\Phi(-1)}}} + \frac{1}{3} \right) \\
& \leq \mathbb{1}\{J \leq \mathbf{j}_k(\zeta)\} 2^{-2\mathbf{j}_k(\zeta)} \cdot 2^{J-\mathbf{j}_k(\zeta)} \sup_{x \geq 1} 2x^2 \Phi(-x) + 2 \cdot \mathbb{1}\{J > \mathbf{j}_k(\zeta)\} 2^{-2\mathbf{j}_k(\zeta)}
\end{aligned} \tag{A.5.211}$$

Let  $\check{C}_0 = \max\{\sup_{x \geq 1} 2x^2 \Phi(-x), 2\}$ , then we have the lemma.

#### **Proof of Lemma A.5.7**

By our stopping rule, apparently  $\check{\mathbf{j}}_k(\zeta) \geq 1$ .

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}} \left( 2^{-2\check{\mathbf{j}}_k(\zeta)} \mathbb{1}\{\check{\mathbf{j}}_k(\zeta) < \infty\} \mathbb{1}\{\tilde{\mathbf{j}}_k > \check{\mathbf{j}}_k(\zeta)\} \right) \\
&= \sum_{j=1}^J 2^{-2j} \mathbb{E}_{\mathbf{f}} \left( \mathbb{E}_{\mathbf{f}} \left( \mathbb{1}\{\tilde{\mathbf{j}}_k > \check{\mathbf{j}}_k(\zeta) = j\} | \nu_{k,i}^l \right) \right) \\
&\leq \sum_{j=1}^J 2^{-2\mathbf{j}_k(\zeta)} \cdot 2^{-2j+2\mathbf{j}_k(\zeta)} \left( \Phi(-2^{\frac{3}{2}(\mathbf{j}_k(\zeta)-j)}(z_{\zeta} + 1)) \mathbb{1}\{j \leq \mathbf{j}_k(\zeta)\} + \mathbb{1}\{j > \mathbf{j}_k(\zeta)\} \right) \\
&\leq \mathbb{1}\{J \leq \mathbf{j}_k(\zeta)\} 2^{-2\mathbf{j}_k(\zeta)} \cdot 2^{-2J+2\mathbf{j}_k(\zeta)} \Phi(-2^{\frac{3}{2}(\mathbf{j}_k(\zeta)-J)}(z_{\zeta} + 1)) \frac{1}{1 - 4^{\frac{\Phi(-2\sqrt{2})}{\Phi(-1)}}} \\
&\quad + \mathbb{1}\{J > \mathbf{j}_k(\zeta)\} 2^{-2\mathbf{j}_k(\zeta)} \left( \frac{1}{1 - 4^{\frac{\Phi(-2\sqrt{2})}{\Phi(-1)}}} + \frac{1}{3} \right) \\
&\leq \mathbb{1}\{J \leq \mathbf{j}_k(\zeta)\} 2^{-2\mathbf{j}_k(\zeta)} \cdot 2^{J-\mathbf{j}_k(\zeta)} \sup_{x \geq 1} 2x^2 \Phi(-x) + 2\mathbb{1}\{J > \mathbf{j}_k(\zeta)\} 2^{-2\mathbf{j}_k(\zeta)}
\end{aligned} \tag{A.5.212}$$

Let  $\check{C}_0 = \max\{\sup_{x \geq 1} 2x^2 \Phi(-x), 2\}$ , then we have the lemma.

### Proof of Lemma A.5.8

Note that  $\check{\mathbf{j}}_k(\zeta) = \infty, \tilde{\mathbf{j}}_k > J$  means that

$$\{i : f_k(\frac{i}{n}) = \min_{l \in \{0,1,\dots,n\}}\} \subset \{\hat{\mathbf{i}}_{k,J} - 3, \hat{\mathbf{i}}_{k,J} - 2, \hat{\mathbf{i}}_{k,J} - 1, \hat{\mathbf{i}}_{k,J}, \hat{\mathbf{i}}_{k,J} + 1\}, \tag{A.5.213}$$

and that

$$Z(f_k) \in [\frac{\hat{\mathbf{i}}_{k,J} - 3}{n}, \frac{\hat{\mathbf{i}}_{k,J} + 1}{n}]. \tag{A.5.214}$$

When  $\mathbf{j}_k(\zeta) \leq J$ , then we have  $2^{-\mathbf{j}_k(\zeta)} \geq 2^{-J} \geq \frac{1}{n+1}$ .

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}} \left( |\hat{Z}_k - Z(f_k)|^2 \mathbb{1}\{\check{\mathbf{j}}_k(\zeta) = \infty, \tilde{\mathbf{j}}_k > J\} \right) \leq \frac{16}{n^2} \\
& \leq 16 \left( \frac{n+1}{n} \right)^2 2^{-2\mathbf{j}_k(\zeta)} \left( 1 \wedge 2^{J-\mathbf{j}_k(\zeta)} \right) \leq 64 \cdot 2^{-2\mathbf{j}_k(\zeta)} \left( 1 \wedge 2^{J-\mathbf{j}_k(\zeta)} \right).
\end{aligned} \tag{A.5.215}$$

When  $j_k(\zeta) \geq J + 1$ , denote  $i_m = \arg \min_{i: f_k(\frac{i}{n}) = \min_{l \in \{0, 1, \dots, n\}} |\frac{i}{n} - \hat{Z}_k|}$ , the index of the position at which  $f_k$  is minimized while being closest to the estimator. Note that this is deterministic when  $f_k$  has unique minimizer among grid points but is a random variable when  $f_k$  has two minimizers among grid points.

Then according to Lemma A.5.5 we know that

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}} \left( |\hat{Z}_k - Z(f_k)|^2 \mathbb{1}\{\check{j}_k(\zeta) = \infty, \tilde{j}_k > J\} \right) \\
& \leq 2\mathbb{E}_{\mathbf{f}} \left( |\hat{Z}_k - \frac{i_m}{n}|^2 \right) + 2\mathfrak{D}_z(f_k; n) \\
& \leq 2 \times \frac{16}{n^2} \times 4\Phi(-2^{\frac{3}{2}(j_k(\zeta)-J)}(z_\zeta + 1)) + 2\mathfrak{D}_z(f_k; n) \\
& \leq 128 \left( \frac{n+1}{n} \right)^2 2^{-2J} \Phi(-2^{\frac{3}{2}(j_k(\zeta)-J)}) + 2\mathfrak{D}_z(f_k; n) \\
& \leq 128 \left( \frac{n+1}{n} \right)^2 2^{-2j_k(\zeta)} \cdot 2^{J-j_k(\zeta)} \cdot 2^3 \Phi(-\sqrt{8}) + 2\mathfrak{D}_z(f_k; n) \\
& < 10 \cdot 2^{-2j_k(\zeta)} \cdot 2^{J-j_k(\zeta)} + 2\mathfrak{D}_z(f_k; n)
\end{aligned} \tag{A.5.216}$$

Hence we concludes the proof.

### Proof of Lemma A.5.9

Denote

$$\Delta_{1,k} = \frac{1}{2} \rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k),$$

and

$$\Delta_{2,k} = \min\{f_k(Z(f_k) + \Delta_{1,k}), f_k(Z(f_k) - \Delta_{1,k})\} - M(f_k).$$

Then we have that

$$\begin{aligned}
& \Delta_{1,k} \Delta_{2,k}^2 \\
& \leq \|f_k - \max\{f_k, M(f_k) + \rho_m((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k)\}\|^2 \\
& = \left( (z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}} \right)^2.
\end{aligned} \tag{A.5.217}$$

Denote

$$\Delta_{3,k} = 2\rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k),$$

and

$$\Delta_{4,k} = \min\{f_k(Z(f_k) + \Delta_{3,k}), f_k(Z(f_k) - \Delta_{3,k})\} - M(f_k).$$

Clearly that

$$\Delta_{4,k} \geq \rho_m((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k).$$

Then we have that

$$\begin{aligned}
& \Delta_{3,k} \Delta_{4,k}^2 \\
& \geq \|f_k - \max\{f_k, M(f_k) + \rho_m((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k)\}\|^2 \\
& = \left( (z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}} \right)^2.
\end{aligned} \tag{A.5.218}$$

#### A.5.15. Proof of Theorem 3.4.2

Note that the coordinates of the hyper cube  $CI_{z,\alpha}$  are independence from each other, so the following two propositions are sufficient to give the statement of the theorem.

**Proposition A.5.11.** *For  $CI_{k,\alpha}$  defined in (3.4.14)*

$$\mathbb{E}_{\mathbf{f}}(\mathbb{1}\{Z(f_k) \notin CI_{k,\alpha}\}) \leq \alpha/s, \tag{A.5.219}$$

for all  $\mathbf{f} \in \mathcal{F}_s$

**Proposition A.5.12.** For  $CI_{k,\alpha}$  defined in (3.4.14)

$$\mathbb{E}_{\mathbf{f}}(|t_{k,hi} - t_{k,lo}|^2) \leq C_5 \rho_z(z_{\alpha/s} \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k)^2 \left(1 \wedge n \rho_z(z_{\alpha/s} \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k)\right) + 9 \mathfrak{D}_z(f_k; n), \quad (\text{A.5.220})$$

for all  $\mathbf{f} \in \mathcal{F}_s$ , for an absolute positive constant  $C_5$ .

The reason Proposition A.5.12 implies the statement of expected volume in Theorem 3.4.2 is as follows. Proposition (A.5.12) implies that

$$\mathbb{E}_{\mathbf{f}}(|t_{k,hi} - t_{k,lo}|) \leq \sqrt{C_5} \cdot (2z_{\alpha/s}) \cdot \varphi_z(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k) + 3\sqrt{\mathfrak{D}_z(f_k; n)}, \quad (\text{A.5.221})$$

where  $\varphi_z(\cdot, \cdot)$  is defined in Equation (A.5.153). This further gives that

$$\mathbb{E}_{\mathbf{f}}(V(CI_{z,\alpha})) \leq \left(3 + \sqrt{C_5} \cdot (2z_{\alpha/s})\right)^s \Pi_{k=1}^s \left(\varphi_z(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k) \vee \sqrt{\mathfrak{D}_z(f_k; n)}\right). \quad (\text{A.5.222})$$

This combined with the lower bound for  $\tilde{\mathcal{L}}_{z,\alpha,n}(\sigma; \mathbf{f})$  given in (A.5.156) gives the statement about expected volume.

Before we continue with the proofs of the propositions, recall the quantities we defined in Equation (A.5.200) and (A.5.199).

And we further introduce the following quantities that will be used frequently

$$i_{m,l} = \min\{i : f(\frac{i}{n}) = \min_{h \in \{0,1,\dots,n\}} f(\frac{h}{n})\}, i_{m,r} = \max\{i : f(\frac{i}{n}) = \min_{h \in \{0,1,\dots,n\}} f(\frac{h}{n})\}. \quad (\text{A.5.223})$$

On the event  $\{\check{\mathbf{j}}_k(\alpha/2s) = \infty\}$ , we define a “bad” event. Let the event that first shrinking step misses the target be

$$B_1 = \{i_l \geq i_{m,l} + 1\} \cup \{i_r \leq i_{m,2} - 2\}. \quad (\text{A.5.224})$$

We will define more “bad” events in the proofs of the propositions, usually denoted by  $B_h$  for  $h = 2, 3, 4, \dots$ .

On the event  $\{\check{\mathbf{j}}_k(\alpha/2s) = \infty\}$ , from our definition, it is clear that  $i_l \leq i_r + 1$ .

We recollect the quantities defined in Equations (A.5.200), (A.5.199).

### Proof of Proposition A.5.11

The event that  $\{Z(f_k) \notin CI_{k,\alpha}\}$  can be partitioned into the followings

$$\begin{aligned} \{Z(f_k) \notin CI_{k,\alpha}\} &\subset \{\check{\mathbf{j}}_k \leq \hat{\mathbf{j}}_k(\alpha/2s) - 1\} \\ &\cup (\{\check{\mathbf{j}}_k \geq \hat{\mathbf{j}}_k(\alpha/2s), \check{\mathbf{j}}_k(\alpha/2s) = \infty\} \cap B_1) \\ &\cup ((\{\check{\mathbf{j}}_k \geq \hat{\mathbf{j}}_k(\alpha/2s), \check{\mathbf{j}}_k(\alpha/2s) = \infty\} \cap B_1^c) \cap \{Z(f_k) \notin CI_{k,\alpha}\}). \end{aligned} \quad (\text{A.5.225})$$

We will bound them separately.

$$\mathbb{E}_{\mathbf{f}}(\mathbb{1}\{\check{\mathbf{j}}_k \leq \hat{\mathbf{j}}_k(\alpha/2s) - 1\}) \leq \mathbb{E}_{\mathbf{f}}\left(\left(\mathbb{1}\{\mathbf{T}_{k,\check{\mathbf{j}}_k} \geq \tilde{\sigma}_{k,\check{\mathbf{j}}_k}(z_{\alpha/2s})\} \middle| \nu_{k,\cdot}^l\right)\right) \leq \alpha/2s. \quad (\text{A.5.226})$$

On event  $\{\check{\mathbf{j}}_k \geq \hat{\mathbf{j}}_k(\alpha/2s), \check{\mathbf{j}}_k(\alpha/2s) = \infty\}$ , we know that  $L_k \leq i_{m,l} \leq i_{m,r} \leq U_k$ . Therefore, we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{f}}(\{\check{\mathbf{j}}_k \geq \hat{\mathbf{j}}_k(\alpha/2s), \check{\mathbf{j}}_k(\alpha/2s) = \infty\} \cap B_1) \\ &\leq P(\nu_{k,i_{m,l}}^e - \nu_{k,i_{m,l}+1}^e + \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (z_{k,i_{m,l}}^3 - z_{k,i_{m,l}+1}^3) > 2\sqrt{3}\frac{\sigma}{(n+1)^{\frac{s-1}{2}}} z_{\alpha_1}) \\ &\quad + P(\nu_{k,i_{m,r}-1}^e - \nu_{k,i_{m,r}}^e + \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (z_{k,i_{m,r}-1}^3 - z_{k,i_{m,r}}^3) < -\frac{2\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} z_{\alpha_1}) \\ &\leq 2\alpha_1 \leq \alpha/4s. \end{aligned} \quad (\text{A.5.227})$$

On the event  $\{\check{\mathbf{j}}_k \geq \hat{\mathbf{j}}_k(\alpha/2s), \check{\mathbf{j}}_k(\alpha/2s) = \infty\} \cap B_1^c$ , we know that only when  $i_l = i_r + 1 \leq n -$

$1, t_{k,hi} < \min\{\frac{i_{m,r}+1}{n}, 1\}$  could happen, and only when  $i_l = i_r + 1 \geq 1, t_{k,lo} > \max\{\frac{i_{m,l}-1}{n}, 0\}$  could happen. And note that  $i_{m,r} \leq i_l = i_r + 1 \leq i_{m,l}$  indicates that  $i_{m,l} = i_{m,r}$ , which we denote as  $i_m$ . So in the following we only consider  $f_k$  with unique minimizer on grids. Also we have in these cases  $i_l = i_m$ . We have that

$$\begin{aligned} & P_{\mathbf{f}} \left( (\{\hat{\mathbf{j}}_k \geq \hat{\mathbf{j}}_k(\alpha/2s), \check{\mathbf{j}}_k(\alpha/2s) = \infty\} \cap B_1^c) \cap \{Z(f_k) \notin CI_{k,\alpha}\} \right) \\ & \leq \mathbb{E}_{\mathbf{f}} (\mathbb{1}\{i_m = i_l = i_r + 1 \leq n-1, t_{k,hi} < Z(f_k)\}) \\ & \quad + \mathbb{E}_{\mathbf{f}} (\mathbb{1}\{i_m = i_l = i_r + 1 \geq 1, t_{k,lo} > Z(f_k)\}). \end{aligned} \quad (\text{A.5.228})$$

The arguments bounding the two terms are similar, so we only show that for the first one.

Use  $t_{k,r}$  to denote the intersection between the two lines

$$l_1 : y = f(\frac{i_m}{n}), l_2 : y(t) = f(\frac{i_m+1}{n}) + \frac{f(\frac{i_m+2}{n}) - f(\frac{i_m+1}{n})}{1/n} (t - \frac{i_m+1}{n}). \quad (\text{A.5.229})$$

It is clear that  $Z(f_k) \leq t_{k,r}$ .

Basic calculation shows that

$$t_{k,r} = \frac{f_k(\frac{i_m}{n}) - f_k(\frac{i_m+1}{n})}{n(f_k(\frac{i_m+2}{n}) - f_k(\frac{i_m+1}{n}))} + \frac{i_m+1}{n}. \quad (\text{A.5.230})$$

It is easy to check that the distribution of

$$\begin{aligned} & \left( \nu_{k,i_m}^e - \nu_{k,i_m+1}^e - \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} \left( z_{k,i_m}^3 - z_{k,i_m+1}^3 - 2\sqrt{2}z_{\alpha_2} \right), \right. \\ & \left. \nu_{k,i_m+2}^e - \nu_{k,i_m+1}^e - \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} \left( z_{k,i_m+2}^3 - z_{k,i_m+1}^3 - 2\sqrt{2}z_{\alpha_2} \right) \right) \end{aligned} \quad (\text{A.5.231})$$



is the same with the following

$$\begin{aligned} & \left( f_k\left(\frac{i_m}{n}\right) + \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}} \cdot \eta_0 - f_k\left(\frac{i_m+1}{n}\right) - \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}} \cdot \eta_1 + \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}} \cdot 2z_{\alpha_2}, \right. \\ & \left. f_k\left(\frac{i_m+2}{n}\right) + \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}} \cdot \eta_2 - f_k\left(\frac{i_m+1}{n}\right) - \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}} \cdot \eta_1 + \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}} \cdot 2z_{\alpha_2} \right), \end{aligned} \quad (\text{A.5.232})$$

where  $\eta_0, \eta_1, \eta_2 \stackrel{i.i.d}{\sim} N(0, 1)$  and also independent from  $i_l, i_r$ .

Note that under the event

$$\{\eta_0 \geq -z_{\alpha_2}, \eta_1 \leq z_{\alpha_2}, \eta_2 \geq -z_{\alpha_2}\},$$

we have  $t_{k,hi} \geq t_{k,r}$ . Hence we have that

$$\begin{aligned} & \mathbb{E}_{\mathbf{f}}(\mathbb{1}\{i_m = i_l = i_r + 1 \leq n-1, t_{k,hi} < Z(f_k)\}) \\ & \leq P(\eta_0 < -z_{\alpha_2}) + P(\eta_1 > z_{\alpha_2}) + P(\eta_2 < -z_{\alpha_2}) \leq 3\alpha_2 = \frac{\alpha}{8s}. \end{aligned} \quad (\text{A.5.233})$$

Similar arguments show that

$$\mathbb{E}_{\mathbf{f}}(\mathbb{1}\{i_m = i_l = i_r + 1 \geq 1, t_{k,lo} > Z(f_k)\}) \leq 3\alpha_2 = \frac{\alpha}{8s}.$$

Therefore we have

$$P_{\mathbf{f}}(Z(f_k) \notin CI_k) \leq \alpha/2s + 2\alpha_1 + 6\alpha_2 = \alpha/s. \quad (\text{A.5.234})$$

**Proof of Proposition A.5.12**

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}} (|CI_k|^2) \\
& \leq 26^2 \mathbb{E}_{\mathbf{f}} \left( \frac{2^{2J-2\hat{j}_k(\alpha/2s)}}{n^2} \mathbb{1}\{\check{j}_k(\alpha/2s) < \infty, \check{j}_k(\alpha/2s) < \tilde{j}_k\} \right) \\
& + 28^2 \mathbb{E}_{\mathbf{f}} \left( \frac{2^{2J-2\tilde{j}_k}}{n^2} \mathbb{1}\{\tilde{j}_k \leq \hat{j}_k(\alpha/2s)\} \right) \\
& + \mathbb{E}_{\mathbf{f}} (|CI_k|^2 \mathbb{1}\{\check{j}_k(\alpha/2s) = \infty, \tilde{j}_k > J\})
\end{aligned} \tag{A.5.235}$$

Recall Lemma A.5.6, A.5.7 and A.5.9, we have first two terms being bounded by multiple times  $\rho_z((z_{\alpha/2s} + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k) \left( 1 \wedge \sqrt{n \rho_z((z_{\alpha/2s} + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k)} \right)$ , specifically,

$$\begin{aligned}
& E_{\mathbf{f}} (|CI_k|^2) \\
& \leq \check{C}_3 \rho_z((z_{\alpha/2s} + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k)^2 \left( 1 \wedge n \rho_z((z_{\alpha/2s} + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k) \right) \\
& + \mathbb{E}_{\mathbf{f}} (|CI_k|^2 \mathbb{1}\{\check{j}_k(\alpha/2s) = \infty, \tilde{j}_k > J\}),
\end{aligned} \tag{A.5.236}$$

where  $\check{C}_3 > 0$  is an absolute constant.

Note that  $\frac{z_{\alpha/2s} + 1}{z_{\alpha/s}} < 4$ , and invoke Proposition 2.2.1 in Chapter 2, it suffices to bound the remaining term.

We proceed to bound the remaining term. Note that

$$\begin{aligned}
\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k) & \leq \left( 2 \frac{z_{\alpha/8s}}{z_{\alpha/s}} \cdot 4\sqrt{3} \sqrt{\frac{n+1}{n}} \right)^{\frac{2}{3}} \rho_z(z_{\alpha/s} \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k), \\
\frac{n+1}{n} & \leq 2, \quad \frac{z_{\alpha/8s}}{z_{\alpha/s}} < 4 \text{ for } \alpha \leq 0.3.
\end{aligned} \tag{A.5.237}$$

So it is sufficient to have the following lemma for concluding the proof.

**Lemma A.5.10.**

$$\begin{aligned} \mathbb{E}_{\mathbf{f}} \left( |CI_k|^2 \mathbb{1} \{ \check{\mathbf{j}}_k(\alpha/2s) = \infty, \check{\mathbf{j}}_k > J \} \right) \leq \\ \check{C}_4 \rho_z \left( z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k \right)^2 \left( 1 \wedge n \rho_z \left( z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k \right) \right) + 9\mathfrak{D}_z(f_k; n) \end{aligned} \quad (\text{A.5.238})$$

where  $\check{C}_4 > 28^2$  is an absolute constant.

*Proof.* When

$$\rho_z \left( z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k \right) \geq \frac{1}{n}, \quad (\text{A.5.239})$$

lemma A.5.10 holds.

Now we consider the case that

$$\rho_z \left( z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k \right) < \frac{1}{n}. \quad (\text{A.5.240})$$

Note that this means that for  $i \geq i_{m,r}$ ,

$$\begin{aligned} f_k \left( \frac{i+1}{n} \right) - f_k \left( \frac{i}{n} \right) &\geq \frac{1}{n} \frac{\rho_m \left( z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k \right)}{\rho_z \left( z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k \right)} \\ &\geq \frac{1}{\sqrt{2}} z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}} \left( n \rho_z \left( z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k \right) \right)^{-\frac{3}{2}}. \end{aligned} \quad (\text{A.5.241})$$

and similarly for  $i \leq i_{m,l}$ , we have

$$f_k \left( \frac{i-1}{n} \right) - f_k \left( \frac{i}{n} \right) \geq \frac{1}{\sqrt{2}} z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}} \left( n \rho_z \left( z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k \right) \right)^{-\frac{3}{2}}. \quad (\text{A.5.242})$$

Note that on the event  $\{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \check{\mathbf{j}}_k > J\}$ , we have that  $L_k \leq i_{m,l} \leq i_{m,r} \leq U_k$ . We define a “bad” event

$$B_2 = \{i_l \leq i_{m,l} - 1\} \cup \{i_r \geq i_{m,r}\}. \quad (\text{A.5.243})$$

Then we know that

$$\begin{aligned}
& P_{\mathbf{f}}(B_2 \cap \{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \tilde{\mathbf{j}}_k > J\}) \\
& \leq 28\Phi \left( -\sqrt{2}z_{\alpha/8s} \left( n\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \right)^{-\frac{3}{2}} + z_{\alpha_1} \right). \tag{A.5.244}
\end{aligned}$$

On the other hand, for the bad event  $B_1$  defined in (A.5.224), we have

$$\begin{aligned}
& P_{\mathbf{f}}(B_1 \cap \{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \tilde{\mathbf{j}}_k > J\}) \\
& \leq \Phi \left( -\sqrt{2}z_{\alpha/8s} \left( n\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \right)^{-\frac{3}{2}} - z_{\alpha_1} \right). \tag{A.5.245}
\end{aligned}$$

Note that we have  $z_{\alpha/8s} > 1$  for  $0 < \alpha \leq 1$ . Hence we have

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}}(|CI_k|^2 \mathbb{1}\{B_1 \cup B_2\} \mathbb{1}\{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \tilde{\mathbf{j}}_k > J\}) \\
& \leq \frac{28^2}{n^2} \times 40\Phi \left( -(\sqrt{2}-1)z_{\alpha/8s} \left( n\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \right)^{-\frac{3}{2}} \right) \\
& \leq \check{C}_5 \rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k)^2 \left( 1 \wedge n\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \right), \tag{A.5.246}
\end{aligned}$$

where  $\check{C}_5 = 28^2 \times 40 \times \sup_{x>1} x^2 \Phi(-(\sqrt{2}-1)x)$ .

On the remaining event

$$(B_1 \cup B_2)^c \cap \{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \tilde{\mathbf{j}}_k > J\},$$

we have that

$$i_l = i_{m,l}, i_r = i_{m,r} - 1.$$

Now we have two cases. Case 1:  $i_{m,l} = i_{m,r} - 1$ , or  $i_{m,l} = i_{m,r} = 1$  or  $i_{m,l} = i_{m,r} = n - 1$ .

Case 2:  $i_{m,l} = i_{m,r}$  and  $i_{m,l} \neq 1$  and  $i_{m,l} \neq n - 1$ .

For the case 1 , we have  $\mathfrak{D}_z(f_k; n) \geq \frac{1}{n^2}$ , so we have

$$\begin{aligned} & \mathbb{E}_{\mathbf{f}} (|CI_k|^2 \mathbb{1}\{(B_1 \cup B_2)^c\} \mathbb{1}\{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \tilde{\mathbf{j}}_k > J\}) \\ & \leq \frac{9}{n^2} \leq 9\mathfrak{D}_z(f_k; n). \end{aligned} \quad (\text{A.5.247})$$

Combining with Inequality (A.5.246), we have lemma A.5.10.

For the case 2, denote  $i_m = i_{m,l} = i_{m,r}$ , we have

$$\begin{aligned} & \mathbb{E}_{\mathbf{f}} (|CI_k|^2 \mathbb{1}\{(B_1 \cup B_2)^c\} \mathbb{1}\{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \tilde{\mathbf{j}}_k > J\}) \\ & \leq \mathbb{E}_{\mathbf{f}} (2(t_{k,hi} - i_m)^2 \mathbb{1}\{(B_1 \cup B_2)^c\} \mathbb{1}\{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \tilde{\mathbf{j}}_k > J, i_m \leq n-2\}) \\ & \quad + \mathbb{E}_{\mathbf{f}} (2(t_{k,lo} - i_m)^2 \mathbb{1}\{(B_1 \cup B_2)^c\} \mathbb{1}\{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \tilde{\mathbf{j}}_k > J, i_m \geq 2\}) . \end{aligned} \quad (\text{A.5.248})$$

The arguments for bounding the two terms are almost identical (flipping everything around  $i_m$ ), we only bound the first and second share the same bound.

Recall  $t_{k,r}$  defined in Equation (A.5.230), for simplicity of notation, denote

$$D = (B_1 \cup B_2)^c \cap \{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \tilde{\mathbf{j}}_k > J, i_m \leq n-2\}$$

we have

$$\begin{aligned} & \mathbb{E}_{\mathbf{f}} (2(t_{k,hi} - i_m)^2 \mathbb{1}\{D\}) \\ & \leq \mathbb{E}_{\mathbf{f}} \left( \left( 4(t_{k,hi} - t_{k,r})_+^2 + 4(t_{k,r} - \frac{i_m}{n})^2 \right) \mathbb{1}\{D\} \right) \\ & \leq 4\mathfrak{D}_z(f_k; n) + 4\mathbb{E}_{\mathbf{f}} ((t_{k,hi} - t_{k,r})_+^2 \mathbb{1}\{D\}) . \end{aligned} \quad (\text{A.5.249})$$

To bound the second term, we will split event  $D$  into  $D \cap A$  and  $D \cap A^c$ , where  $A$  is an event define later. We will consider the expectation on these two events.

Recall the joint distribution of the quantities in the numerator and denominator of  $t_{k,hi}$  under  $(B_1 \cup B_2)^c \cap \{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \tilde{\mathbf{j}}_k > J, i_m \leq n-2\}$ , as explained in Equation (A.5.232),

denote  $\varepsilon = \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}}$ , when further under the event  $t_{k,hi} > \frac{i_m}{n}$  (the only one we need to consider),  $t_{k,hi} - t_{k,r}$  is upper bounded:

$$\begin{aligned}
t_{k,hi} - t_{k,r} \leq & \frac{\varepsilon\eta_0 \left(f_k\left(\frac{i_m+2}{n}\right) - f_k\left(\frac{i_m+1}{n}\right)\right) + \varepsilon\eta_1 \left(f_k\left(\frac{i_m}{n}\right) - f_k\left(\frac{i_m+2}{n}\right)\right) + \varepsilon\eta_2 \left(f_k\left(\frac{i_m+1}{n}\right) - f_k\left(\frac{i_m}{n}\right)\right)}{n \left(f_k\left(\frac{i_m+2}{n}\right) - f_k\left(\frac{i_m+1}{n}\right) + \varepsilon\eta_2 - \varepsilon\eta_1 + 2\varepsilon z_{\alpha_2}\right) \left(f_k\left(\frac{i_m+2}{n}\right) - f_k\left(\frac{i_m+1}{n}\right)\right)} \\
& + \frac{2z_{\alpha_2}\varepsilon \left(f_k\left(\frac{i_m+2}{n}\right) - f_k\left(\frac{i_m}{n}\right)\right)}{n \left(f_k\left(\frac{i_m+2}{n}\right) - f_k\left(\frac{i_m+1}{n}\right) + \varepsilon\eta_2 - \varepsilon\eta_1 + 2\varepsilon z_{\alpha_2}\right) \left(f_k\left(\frac{i_m+2}{n}\right) - f_k\left(\frac{i_m+1}{n}\right)\right)}.
\end{aligned} \tag{A.5.250}$$

The reason it is not an equation is due to the possibility of upper truncation if  $t_{k,hi}$  by  $\frac{i_m+1}{n}$

Recall that we define  $\eta_0, \eta_1, \eta_2$  in Equation (A.5.232).

Now we consider a “good” event

$$A = \{\eta_1 \leq \frac{f_k\left(\frac{i_m+2}{n}\right) - f_k\left(\frac{i_m+1}{n}\right)}{6\varepsilon} + \frac{1}{2}\varepsilon z_{\alpha_2}, \eta_2 \geq -\frac{f_k\left(\frac{i_m+2}{n}\right) - f_k\left(\frac{i_m+1}{n}\right)}{6\varepsilon} - \frac{1}{2}\varepsilon z_{\alpha_2}\}. \tag{A.5.251}$$

Under this good event  $A$ , we have

$$f_k\left(\frac{i_m+2}{n}\right) - f_k\left(\frac{i_m+1}{n}\right) + \varepsilon\eta_2 - \varepsilon\eta_1 + 2\varepsilon z_{\alpha_2} \geq \frac{2}{3} \left(f_k\left(\frac{i_m+2}{n}\right) - f_k\left(\frac{i_m+1}{n}\right)\right) + \varepsilon z_{\alpha_2}. \tag{A.5.252}$$

Then we have that

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}} \left( (t_{k,hi} - t_{k,r})_+^2 \mathbb{1}\{D \cap A\} \right) \\
& \leq 4 \frac{1}{n^2} \left( \frac{\varepsilon}{\frac{2}{3} (f_k(\frac{i_m+2}{n}) - f_k(\frac{i_m+1}{n})) + \varepsilon z_{\alpha_2}} \right)^2 (1 + 4 + 1 + 16z_{\alpha_2}^2) \\
& \leq 4 \frac{1}{n^2} \left( \frac{1}{\frac{2}{3} \cdot 2z_{\alpha/8s} \left( n\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \right)^{-\frac{3}{2}} + z_{\alpha/24s}} \right)^2 (6 + 16z_{\alpha/24s}^2) \\
& \leq \rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k)^2 \cdot \left( n\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \right) \left( 13.5 + 36 \left( \frac{z_{\alpha/24s}}{z_{\alpha/8s}} \right)^2 \right).
\end{aligned} \tag{A.5.253}$$

The second inequality is due to Inequality (A.5.241).

Also note that  $\frac{z_{\alpha/24s}}{z_{\alpha/8s}} < 2$  for  $\alpha < 1$ . Hence we have that

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}} \left( (t_{k,hi} - t_{k,r})_+^2 \mathbb{1}\{D \cap A\} \right) \\
& < 86\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k)^2 \cdot \left( n\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \right).
\end{aligned} \tag{A.5.254}$$

For event  $A^c \cap D$ , we have

$$P(A^c \cap D) \leq 2\Phi \left( -\frac{z_{\alpha/8s}}{3} \left( n\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \right)^{-\frac{3}{2}} \right). \tag{A.5.255}$$

Therefore we have

$$\begin{aligned}
& \mathbb{E}_{\mathbf{f}} \left( (t_{k,hi} - t_{k,r})_+^2 \mathbb{1}\{D \cap A^c\} \right) \\
& \leq 18\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k)^2 \cdot \left( n\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \right).
\end{aligned} \tag{A.5.256}$$

Adding up the expectation on event  $D \cap A^c$  and  $D \cap A$  and going back to Inequality (A.5.249), we have the first term in (A.5.248) bounded. Using similar arguments, the second term can

be bounded by the same bound. So we have

$$\begin{aligned} & \mathbb{E}_{\mathbf{f}} (|CI_k|^2 \mathbb{1}\{(B_1 \cup B_2)^c\} \mathbb{1}\{\check{\mathbf{j}}_k(\alpha/2s) = \infty, \check{\mathbf{j}}_k > J\}) \\ & \leq 8D(f_k; n) + 832\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k)^2 \cdot \left( n\rho_z(z_{\alpha/8s} \frac{2\sqrt{12}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \right). \end{aligned} \quad (\text{A.5.257})$$

This concludes case 2, thus the proof of the lemma. □

### A.5.16. Proof of Theorem 3.4.3

Note that  $\mathfrak{D}_m(\mathbf{f}; n) \geq \sum_{k=1}^s (\min\{f_k(\frac{i}{n}) : 0 \leq i \leq n\} - M(f_k))$ . Recall the lower bound of  $\tilde{\mathbf{L}}_{m,\alpha,n}(\sigma; \mathbf{f})$  given in Equation (A.5.157). Note that  $\rho_z(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k) \leq 1$  for all  $k \in \{1, 2, \dots, s\}$ . Using Cauchy-Schwartz inequality, we know that it suffices to prove that

$$\begin{aligned} \mathbb{E} \left( \hat{M} - M(\mathbf{f}) \right)^2 & \leq \left( C_m \sum_{k=1}^s \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \left( 1 \wedge \sqrt{n\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)} \right) \right. \\ & \quad \left. + \sum_{k=1}^s \left( \min\{f_k(\frac{i}{n}) : 0 \leq i \leq n\} - M(f_k) \right) \right)^2, \end{aligned} \quad (\text{A.5.258})$$

for some positive absolute constant  $C_m$ .

Now we will prove this statement.

Recall that  $\zeta = \Phi(-2) < 0.1$ .

For simplicity of notation, denote

$$\hat{\mathbf{f}}_{k,i} = \frac{1}{2^{\hat{\mathbf{j}}_k(\zeta)}} \sum_{w=2^{\hat{\mathbf{j}}_k(\zeta)} \cdot (i-1)}^{2^{\hat{\mathbf{j}}_k(\zeta)} \cdot i-1} f_k\left(\frac{w}{n}\right).$$

Note that  $\{\nu_{k,h}^u : 1 \leq k \leq s, 0 \leq h \leq n, u = l, r, e\}$  are independent. So we have that



$$2^{\hat{\mathbf{j}}_k(\zeta)-J} \mathbf{Y}_{k,\hat{\mathbf{j}}_k(\zeta),\hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\zeta)}+2\Delta_k}^e - \hat{\mathbf{f}}_{k,\hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\zeta)}+2\Delta_k} \left| (\nu_{\cdot,\cdot}^l, \nu_{\cdot,\cdot}^r) \sim N(0, (1 - 2^{\hat{\mathbf{j}}_k(\zeta)-J}) 2^{\hat{\mathbf{j}}_k(\zeta)-J} \cdot 3 \frac{\sigma^2}{(n+1)^{s-1}}) \right|.$$

Also recall the independence between  $\mathbf{er}(\{y_i\})$  and  $\{\nu_{k,h}^u : 1 \leq k \leq s, 0 \leq h \leq n, u = l, r, e\}$ .

So we have that

$$\begin{aligned} & \mathbb{E} \left( \hat{M} - M(\mathbf{f}) \right)^2 \leq \\ & \leq \left( \sqrt{\mathbb{E} \left( \frac{1}{(n+1)^s} \sum_{\mathbf{i} \in \{0,1,2,\dots,n\}^s} \mathbf{er}(\{y_i\}) - f_0 \right)^2} + \sum_{k=1}^s \sqrt{\mathbb{E} \left( \hat{M}_k - M(f_k) \right)^2} \right)^2 \\ & \leq \left( \sqrt{\mathbb{E} \left( \left( \frac{1}{(n+1)^s} \sum_{\mathbf{i} \in \{0,1,2,\dots,n\}^s} \mathbf{er}(\{y_i\}) \right) - f_0 \right)^2} \right. \\ & \quad + \sum_{k=1}^s \left( \sqrt{\mathbb{E} \left( \left( \hat{M}_k - \hat{\mathbf{f}}_{k,\hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\zeta)}+2\Delta_k} \right)^2 \mathbb{1}_{\{\check{\mathbf{j}}_k(\zeta) < \infty\}} \right)} \right. \\ & \quad \left. + \sqrt{\mathbb{E} \left( \left( \hat{\mathbf{f}}_{k,\hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\zeta)}+2\Delta_k} - M(f_k) \right)^2 \mathbb{1}_{\{\check{\mathbf{j}}_k(\zeta) < \infty\}} \right)} \right. \\ & \quad \left. + \sqrt{\mathbb{E} \left( \mathbb{1}_{\{\check{\mathbf{j}}_k(\zeta) = \infty\}} (\hat{M}_k - M(f_k))^2 \right)} \right)^2 \\ & \leq \left( \frac{\sigma}{(n+1)^{\frac{s}{2}}} + \sum_{k=1}^s \left( \sqrt{\frac{3\sigma^2}{(n+1)^{s-1}}} \sqrt{\mathbb{E}(2^{\hat{\mathbf{j}}_k(\zeta)-J} \mathbb{1}_{\{\check{\mathbf{j}}_k(\zeta) < \infty\}})} + \right. \right. \\ & \quad \left. \sqrt{\mathbb{E} \left( \left( \hat{\mathbf{f}}_{k,\hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\zeta)}+2\Delta_k} - M(f_k) \right)^2 \mathbb{1}_{\{\check{\mathbf{j}}_k(\zeta) < \infty\}} \right)} + \right. \\ & \quad \left. \left. \sqrt{\mathbb{E} \left( \left( \hat{M}_k - M(f_k) \right)^2 \mathbb{1}_{\{\check{\mathbf{j}}_k(\zeta) = \infty\}} \right)} \right) \right)^2. \end{aligned} \tag{A.5.259}$$

Now we will continue with bounding the terms in Inequality (A.5.259) separately.

We introduce the following lemma, which we will prove later, to bound the first term in the summation.

**Lemma A.5.11.** *For  $\zeta \leq 0.1$ , we have*

$$\mathbb{E}(2^{\hat{\mathbf{j}}_k(\zeta)} \mathbb{1}_{\{\check{\mathbf{j}}_k(\zeta) < \infty\}}) \leq 37 \cdot 2^{\hat{\mathbf{j}}_k(\zeta)} \tag{A.5.260}$$

for  $k = 1, 2, \dots, s$ , where  $j_k(\zeta)$  is defined in Equation A.5.199.

By definition of  $j_k(\zeta)$ , we know that

$$\frac{2^{J-j_k(\zeta)}}{n} > \xi_k(\zeta). \quad (\text{A.5.261})$$

By Lemma A.5.9, we have that

$$\frac{2^{J-j_k(\zeta)}}{n} > \xi_k(\zeta) \geq \frac{1}{2} \rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k). \quad (\text{A.5.262})$$

Recall that we have  $\zeta < 0.1$  (because  $\zeta = \Phi(-2)$  here).

This combined with Lemma A.5.11 we have that

$$\mathbb{E}(2^{\hat{j}_k(\zeta)-J} \mathbb{1}_{\{\hat{j}_k(\zeta) < \infty\}}) \leq 37 \cdot 2^{j_k(\zeta)-J} \leq \frac{148}{n} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2 \cdot \left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}\right)^{-2}. \quad (\text{A.5.263})$$

The second inequality is due to that

$$\begin{aligned} & \frac{1}{2} \left( (z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}} \right)^2 \\ & \leq \rho_z\left((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k\right) \rho_m\left((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k\right)^2 \\ & \leq \rho_z\left((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k\right) \cdot \left( (z_\zeta + 1) \sqrt{6} \sqrt{\frac{n+1}{n}} \right)^2 \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2. \end{aligned} \quad (\text{A.5.264})$$

Therefore, we have the first term in the summation in Inequality (A.5.259) upper bounded, which we summarize into the following lemma.

**Lemma A.5.12.**

$$\begin{aligned}
& \sqrt{\frac{3\sigma^2}{(n+1)^{s-1}}} \sqrt{\mathbb{E}(2\hat{\mathbf{j}}_k(\zeta) - J \mathbb{1}\{\check{\mathbf{j}}_k(\zeta) < \infty\})} \\
& \leq \min \left\{ \sqrt{\frac{3\sigma^2}{(n+1)^{s-1}}}, \sqrt{\frac{3 \cdot 148(n+1)}{n}} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \right\} \\
& \leq \min \left\{ \sqrt{\frac{6(n+1)}{n}} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right), \sqrt{n\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}, \sqrt{\frac{444(n+1)}{n}} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \right\}.
\end{aligned} \tag{A.5.265}$$

Now we continue with bounding the second term in the summation in Inequality (A.5.259).

Note that our localization step and stopping rule for each coordinate parallel that in Chapter 2, but with noise level  $\frac{\sigma}{(n+1)^{\frac{s}{2}}}$ . So according to Lemma A.1.39 and Lemma A.1.42, we have that

$$\begin{aligned}
& \mathbb{E} \left( \left( \hat{\mathbf{f}}_{k, \hat{\mathbf{i}}_k, \hat{\mathbf{j}}_k(\zeta) + 2\Delta_k} - M(f_k) \right)^2 \mathbb{1}\{\check{\mathbf{j}}_k(\zeta) < \infty\} \right) \\
& \leq \min \left\{ c_{m2} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s-1}{2}}}; \frac{1}{\sqrt{n}}; f_k\right)^2, \check{c}_{m2} \frac{\sigma^2}{(n+1)^{s-1}} \right\} \\
& \leq c_m \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2 \left( 1 \wedge n\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \right),
\end{aligned} \tag{A.5.266}$$

where  $c_{m2}$  and  $\check{c}_{m2}$  are from Lemma A.1.39 and A.1.42, and  $c_m$  is an absolute positive constant.

Now we turn to the third term in the summation in Inequality (A.5.259).

Recall that  $\{\nu_{k,h}^e\}$  is independent from  $\{\nu_{k,h}^l\} \cup \{\nu_{k,h}^r\}$ . Let  $\tilde{f}_k = \min_{\hat{\mathbf{i}}_k, J-2 \leq i \leq \hat{\mathbf{i}}_k, J+2} f_k(\frac{i-1}{n})$ .

Elementary calculation show that

$$\begin{aligned}
& \mathbb{E} \left( \left( \hat{M}_k - M(f_k) \right)^2 \mathbb{1}\{\check{\mathbf{j}}_k(\zeta) = \infty\} \right) \\
& \leq 2 \cdot 5 \cdot \frac{\sigma^2}{(n+1)^{s-1}} P(\check{\mathbf{j}}_k(\zeta) = \infty) + 2\mathbb{E} \left( \left( \tilde{f}_k - M(f_k) \right)^2 \mathbb{1}\{\check{\mathbf{j}}_k(\zeta) = \infty\} \right).
\end{aligned} \tag{A.5.267}$$

Again, note that the localization procedure and stopping rule for each coordinate parallels that in Chapter 2, by Lemma A.1.43 and Lemma A.1.40, we have that

$$\begin{aligned} & \mathbb{E} \left( \left( \tilde{f}_k - M(f_k) \right)^2 \mathbb{1}_{\{\check{\mathbf{j}}_k(\zeta) = \infty\}} \right) \\ & \leq \min \left\{ \check{c}_{m3}^2 \frac{\sigma^2}{(n+1)^{s-1}}, c_{m6} \cdot 2\rho_m \left( \sqrt{\frac{\sigma^2}{(n+1)^s}}; f_k \right)^2 \right\} \\ & \quad + \left( \min \left\{ f_k \left( \frac{i}{n} \right) : 0 \leq i \leq n \right\} - M(f_k) \right)^2. \end{aligned} \quad (\text{A.5.268})$$

And by Lemma A.1.41 we have that

$$\frac{\sigma^2}{(n+1)^{s-1}} P(\check{\mathbf{j}}_k(\zeta) = \infty) \leq 64\rho_m \left( \sqrt{\frac{\sigma^2}{(n+1)^s}}; f_k \right)^2. \quad (\text{A.5.269})$$

Also note that  $\frac{\sigma^2}{(n+1)^{s-1}} \leq 4\rho_m \left( \sqrt{\frac{\sigma^2}{(n+1)^s}}; f_k \right)^2 \cdot n\rho_z \left( \sqrt{\frac{\sigma^2}{(n+1)^s}}; f_k \right)$  and that

$$\frac{\sigma}{(n+1)^{\frac{s}{2}}} \leq \sqrt{2}\rho_m \left( \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k \right) \sqrt{\rho_z \left( \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k \right)}.$$

Adding the three parts together, and going back to Inequality (A.5.259), we have that

$$\begin{aligned} \mathbb{E} \left( \hat{M} - M(\mathbf{f}) \right)^2 & \leq \left( C_m \sum_{k=1}^s \rho_m \left( \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k \right) \left( 1 \wedge \sqrt{n\rho_z \left( \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k \right)} \right) \right. \\ & \quad \left. + \sum_{k=1}^s \left( \min \left\{ f_k \left( \frac{i}{n} \right) : 0 \leq i \leq n \right\} - M(f_k) \right) \right)^2, \end{aligned} \quad (\text{A.5.270})$$

where  $C_m$  is a positive absolute constant. This concludes the proof of the theorem.

Now we give the proof of Lemma A.5.11.

### Proof of Lemma A.5.11

By the definition of  $j_k(\zeta)$ , we immediately have the following facts that we summarize into a lemma

**Lemma A.5.13.** *For  $J \geq j \geq j_k(\zeta) + 5$ , we have that*

$$\frac{1}{\tilde{\sigma}_{k,j}} \sum_{h=(i_{k,j}^*+13)2^{J-j}}^{(i_{k,j}^*+14)2^{J-j}-1} \left( f_k\left(\frac{h}{n}\right) - f_k\left(\frac{h-2^{J-j}}{n}\right) \right) \leq 2^{-2} \times 2^{\frac{3}{2}(5+j_k(\zeta)-j)} (z_\zeta + 1), \quad (\text{A.5.271})$$

or

$$\frac{1}{\tilde{\sigma}_{k,j}} \sum_{h=(i_{k,j}^*-14)2^{J-j}}^{(i_{k,j}^*-13)2^{J-j}-1} \left( f_k\left(\frac{h-2^{J-j}}{n}\right) - f_k\left(\frac{h}{n}\right) \right) \leq 2^{-2} \times 2^{\frac{3}{2}(5+j_k(\zeta)-j)} (z_\zeta + 1). \quad (\text{A.5.272})$$

Therefore, we have that

$$\begin{aligned} & \mathbb{E}(2^{\hat{j}_k(\zeta)} \mathbb{1}_{\{\check{j}_k(\zeta) < \infty\}}) \\ & \leq 2^{j_k(\zeta)} \leq 16 \cdot 2^{j_k(\zeta)} + \sum_{j=j_k(\zeta)+5}^J 2^j \Phi\left(-z_\zeta + \frac{z_\zeta + 1}{4} \cdot 2^{\frac{3}{2}(5+j_k(\zeta)-j)}\right) \leq 37 \cdot 2^{j_k(\zeta)}. \end{aligned} \quad (\text{A.5.273})$$

The last inequality is based on elementary calculation.

### A.5.17. Proof of Theorem 3.4.4

Recall the lower bound of  $\tilde{L}_{m,\alpha,n}(\sigma; \mathbf{f})$  given in Inequality (A.5.158). Using Cauchy-Schwartz inequality, it suffices to prove the following two propositions.

**Proposition A.5.13** (Coverage). *For  $0 < \alpha \leq 0.3$ ,  $CI_{m,\alpha}$  defined in (3.4.25) is a  $1 - \alpha$  level confidence interval for  $M(\mathbf{f})$ .*

**Proposition A.5.14** (Expected Length). *Suppose  $\alpha \leq 0.3$ . For  $CI_{m,\alpha}$  defined in (3.4.25),*

we have

$$\mathbb{E}(|CI_{m,\alpha}|) \leq \mathfrak{D}_m(\mathbf{f}; n) + \bar{C}_{m,\alpha,s} \sum_{k=1}^s \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \left(1 \wedge \sqrt{n\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}\right), \quad (\text{A.5.274})$$

where

$$\begin{aligned} \bar{C}_{m,\alpha,s} = & \left(2\sqrt{3}S_{210,\alpha/8s} + 3(z_{\alpha/4s} + 1)\right) \sqrt{8 \cdot 148 \cdot 2} + \left(\sqrt{3}S_{210,\alpha/8s} + 2\right) \cdot 32 + \\ & (6 + S_{212,\alpha/24s} + z_{\alpha/48s}/\sqrt{2}) \cdot 210 \cdot \sqrt{3} \cdot 32 + z_{\alpha/8} 4\sqrt{6}, \end{aligned} \quad (\text{A.5.275})$$

and  $\mathfrak{D}_m(\mathbf{f}; n)$  is defined in (A.5.150).

### Proof of Proposition A.5.13

Denote

$$\mathbf{j}_k^* = j_{F,k} \wedge J. \quad (\text{A.5.276})$$

Note that  $\zeta = \alpha/4s$  and recall Theorem 3.4.15, we have that for the event  $A_1$  defined by

$$\begin{aligned} A_1 = & \\ \{Z(\mathbf{f})_k \in & \left[\frac{2^{J-\hat{\mathbf{j}}_k(\alpha/4s)+1}}{n} \times (\hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\alpha/4s)-1} - 7) - \frac{1}{2n}, \frac{2^{J-\hat{\mathbf{j}}_k(\alpha/4s)+1}}{n} \times (\hat{\mathbf{i}}_{k,\hat{\mathbf{j}}_k(\alpha/4s)-1} + 6) - \frac{1}{2n}\right] \\ & \text{for } k = 1, 2, \dots, s\}, \end{aligned} \quad (\text{A.5.277})$$

its probability satisfies

$$P(A_1) \geq 1 - \alpha/4. \quad (\text{A.5.278})$$

Note that

$$\left\{ 2^{j_k^*-J} \left( \mathbf{y}_{k, \mathbf{j}_k^*, i}^e - \sum_{w=2^{J-j_k^*}^{i-1}}^{i \cdot 2^{J-j_k^*}-1} f_k\left(\frac{w}{n}\right) \right) + \sqrt{2} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \frac{\sum_{l=0}^n z_{k,l}^1}{n+1} + \right. \\ \left. \frac{1}{(n+1)^s} \sum_{\mathbf{i}} \mathbf{er}(\{y_{\mathbf{i}}\}) - f_0 : 0 \leq i \leq n \right\} \left| \left( \hat{\mathbf{j}}_k(\zeta), \hat{\mathbf{i}}_{k, \hat{\mathbf{j}}_k(\zeta)} \right) \stackrel{i.i.d}{\sim} N(0, 2^{j_k^*-J} \cdot 3 \frac{\sigma^2}{(n+1)^{s-1}}) \right. \right. \quad (\text{A.5.279})$$

for  $i = 0, 1, 2, \dots, n$ . This fact together with the fact that on event  $A_1$ ,

$$\min_{I_{k, lo} \leq i \leq I_{k, hi}} 2^{j_k^*-J} \sum_{w=2^{J-j_k^*}^{i-1}}^{i \cdot 2^{J-j_k^*}-1} f_k\left(\frac{w}{n}\right) = \min_{0 \leq i \leq n} 2^{j_k^*-J} \sum_{w=2^{J-j_k^*}^{i-1}}^{i \cdot 2^{J-j_k^*}-1} f_k\left(\frac{w}{n}\right), \quad (\text{A.5.280})$$

gives

$$P \left( \tilde{M}_{k, md} + \frac{1}{(n+1)^s} \sum_{\mathbf{i}} \mathbf{er}(\{y_{\mathbf{i}}\}) - f_0 - M(f_k) + \sqrt{2} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \frac{\sum_{l=0}^n z_{k,l}^1}{n+1} \right. \\ \left. \leq -S_{210, \alpha/8s} \times \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} \times 2^{\frac{j_k^*-J}{2}} \middle| A_1 \right) \leq \alpha/8s. \quad (\text{A.5.281})$$

Also note that  $\frac{1}{(n+1)^s} \sum_{\mathbf{i}} \mathbf{er}(\{y_{\mathbf{i}}\}) - f_0 \sim N(0, \frac{\sigma^2}{(n+1)^s})$ , elementary calculation on the remainder terms of  $\tilde{M}_{hi}$  gives

$$P \left( \tilde{M}_{hi} \leq M(\mathbf{f}) | A_1 \right) \leq \frac{\alpha}{8} + \frac{\alpha}{8}. \quad (\text{A.5.282})$$

Recollect quantities introduced in (A.5.200) and (A.5.199).

Lemma A.5.9 and the definition of  $\mathbf{j}_k(\zeta)$  gives

$$\frac{2^{J-j_k(\zeta)}}{n} \leq 4\rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k).$$

Therefore

$$2\sqrt{3}(z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}} \sqrt{\frac{n}{2^{J-j_k(\zeta)}}} \geq \rho_m((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k). \quad (\text{A.5.283})$$

This means for  $j \geq j_k(\zeta) + 3$ ,

$$\frac{3\sigma(z_\zeta + 1)}{(n+1)^{\frac{s-1}{2}}} \sqrt{\frac{1}{2^{J-j}}} \geq \rho_m((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k), \quad (\text{A.5.284})$$

and if further  $j \leq J$ ,

$$\min_{w \in \{-2, -1, 0\}} \sum_{h=(i_{k,j}^*+w)2^{J-j}}^{(i_{k,j}^*+w+1)2^{J-j-1}} f_k\left(\frac{h}{n}\right) \leq M(f_k) + \rho_m((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k). \quad (\text{A.5.285})$$

Now we define an event

$$D_{2,k} = \{\check{j}_k(\zeta) \leq j_k(\zeta) - 1\}. \quad (\text{A.5.286})$$

Lemma A.5.5 gives that for  $\zeta \leq 0.1$

$$\begin{aligned} P(D_{2,k}) &\leq P(\tilde{j}_k \leq j_k(\zeta) - 1) + P(\check{j}_k(\zeta) \leq j_k(\zeta) - 1, \tilde{j}_k \geq j_k(\zeta)) \\ &\leq 6\Phi(-z_\zeta - 2) \times \frac{1}{1 - 0.001} + \Phi(-z_\zeta - 2) \frac{1}{1 - 0.001} \leq \zeta \cdot \frac{7}{1 - 0.001} \cdot \exp(-4) \cdot \frac{4}{3} \leq 0.5\zeta. \end{aligned} \quad (\text{A.5.287})$$

Note that  $\zeta = \alpha/4s$ , hence  $P(D_{2,k}) \leq \alpha/8s$  and  $P(\cup_{k=1}^s D_{2,k}) \leq \alpha/8$ .

Equations (A.5.279), (A.5.280), (A.5.284), (A.5.285) together with the apparent fact that

$$\min\{v_1, \dots, v_w\} \leq \max\{v_1, \dots, v_w\}$$



, we have that

$$P\left(\tilde{M}_{k,lo} + \frac{1}{(n+1)^s} \sum_{\mathbf{i}} \mathbf{er}(\{y_{\mathbf{i}}\}) - f_0 + \sqrt{2} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \frac{\sum_{l=0}^n z_{k,l}^1}{n+1} - M(f_k) \geq 0 \right. \\ \left. \left| A_1 \cap D_{2,k}^c \cap \{j_{F,k} \leq J\} \right) \leq \alpha/8s. \quad (\text{A.5.288})$$

Now we introduce a lemma.

**Lemma A.5.14.**

$$P\left(\tilde{M}_{k,lo} \geq M(f_k) \left| A_1 \cap D_{2,k}^c \cap \{j_{F,k} \geq J+1\} \right) \leq \alpha/8s, \quad (\text{A.5.289})$$

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

*Proof.* We prove the inequality for any fixed  $k \in \{1, 2, \dots, s\}$ . Denote  $\delta_i = \nu_{k,i}^e - f_k(\frac{i}{n})$

Note that  $\{\nu_{\cdot,\cdot}^e\}$  is independent with  $\{\nu_{\cdot,\cdot}^l, \nu_{\cdot,\cdot}^r\}$ , elementary calculation show that

$$P(\max\{|\delta_i| : (k_l - 1) \vee 0 \leq i \leq (k_r + 2) \wedge n\} \leq H \left| \nu_{\cdot,\cdot}^l, \nu_{\cdot,\cdot}^r \right) \geq 1 - 2 \cdot \alpha/24s - 2 \cdot \alpha/48s = 1 - \alpha/8s. \quad (\text{A.5.290})$$

Denote event

$$B = \max\{|\delta_i| : k_l \vee 0 \leq i \leq k_r + 2 \wedge n\} \leq H$$

.

On event  $A_1$ , we know that  $\frac{k_l}{n} \leq Z(f_k) \leq \frac{k_r+1}{n}$ .

Recall a geometric fact: for  $t \in [i/n, (i+1)/n]$ , where  $1 \leq i \leq n-2$ , we have that

$$f_k(t) \geq \max\left\{ \frac{f_k(\frac{i}{n}) - f_k(\frac{i-1}{n})}{1/n} \left(t - \frac{i}{n}\right) + f_k\left(\frac{i}{n}\right), \frac{f_k(\frac{i+2}{n}) - f_k(\frac{i+1}{n})}{1/n} \left(t - \frac{i+1}{n}\right) + f_k\left(\frac{i+1}{n}\right) \right\} \quad (\text{A.5.291})$$

and the right hand side are also attainable for some  $f_k$  when  $\{f_k(\frac{i}{n}) : i = 0, 1, \dots, n\}$  are

given.

For  $0 < t \leq 1/n$ , we have that

$$f_k(t) \geq \frac{f(2/n) - f(1/n)}{1/n}(t - 1/n) + f(1/n) \quad (\text{A.5.292})$$

and the right hand side is attainable for some  $f_k$  when  $\{f_k(\frac{i}{n}) : i = 0, 1, \dots, n\}$  are given.

For  $1 > t \geq n - 1/n$ , we have that

$$f_k(t) \geq \frac{f((n-2)/n) - f((n-1)/n)}{1/n}(t - (n-1)/n) + f((n-1)/n). \quad (\text{A.5.293})$$

On event  $B$ , we have that

$$h(i) \leq \min_{t \in [\frac{i}{n}, \frac{i+1}{n}]} f_k(t), \quad (\text{A.5.294})$$

for  $i = t_l, \dots, t_r$ .

Therefore, on event  $A_1 \cap B$ , we have that

$$\tilde{M}_{k,lo} \leq f_k(t). \quad (\text{A.5.295})$$

Also we have

$$\begin{aligned} & P(A_1 \cap B | A_1 \cap D_{2,k}^c \cap \{j_{F,k} \geq J+1\}) \\ &= \mathbb{E} \left( \mathbb{E}(\mathbb{1}\{B\} | \{\nu_{k,\cdot}^l, \nu_{k,\cdot}^r\}) \mathbb{1}\{A_1 \cap D_{2,k}^c \cap \{j_{F,k} \geq J+1\}\} \right) / P(A_1 \cap D_{2,k}^c \cap \{j_{F,k} \geq J+1\}) \\ &\geq 1 - \alpha/8s, \end{aligned} \quad (\text{A.5.296})$$

which gives the statement of the lemma.

□

Write  $\tilde{M}_{lo}$  in the form

$$\begin{aligned}
\tilde{M}_{lo} = & f_0 + \left( (|\{k : j_{F,k} \leq J\}| - 1) \cdot \left( f_0 - \frac{1}{(n+1)^s} \sum_{\mathbf{i} \in \{0,1,2,\dots,n\}^s} \mathbf{er}(\{y_{\mathbf{i}}\}) \right) - \right. \\
& \sum_{k=1}^s \mathbb{1}\{j_{F,k} \leq J\} \sqrt{2} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \frac{\sum_{l=0}^n z_{k,l}^1}{n+1} - z_{\alpha/8} \cdot 2\sqrt{3} \frac{\sigma}{(n+1)^{\frac{s}{2}}} s \Bigg) \\
& + \sum_{k=1}^s \left( \tilde{M}_{k,lo} + \right. \\
& \left. \mathbb{1}\{j_{F,k} \leq J\} \left( \frac{1}{(n+1)^s} \sum_{\mathbf{i} \in \{0,1,2,\dots,n\}^s} \mathbf{er}(\{y_{\mathbf{i}}\}) - f_0 + \sqrt{2} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \frac{\sum_{l=0}^n z_{k,l}^1}{n+1} \right) \right),
\end{aligned} \tag{A.5.297}$$

we have

$$\begin{aligned}
& P \left( \tilde{M}_{lo} > M(\mathbf{f}) \middle| A_1 \cap (\cap_{k=1}^s D_{2,k}^c) \right) \\
& \leq P \left( (|\{k : j_{F,k} \leq J\}| - 1) \left( f_0 - \frac{1}{(n+1)^s} \sum_{\mathbf{i} \in \{0,1,2,\dots,n\}^s} \mathbf{er}(\{y_{\mathbf{i}}\}) \right) \right. \\
& \quad \left. - \sum_{k=1}^s \mathbb{1}\{j_{F,k} \leq J\} \sqrt{2} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \frac{\sum_{l=0}^n z_{k,l}^1}{n+1} - z_{\alpha/8} \cdot 2\sqrt{3} \frac{\sigma}{(n+1)^{\frac{s}{2}}} s > 0 \right. \\
& \quad \left. \middle| A_1 \cap (\cap_{k=1}^s D_{2,k}^c) \right) \\
& + \sum_{k=1}^s \left( P \left( \tilde{M}_{k,lo} + \frac{1}{(n+1)^s} \sum_{\mathbf{i}} \mathbf{er}(\{y_{\mathbf{i}}\}) - f_0 - M(f_k) \geq 0 \middle| A_1 \cap D_{2,k}^c \cap \{j_{F,k} \leq J\} \right) \right. \\
& \quad \times P(A_1 \cap D_{2,k}^c \cap \{j_{F,k} \leq J\} \middle| A_1 \cap (\cap_{k=1}^s D_{2,k}^c)) \\
& + P \left( \tilde{M}_{k,lo} \geq M(f_k) \middle| A_1 \cap D_{2,k}^c \cap \{j_{F,k} \geq J+1\} \right) \\
& \quad \left. \times P(A_1 \cap D_{2,k}^c \cap \{j_{F,k} \geq J+1\} \middle| A_1 \cap (\cap_{k=1}^s D_{2,k}^c)) \right).
\end{aligned} \tag{A.5.298}$$

Inequality (A.5.288) and Lemma A.5.14 gives that the sum of the terms in the summation is upper bounded by  $\alpha/8s$  for each  $k$ .

For the first term, split it into summation of conditional probability on  $A_1 \cap \left(\cap_{k=1}^s D_{2,k}^c\right) \cap \{j_{F,k} = j_k : k = 1, 2, \dots, s\}$  times  $P(A_1 \cap \left(\cap_{k=1}^s D_{2,k}^c\right) \cap \{j_{F,k} = j_k : k = 1, 2, \dots, s\} \mid A_1 \cap \left(\cap_{k=1}^s D_{2,k}^c\right))$  for legitimate  $j$ . Elementary calculation show that the conditional probability on  $A_1 \cap \left(\cap_{k=1}^s D_{2,k}^c\right) \cap \{j_{F,k} = j_k : k = 1, 2, \dots, s\}$  is upper bounded by  $\alpha/8$ .

Therefore

$$P\left(\tilde{M}_{lo} > M(\mathbf{f}) \mid A_1 \cap \left(\cap_{k=1}^s D_{2,k}^c\right)\right) \leq \alpha/8 + \alpha/8 = \alpha/4.$$

Therefore,

$$\begin{aligned} P(M(\mathbf{f}) \notin [\tilde{M}_{lo}, \tilde{M}_{hi}]) &\leq P(A_1^c) + \sum_{k=1}^s P(D_{2,k}) + P\left(\tilde{M}_{lo} > M(\mathbf{f}) \mid A_1 \cap \left(\cap_{k=1}^s D_{2,k}^c\right)\right) \\ &\quad + P\left(\tilde{M}_{hi} < M(\mathbf{f}) \mid A_1 \cap \left(\cap_{k=1}^s D_{2,k}^c\right)\right) \leq \alpha. \end{aligned} \tag{A.5.299}$$

#### Proof of Proposition A.5.14

$$\begin{aligned} \mathbb{E}(\tilde{M}_{hi} - \tilde{M}_{lo}) &= z_{\alpha/8} \frac{4\sqrt{3}\sigma}{(n+1)^{\frac{s}{2}}} s + \sum_{k=1}^s \mathbb{E}(\tilde{M}_{k,hi} - \tilde{M}_{k,lo}) \\ &\leq z_{\alpha/8} 4\sqrt{6} \sum_{k=1}^s \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \sqrt{\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)} + \sum_{k=1}^s \mathbb{E}(\tilde{M}_{k,hi} - \tilde{M}_{k,lo}). \end{aligned} \tag{A.5.300}$$

Recall that  $\mathfrak{D}_m(\mathbf{f}; n)$  defined in (A.5.150) also applies to univariate case by setting  $s = 1$ , more specifically,

$$\mathfrak{D}_m(f_k; n) = \min\{f_k(\frac{i}{n}) : 0 \leq i \leq n\} - \min\{M(h) : h(\frac{i}{n}) = f_k(\frac{i}{n}) \text{ for } 0 \leq i \leq n, h \in \mathcal{F}\}. \tag{A.5.301}$$

Then it is easy to see that

$$\mathfrak{D}_m(\mathbf{f}; n) = \sum_{k=1}^s \mathfrak{D}_m(f_k; n). \quad (\text{A.5.302})$$

So it is sufficient to prove that the following holds for any  $k \in \{1, 2, \dots, s\}$

$$\mathbb{E}(\tilde{M}_{k,hi} - \tilde{M}_{k,lo}) \leq \mathfrak{D}_m(f_k; n) + \tilde{C}_{m,\alpha,s} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \left(1 \wedge \sqrt{n \rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}\right), \quad (\text{A.5.303})$$

where

$$\begin{aligned} \tilde{C}_{m,\alpha,s} = & \left(2\sqrt{3}S_{210,\alpha/8s} + 3(z_{\alpha/4s} + 1)\right) \sqrt{8 \cdot 148 \cdot 2} + \left(\sqrt{3}S_{210,\alpha/8s} + 2\right) \cdot 32 + \\ & (6 + S_{212,\alpha/24s} + z_{\alpha/48s}/\sqrt{2}) \cdot 210 \cdot \sqrt{3} \cdot 32. \end{aligned} \quad (\text{A.5.304})$$

This gives the statement of the proposition by taking  $\bar{C}_{m,\alpha,s} = z_{\alpha/8}4\sqrt{6} + \tilde{C}_{m,\alpha,s}$ .

Next we will prove Inequality (A.5.303).

We have

$$\begin{aligned} \mathbb{E}(\tilde{M}_{k,hi} - \tilde{M}_{k,lo}) & \leq \\ \mathbb{E}((\tilde{M}_{k,hi} - \tilde{M}_{k,lo}) \mathbb{1}\{j_{F,k} \leq J\}) & + \mathbb{E}((\tilde{M}_{k,hi} - \tilde{M}_{k,lo}) \mathbb{1}\{j_{F,k} > J\}). \end{aligned} \quad (\text{A.5.305})$$

For the first term we have

$$\begin{aligned}
& \mathbb{E}((\tilde{M}_{k,hi} - \tilde{M}_{k,lo}) \mathbb{1}\{j_{F,k} \leq J\}) \\
&= \left(2\sqrt{3}S_{210,\alpha/8s} + 3(z_{\alpha/4s} + 1)\right) \frac{\sigma}{(n+1)^{s-1}} \mathbb{E}(2^{\frac{j_{F,k}-J}{2}} \mathbb{1}\{j_{F,k} \leq J\}) \\
&\leq \left(2\sqrt{3}S_{210,\alpha/8s} + 3(z_{\alpha/4s} + 1)\right) \frac{\sigma}{(n+1)^{s-1}} \left(\mathbb{E}(2^{\frac{j_k(\zeta)+3-J}{2}}) \wedge 1\right) \\
&\leq \left(2\sqrt{3}S_{210,\alpha/8s} + 3(z_{\alpha/4s} + 1)\right) \frac{\sigma}{(n+1)^{s-1}} \\
&\quad \left(\sqrt{8 \cdot \frac{148}{n} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)^2 \cdot \left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}\right)^{-2}} \wedge 1\right) \\
&\leq \left(2\sqrt{3}S_{210,\alpha/8s} + 3(z_{\alpha/4s} + 1)\right) \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \\
&\quad \left(\sqrt{8 \cdot \frac{148(n+1)}{n}} \wedge \sqrt{\frac{2(n+1)}{n}} \sqrt{n \rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}\right).
\end{aligned} \tag{A.5.306}$$

The second to last inequality is due to Inequality (A.5.263).

Let  $\tilde{C}_{m,s,\alpha,0} = (2\sqrt{3}S_{210,\alpha/8s} + 3(z_{\alpha/4s} + 1)) \sqrt{8 \cdot 148 \cdot 2}$ , we have

$$\mathbb{E}((\tilde{M}_{k,hi} - \tilde{M}_{k,lo}) \mathbb{1}\{j_{F,k} \leq J\}) \leq \tilde{C}_{m,s,\alpha,0} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \left(1 \wedge \sqrt{n \rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}\right). \tag{A.5.307}$$

Now we turn to the second term in Equation (A.5.305). We introduce two quantities first.

$$\tilde{f}_k = \min_{(I_{k,lo}-1) \wedge 0 \leq i \leq (I_{k,hi}-1) \vee n} f_k\left(\frac{i}{n}\right), \quad \tilde{i}_{k,m} = \arg \min_{(I_{k,lo}-1) \wedge 0 \leq i \leq (I_{k,hi}-1) \vee n} f_k\left(\frac{i}{n}\right). \tag{A.5.308}$$

Note that these two quantities depend on  $\{\nu_{k,\cdot}^l, \nu_{k,\cdot}^r\}$ .

$$\begin{aligned}
& \mathbb{E}((\tilde{M}_{k,hi} - \tilde{M}_{k,lo}) \mathbb{1}\{j_{F,k} > J\}) \\
&\leq \mathbb{E}\left(\left(\tilde{M}_{k,hi} - \tilde{f}_k\right)_+ \mathbb{1}\{j_{F,k} > J\}\right) + \mathbb{E}\left(\left(\tilde{f}_k - \tilde{M}_{k,lo}\right)_+ \mathbb{1}\{j_{F,k} > J\}\right).
\end{aligned} \tag{A.5.309}$$

Note that

$$\tilde{M}_{k,hi} \leq \nu_{k,\tilde{b}_{k,m}}^e + S_{210,\alpha/8s} \times \sqrt{3} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}}, \quad (\text{A.5.310})$$

hence we have that

$$\mathbb{E} \left( \left( \tilde{M}_{k,hi} - \tilde{f}_k \right)_+ \mathbb{1}\{j_{F,k} > J\} \right) \leq P(j_{F,k} > J) \left( \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s}{2}}} + S_{210,\alpha/8s} \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} \right). \quad (\text{A.5.311})$$

**Lemma A.5.15.**

$$\frac{\sigma}{(n+1)^{\frac{s-1}{2}}} P(j_{F,k} > J) \leq 32\rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \left( 1 \wedge \sqrt{n\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)} \right). \quad (\text{A.5.312})$$

*Proof.* Recall that  $\zeta = \alpha/4s \leq 0.25$ . According to Lemma A.5.13, we know that when  $J \geq j_k(\zeta) + 8$ ,

$$P(j_{F,k} > J) \leq \Pi_{j=j_k(\zeta)+5}^{J-3} \Phi(-z_\zeta + 2^{\frac{3}{2}(j_k(\zeta)+5-j)} \frac{z_\zeta+1}{4}) < 0.4^{J-j_k(\zeta)-7}. \quad (\text{A.5.313})$$

By Lemma A.5.9 and the definition of  $j_k(\zeta)$ , we have that

$$0.4^{J-j_k(\zeta)-7} < 2^7 \cdot 2^{j_k(\zeta)-J} < 2^7 \cdot \frac{1}{n\xi_k(\zeta)} \leq 2^8 \frac{1}{n\rho_z((z_\zeta+1)\frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k)} \quad (\text{A.5.314})$$

When  $n\rho_z((z_\zeta+1)\frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k) \geq 2^8$ , we have that

$$2^{j_k(\zeta)-J+8} < \frac{1}{n\xi_k(\zeta)} \cdot 2^8 \leq 2^9 \cdot \frac{1}{n\rho_z((z_\zeta+1)\frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}}\sqrt{n}}; f_k)} \leq 2. \quad (\text{A.5.315})$$

Note that  $2^{j_k(\zeta)-J+8}$  only takes integer value, hence we have  $j_k(\zeta) - J + 8 \leq 0$ . Hence

$$\begin{aligned} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} P(j_{F,k} > J) &\leq \sqrt{2} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \cdot 2^8 \frac{1}{\sqrt{n \rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k)}} \cdot \sqrt{2} \\ &\leq 32 \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right). \end{aligned} \quad (\text{A.5.316})$$

Also, we always have

$$\begin{aligned} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} P(j_{F,k} > J) &\leq \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} \\ &\leq \sqrt{2} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \sqrt{\frac{n+1}{n}} \sqrt{n \rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)} \quad (\text{A.5.317}) \\ &\leq 32 \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \sqrt{\frac{n \rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}{2^8}} \end{aligned}$$

Note that when  $\sqrt{\frac{n \rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}{2^8}} \geq 1$ , we have  $n \rho_z((z_\zeta + 1) \frac{\sqrt{6}\sigma}{(n+1)^{\frac{s-1}{2}} \sqrt{n}}; f_k) \geq 2^8$ , in which case we have Inequality (A.5.316) holds.

So we have

$$\begin{aligned} \frac{\sigma}{(n+1)^{\frac{s-1}{2}}} P(j_{F,k} > J) &\leq 32 \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \left(1 \wedge \sqrt{\frac{n \rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}{2^8}}\right) \\ &\leq 32 \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \left(1 \wedge \sqrt{n \rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}\right). \end{aligned} \quad (\text{A.5.318})$$

□



With Lemma A.5.15, going back to inequality (A.5.311), we have

$$\begin{aligned} \mathbb{E} \left( \left( \tilde{M}_{k,hi} - \tilde{f}_k \right)_+ \mathbb{1}\{j_{F,k} > J\} \right) \leq \\ \left( \sqrt{3}S_{210,\alpha/8s} + 2 \right) \cdot 32\rho_m \left( \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k \right) \left( 1 \wedge \sqrt{n\rho_z \left( \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k \right)} \right). \end{aligned} \quad (\text{A.5.319})$$

Now we turn to the second term in Inequality (A.5.309).

We have the following lemma

**Lemma A.5.16.** *Let  $\mathfrak{D}_m(f_k; n)$  be defined in (A.5.301). Then we have*

$$\begin{aligned} \mathbb{E} \left( \left( \tilde{f}_k - \tilde{M}_{k,lo} \right)_+ \mathbb{1}\{j_{F,k} > J\} \right) \leq \mathfrak{D}_m(f_k; n) + (6 + S_{212,\alpha/24s} + z_{\alpha/48s}/\sqrt{2}) \cdot 210 \cdot \sqrt{3} \times \\ 32\rho_m \left( \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k \right) \left( 1 \wedge \sqrt{n\rho_z \left( \frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k \right)} \right) \end{aligned} \quad (\text{A.5.320})$$

*Proof.* We first recall a basic geometry property of univariate convex functions. Suppose  $f$  is a convex function. For any  $0 \leq i \leq j \leq n$ , we have that

$$\min_{i \leq l \leq j} \left\{ f\left(\frac{l}{n}\right) \right\} - \min_{\frac{i}{n} \leq t \leq \frac{j}{n}} f(t) \leq \min_{0 \leq l \leq n} \left\{ f\left(\frac{l}{n}\right) \right\} - \min_{0 \leq t \leq 1} f(t). \quad (\text{A.5.321})$$

For  $0 \leq i \leq n-1$ , we define a reference number  $\tilde{h}(i)$ , which is the smallest number a function  $h$  could achieve on  $[i/n, (i+1)/n]$  when it has the same values with  $f_k$  on the grid points (i.e  $0, 1/n, 2/n, \dots, 1$ ).

$$\begin{aligned} \tilde{h}(i) = \min_{i/n \leq t \leq (i+1)/n} \max \left\{ f_k\left(\frac{i+1}{n}\right) + \frac{f_k\left(\frac{i+2}{n}\right) - f_k\left(\frac{i+1}{n}\right)}{1/n} \left(t - \frac{i+1}{n}\right), \right. \\ \left. f_k\left(\frac{i}{n}\right) + \frac{f_k\left(\frac{i-1}{n}\right) - f_k\left(\frac{i}{n}\right)}{1/n} \left(t - \frac{i}{n}\right) \right\}, \end{aligned} \quad (\text{A.5.322})$$

where  $f(-1/n) = \infty = f(n+1/n)$  and  $\infty \times 0$  is set to 0.

Therefore, we have that

$$\begin{aligned}
& \mathbb{E} \left( \left( \tilde{f}_k - \tilde{M}_{k,lo} \right)_+ \mathbb{1}\{j_{F,k} > J\} \right) \\
& \leq \mathbb{E} \left( \mathbb{E} \left( \left( \tilde{f}_k - \min_{t_l \leq i \leq t_r} \tilde{h}(i) \right)_+ + \sum_{i=k_l}^{k_r} (\tilde{h}(i) - h(i))_+ \middle| \{\nu_{\cdot,\cdot}^r, \nu_{\cdot,\cdot}^l\} \right) \mathbb{1}\{j_{F,k} > J\} \right) \\
& \leq \mathfrak{D}_m(f_k; n) P(j_{F,k} > J) + \mathbb{E} \left( \sum_{i=k_l}^{k_r} \mathbb{E} \left( (\tilde{h}(i) - h(i))_+ \middle| \{\nu_{\cdot,\cdot}^r, \nu_{\cdot,\cdot}^l\} \right) \mathbb{1}\{j_{F,k} > J\} \right).
\end{aligned} \tag{A.5.323}$$

Now we are left with bounding the second term.

Recollect the notation  $\delta_i = \nu_{k,i}^e - f_k(\frac{i}{n})$  for  $0 \leq i \leq n$ , and  $\delta_i = 0$  for  $i \notin \{0, 1, \dots, n\}$ .

Elementary calculation shows that

$$(\tilde{h}(i) - h(i))_+ \leq 2|\delta_i| + 2|\delta_{i+1}| + |\delta_{i-1}| + |\delta_{i+2}| + 3H. \tag{A.5.324}$$

And note that for fixed  $i$ ,  $\delta_{i-1}, \delta_i, \delta_{i+1}, \delta_{i+2}$  are independent from  $\{\nu_{\cdot,\cdot}^l, \nu_{\cdot,\cdot}^r\}$ .

Also  $\delta_i \sim N(0, \frac{n}{n+1} \frac{3\sigma^2}{(n+1)^{s-1}})$ .

Therefore, we have that

$$\begin{aligned}
& \mathbb{E} \left( \sum_{i=k_l}^{k_r} \mathbb{E} \left( (\tilde{h}(i) - h(i))_+ \middle| \{\nu_{\cdot,\cdot}^r, \nu_{\cdot,\cdot}^l\} \right) \mathbb{1}\{j_{F,k} > J\} \right) \\
& \leq \frac{\sqrt{3}\sigma}{(n+1)^{\frac{s-1}{2}}} (6 + S_{212,\alpha/24s} + z_{\alpha/48s}/\sqrt{2}) \cdot 210 P(j_{F,k} > J) \\
& \leq (6 + S_{212,\alpha/24s} + z_{\alpha/48s}/\sqrt{2}) \cdot 210 \cdot \sqrt{3} \times \\
& \quad 32\rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \left(1 \wedge \sqrt{n\rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}\right).
\end{aligned} \tag{A.5.325}$$

The last inequality comes from Lemma A.5.15.

This concludes the proof of Lemma A.5.16. □

Now, combining Lemma A.5.16, Inequality (A.5.309), Inequality (A.5.319) and Inequality (A.5.307), we have that

$$\mathbb{E}(\tilde{M}_{k,hi} - \tilde{M}_{k,lo}) \leq \mathfrak{D}_m(f_k; n) + \tilde{C}_{m,\alpha,s} \rho_m\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right) \left(1 \wedge \sqrt{n \rho_z\left(\frac{\sigma}{(n+1)^{\frac{s}{2}}}; f_k\right)}\right), \quad (\text{A.5.326})$$

where

$$\begin{aligned} \tilde{C}_{m,\alpha,s} = & \left(2\sqrt{3}S_{210,\alpha/8s} + 3(z_{\alpha/4s} + 1)\right) \sqrt{8 \cdot 148 \cdot 2} + \left(\sqrt{3}S_{210,\alpha/8s} + 2\right) \cdot 32 + \\ & (6 + S_{212,\alpha/24s} + z_{\alpha/48s}/\sqrt{2}) \cdot 210 \cdot \sqrt{3} \cdot 32. \end{aligned} \quad (\text{A.5.327})$$

## A.6. Proofs of the Results in Chapter 4

In this section, we give all the proofs of the results in Chapter 4. We start with proving three overall results for our examples using statistical-optimization-interplay results and optimization results, which are proved later. Next we prove the statistical-optimization interplay results for our examples. Then we prove optimization results for our general optimization template. In the end, we prove the optimization results for our examples.

### A.6.1. Proof of Theorem 4.3.3

Recall Theorem 4.3.1, Theorem 4.3.2.

According to Theorem 4.3.2, we have

$$\begin{aligned}\delta &\leq \frac{\alpha^2 \tilde{L}_\alpha d_1 d_2}{T} + (4\alpha \tilde{L}_\alpha \sqrt{d_1 d_2} + 2L_\alpha) \sqrt{\frac{1}{t}} \sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} + 2\tilde{L}_\alpha \frac{1}{t} \left( q(\beta) + \frac{2d_1 d_2}{\beta} \right), \\ \max\{\delta_1, \delta_2, \delta_0\} &\leq \sqrt{\frac{1}{t}} \sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} \leq u_0,\end{aligned}\tag{A.6.1}$$

Therefore,  $L_{\alpha+\delta_1} \leq 2L_\alpha$ .

Combing with Theorem 4.3.1 through plugging in the bounds of  $\delta, \delta_1, \delta_2$ , we have the following holds with probability at least  $1 - \frac{c_1}{d_1+d_2}$ .

$$\begin{aligned}D(l(M) \| l(\tilde{M})) &\leq 2c_0 L_\alpha \left( \alpha \sqrt{r d_1 d_2} + \sqrt{\frac{1}{t}} \sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} \right) \sqrt{\frac{d_1 + d_2}{n d_1 d_2}} \sqrt{1 + \frac{(d_1 + d_2) \log(d_1 d_2)}{n}} \\ &\quad + \frac{\alpha^2 \tilde{L}_\alpha d_1 d_2}{T n} + \frac{4\alpha \tilde{L}_\alpha \sqrt{d_1 d_2} + 2L_\alpha}{n} \sqrt{\frac{1}{t}} \sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} + \frac{2\tilde{L}_\alpha}{n} \frac{1}{t} \left( q(\beta) + \frac{2d_1 d_2}{\beta} \right).\end{aligned}\tag{A.6.2}$$

### A.6.2. Proof of Theorem 4.4.3

Note that according to Theorem 4.4.2, we have

$$\delta_1 \leq \sqrt{\frac{q_1(\beta)(\lambda|\mathcal{O}|)^2 \min\{N, T\} + q_2(\beta)C(Y)^2 + q_3(\beta)(\|Y\|^2 + 2(NT - |\mathcal{O}|)L_{\max}^2)}{k - q_0(\beta)}},\tag{A.6.3}$$

where  $q_0(\beta), q_1(\beta), q_2(\beta), q_3(\beta)$  are defined in Theorem 4.4.2.

Noting that

$$C(Y) = \sup_{L \in C_1} \|\mathbf{P}_{\mathcal{O}}(Y - L)\| \leq \|Y\| + L_{\max} \sqrt{|\mathcal{O}|} \leq \sqrt{2\|Y\|^2 + 2L_{\max}^2 |\mathcal{O}|},\tag{A.6.4}$$

we have

$$\begin{aligned}
\delta_1 &\leq \\
&\sqrt{\frac{q_1(\beta)(\lambda|\mathcal{O}|)^2 \min\{N, T\} + (2q_2(\beta) + q_3(\beta)) \|Y\|^2 + (2|\mathcal{O}|(q_2(\beta) - q_3(\beta)) + 2NTq_3(\beta)) L_{\max}^2}{k - q_0(\beta)}} \\
&\leq \sqrt{\frac{q_1(\beta)(\lambda|\mathcal{O}|)^2 \min\{N, T\} + (2q_2(\beta) + q_3(\beta)) \|Y\|^2 + 2NT \max\{q_2(\beta), q_3(\beta)\} L_{\max}^2}{k - q_0(\beta)}}.
\end{aligned} \tag{A.6.5}$$

Note that we have  $\lambda|\mathcal{O}| \leq 13 \times 8\sigma \max\{\sqrt{N}, \sqrt{T}\} \log^{\frac{3}{2}}(N + T)$ , we have

$$\begin{aligned}
\delta_1 &\leq \\
&\sqrt{\frac{104^2 q_1(\beta) \sigma^2 NT \log^3(N + T) + (2q_2(\beta) + q_3(\beta)) \|Y\|^2 + 2NT \max\{q_2(\beta), q_3(\beta)\} L_{\max}^2}{k - q_0(\beta)}}.
\end{aligned} \tag{A.6.6}$$

Let  $\widetilde{q_1(\beta)} = 104^2 q_1(\beta)$ ,  $\widetilde{q_2(\beta)} = 2q_2(\beta) + q_3(\beta)$ ,  $\widetilde{q_3(\beta)} = 2 \max\{q_2(\beta), q_3(\beta)\}$ , then we have

$$\delta_1 \leq \sqrt{\frac{\widetilde{q_1(\beta)} \sigma^2 NT \log^3(N + T) + \widetilde{q_2(\beta)} \|Y\|^2 + \widetilde{q_3(\beta)} NT L_{\max}^2}{k - q_0(\beta)}}. \tag{A.6.7}$$

In the proof of Theorem 4.4.2, we derive the bound for  $\delta_1$  through that of  $\delta_0$ , the  $L_2$  distance between the resulting approximate solution of inner loop and the target exact solution of the inner loop. So the bound in Inequality (A.6.7) also holds for  $\delta_0$ . We set the upper bound for inner loop error  $\delta_0$  at iteration number  $k$  as

$$\delta(k) = \sqrt{\frac{\widetilde{q_1(\beta)} \sigma^2 NT \log^3(N + T) + \widetilde{q_2(\beta)} \|Y\|^2 + \widetilde{q_3(\beta)} NT L_{\max}^2}{k - q_0(\beta)}}.$$

Invoking outer loop convergence rate, Proposition 4.4.1, similarly to the proof in Theorem 4.4.2, we have the optimization error for objective function as defined in (4.4.5) is upper

bounded as follows.

$$\delta \leq \frac{NTL_{\max}^2}{|\mathcal{O}|K} + \frac{2\delta(k)^2}{|\mathcal{O}|} + \delta(k) \left( \frac{4L_{\max}\sqrt{NT}}{|\mathcal{O}|} + \frac{2C(Y)}{|\mathcal{O}|} + \min\{\sqrt{N}, \sqrt{T}\}\lambda \right). \quad (\text{A.6.8})$$

Using

$$C(Y) \leq \|Y\| + L_{\max}\sqrt{|\mathcal{O}|} \leq \|Y\| + L_{\max}\sqrt{NT} \quad (\text{A.6.9})$$

and invoking Theorem 4.4.1 we get the statement of Theorem 4.3.3.

### A.6.3. Proof of Theorem 4.5.4

Denote  $F(\theta) = \frac{\|\mathbf{X}\theta\|_2^2}{2n} - \lambda_n\|\theta\|_1^2$ .

From Inequality (A.6.145), Lemma (A.6.7), Theorem (4.5.1), we know that with probability at least  $1 - \frac{\exp(-n/32)}{1 - \exp(-n/32)} - \exp(-\frac{n}{2}) - \frac{1}{2(n+d)}$  the following holds.

$$\|\mathbf{X}^T w\|_{\infty} < 4\rho(\Sigma) \sqrt{1 + \frac{\log d}{n}} \sqrt{\frac{\log 2(n+d)}{n}}, \quad (\text{A.6.10})$$

$$\frac{\|\mathbf{X}\theta\|_2^2}{n} \geq c_1\kappa\|\theta\|_2^2 - c_2\rho^2(\Sigma) \frac{\log d}{n} \|\theta\|_1^2, \quad (\text{A.6.11})$$

$$\|\theta - \theta^*\| \leq \frac{F(\theta) - F(\hat{\theta})}{2\lambda_n\sqrt{s}} + \frac{\|\theta_{Sc}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa}, \quad (\text{A.6.12})$$

where  $c_1, c_2$  are constants and can be taken as  $c_1 = 1/8, c_2 = 50$ .

Therefore, the condition in Theorem 4.5.3 is satisfied with

$$a_1 = c_1\kappa, a_2 = c_2\rho^2(\Sigma) \frac{\log d}{n}. \quad (\text{A.6.13})$$

We only need to prove that under these conditions  $F(\theta_k) - F(\hat{\theta}) \leq \delta_k$  holds.

By Inequality (4.5.11) in Theorem 4.5.3 and Inequality (A.6.12) we have

$$F(\theta_k) - F(\hat{\theta}) \leq \frac{\|\mathbf{X}^T \mathbf{X}/n\|_s}{2k} \|\hat{\theta}\|_2^2 \leq \frac{\|\mathbf{X}^T \mathbf{X}/n\|_s}{2k} \left( \|\theta^*\|_2 + \frac{\|\theta_{Sc}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa} \right)^2. \quad (\text{A.6.14})$$

According to Inequality (A.6.160) we know that

$$F(\theta_k) - F(\hat{\theta}) \leq F(\theta_0) - F(\hat{\theta}) \leq \frac{\|y\|_2^2}{2n}. \quad (\text{A.6.15})$$

For  $k \geq K_0$ , Inequality (4.5.11) and Inequality (A.6.160) gives

$$F(\theta_k) - F(\hat{\theta}) \leq F(\theta_{K_0}) - F(\hat{\theta}) \leq \frac{\lambda_n^2}{48c_2\rho(\Sigma)^{\frac{\log d}{n}}}. \quad (\text{A.6.16})$$

Now we are only left to prove for  $k \geq K_0$ ,

$$F(\theta_k) - F(\hat{\theta}) \leq \max \left\{ 2^{-T_k} \frac{\lambda_n^2}{48c_2\rho^2(\Sigma)^{\frac{\log d}{n}}}, \rho^2(\Sigma) \frac{\log d}{n} s \cdot \left( \frac{2\|\theta_{Sc}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa} \right)^2 \cdot 768c_2 \right\}, \quad (\text{A.6.17})$$

which is also Inequality (4.5.14) in Remark 4.5.4.

To prove this, we only need to prove that for  $k_1 \geq k_0$  and  $k$  satisfying

$$k \geq \begin{cases} k_1 + \lceil \frac{\log 1/6}{\log(1 - \frac{c_1 \kappa}{8\|\mathbf{X}^T \mathbf{X}/n\|_s})} \rceil, & \text{when } c_1 \kappa < 8\|\mathbf{X}^T \mathbf{X}/n\|_s \\ k_1 + 1, & \text{otherwise} \end{cases}, \quad (\text{A.6.18})$$

the following holds

$$F(\theta_k) - F(\hat{\theta}) \leq \max\left\{\frac{1}{2} \left(F(\theta_{k_1}) - F(\hat{\theta})\right), \rho^2(\Sigma) \frac{\log d}{n} s \cdot \left(\frac{2\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa}\right)^2 \cdot 768c_2\right\}. \quad (\text{A.6.19})$$

If

$$F(\theta_k) - F(\hat{\theta}) \geq \rho^2(\Sigma) \frac{\log d}{n} s \cdot \left(\frac{2\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa}\right)^2 \cdot 768c_2, \quad (\text{A.6.20})$$

$$\text{then } F(\theta_{k_1}) - F(\hat{\theta}) \geq \rho^2(\Sigma) \frac{\log d}{n} s \cdot \left(\frac{2\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa}\right)^2 \cdot 768c_2.$$

Also, since  $k_1 \geq K_0$ , we have

$$F(\theta_{k_1}) - F(\hat{\theta}) \leq F(\theta_{K_0}) - F(\hat{\theta}) \leq \frac{\lambda_n^2}{48c_2\rho(\Sigma) \frac{\log d}{n}}. \quad (\text{A.6.21})$$

By Inequality (4.5.10), we have

$$\begin{aligned} F(\theta_k) - F(\hat{\theta}) &\leq \frac{1}{6} \left(F(\theta_{k_1}) - F(\hat{\theta})\right) \\ &\quad + \rho^2(\Sigma) \frac{\log d}{n} s \cdot \left(\frac{2\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa}\right)^2 \cdot 128c_2 \\ &\quad + \frac{8c_2\rho^2(\Sigma) \frac{\log d}{n}}{\lambda_n^2} \frac{\lambda_n^2}{48c_2\rho(\Sigma) \frac{\log d}{n}} \left(F(\theta_{k_1}) - F(\hat{\theta})\right) \\ &\leq \frac{1}{2} \left(F(\theta_{k_1}) - F(\hat{\theta})\right). \end{aligned} \quad (\text{A.6.22})$$

Thus we concludes the proof.



#### A.6.4. Proof of Theorem 4.3.1

The structure of the proof is similar to the proof of Theorem 2 in Davenport et al. (2014), but to show how the statistical-optimization interface work, we will show in details how the optimization error terms get into the statistical accuracy.

Let

$$\bar{\mathcal{L}}_{\Omega,Y}(X) = \mathcal{L}_{\Omega,Y}(X) - \mathcal{L}_{\Omega,Y}(\mathbf{0}). \quad (\text{A.6.23})$$

Then we know that

$$-\bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) \leq -\bar{\mathcal{L}}_{\Omega,Y}(\hat{M}) + \delta \leq -\bar{\mathcal{L}}_{\Omega,Y}(M) + \delta. \quad (\text{A.6.24})$$

We also know that

$$\|\tilde{M}\|_* \leq \alpha\sqrt{rd_1d_2} + \delta_2, \|\tilde{M}\|_\infty \leq \alpha + \delta_1. \quad (\text{A.6.25})$$

We have the following lemma, which we will proof later in this section.

**Lemma A.6.1.** *Let  $G \in \mathbb{R}^{d_1 \times d_2}$  be*

$$G = \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_* \leq \alpha\sqrt{rd_1d_2} + \delta_2, \|\tilde{M}\|_\infty \leq \alpha + \delta_1\} \quad (\text{A.6.26})$$

*for some  $r \leq \min\{d_1, d_2\}$  and  $\alpha \geq 0$ . Then*

$$\begin{aligned} & \mathbb{P} \left( \sup_{X \in G} |\bar{\mathcal{L}}_{\Omega,Y}(X) - \mathbb{E}\bar{\mathcal{L}}_{\Omega,Y}(X)| \geq \tilde{c}_0 L_{\alpha+\delta_1} \left( \alpha\sqrt{rd_1d_2} + \delta_2 \right) \sqrt{\frac{n(d_1+d_2)}{d_1d_2} + \log(d_1d_2)} \right) \\ & \leq \frac{c_1}{d_1+d_2}, \end{aligned} \quad (\text{A.6.27})$$

*where  $\tilde{c}_0, c_1$  are absolute constants and the probability and the expectation are both over the choice of  $\Omega$  and draw of  $Y$ .*

Note that for any  $X$  we have

$$\begin{aligned}
& \mathbb{E} \left( \bar{\mathcal{L}}_{\Omega,Y}(X) - \bar{\mathcal{L}}_{\Omega,Y}(M) \right) \\
&= \frac{n}{d_1 d_2} \sum_{i,j} \left( l(M_{i,j}) \log \left( \frac{l(X_{i,j})}{l(M_{i,j})} \right) + \log \left( \frac{1 - l(X_{i,j})}{1 - l(M_{i,j})} \right) \right) \\
&= -nD(l(M) \| l(X)).
\end{aligned} \tag{A.6.28}$$

Therefore, we have

$$\begin{aligned}
& -\delta \\
& \leq \bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) - \bar{\mathcal{L}}_{\Omega,Y}(M) \\
&= \mathbb{E} \left( \bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) - \bar{\mathcal{L}}_{\Omega,Y}(M) \right) + \left( \bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) - \mathbb{E} \left( \bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) \right) \right) - \left( \bar{\mathcal{L}}_{\Omega,Y}(M) - \mathbb{E} \left( \bar{\mathcal{L}}_{\Omega,Y}(M) \right) \right) \\
&\leq \mathbb{E} \left( \bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) - \bar{\mathcal{L}}_{\Omega,Y}(M) \right) + 2 \sup_{X \in G} \left| \bar{\mathcal{L}}_{\Omega,Y}(X) - \mathbb{E} \left( \bar{\mathcal{L}}_{\Omega,Y}(X) \right) \right| \\
&= -nD(l(M) \| l(\tilde{M})) + 2 \sup_{X \in G} \left| \bar{\mathcal{L}}_{\Omega,Y}(X) - \mathbb{E} \left( \bar{\mathcal{L}}_{\Omega,Y}(X) \right) \right|,
\end{aligned} \tag{A.6.29}$$

where  $G$  is defined in (A.6.26).

Applying Lemma A.6.1, we have that with probability at least  $1 - \frac{c_1}{d_1 + d_2}$

$$\begin{aligned}
& D(l(M) \| l(\tilde{M})) \\
&\leq \frac{2}{n} \tilde{c}_0 L_{\alpha + \delta_1} \left( \alpha \sqrt{r d_1 d_2} + \delta_2 \right) \sqrt{\frac{n(d_1 + d_2)}{d_1 d_2} + \log(d_1 d_2)} + \frac{\delta}{n} \\
&\leq 2\tilde{c}_0 L_{\alpha + \delta_1} \left( \alpha \sqrt{r d_1 d_2} + \delta_2 \right) \sqrt{\frac{d_1 + d_2}{n d_1 d_2}} \sqrt{1 + \frac{(d_1 + d_2) \log(d_1 d_2)}{n}} + \frac{\delta}{n}.
\end{aligned} \tag{A.6.30}$$

Let  $c_0 = 2\tilde{c}_0$  we have the theorem.

### Proof of Lemma A.6.1

Noting that

$$\begin{aligned} \bar{\mathcal{L}}_{\Omega,Y}(X) = \\ \sum_{(i,j)} \mathbb{1}\{(i,j) \in \Omega\} \left( \mathbb{1}\{Y_{i,j} = 1\} \log \left( \frac{l(X_{i,j})}{l(0)} \right) + \mathbb{1}\{Y_{i,j} = -1\} \log \left( \frac{1-l(X_{i,j})}{1-l(0)} \right) \right), \end{aligned} \quad (\text{A.6.31})$$

by symmetrization (i.e Lemma 6.3 in Ledoux and Talagrand (1991)) we have

$$\begin{aligned} \mathbb{E} \left( \sup_{X \in G} |\bar{\mathcal{L}}_{\Omega,Y}(X) - \mathbb{E} \bar{\mathcal{L}}_{\Omega,Y}(X)|^h \right) \leq 2^h \mathbb{E} \left( \sup_{X \in G} \left| \sum_{(i,j)} \zeta_{i,j} \mathbb{1}\{(i,j) \in \Omega\} \right. \right. \\ \left. \left. \left( \mathbb{1}\{Y_{i,j} = 1\} \log \left( \frac{l(X_{i,j})}{l(0)} \right) + \mathbb{1}\{Y_{i,j} = -1\} \log \left( \frac{1-l(X_{i,j})}{1-l(0)} \right) \right) \right|^h \right), \end{aligned} \quad (\text{A.6.32})$$

where  $\zeta_{i,j}$  are i.i.d. Rademacher random variables and the expectation is with respect to  $\Omega$ ,  $Y$  and  $\zeta_{i,j}$ . Next is to apply the contraction principle (i.e. Theorem 4.12 in Ledoux and Talagrand (1991)). By the definition of  $L_{\alpha+\delta_1}$  and definition of  $G$ , we know that

$$\frac{1}{L_{\alpha+\delta_1}} \log \left( \frac{l(x)}{l(0)} \right) \text{ and } \frac{1}{L_{\alpha+\delta_1}} \log \left( \frac{1-l(x)}{1-l(0)} \right)$$

are contractions that vanish at 0 within the domain of any  $X_{i,j}$  such that  $X \in G$ . Invoking contraction principle gives

$$\begin{aligned} \mathbb{E} \left( \sup_{X \in G} |\bar{\mathcal{L}}_{\Omega,Y}(X) - \mathbb{E} \bar{\mathcal{L}}_{\Omega,Y}(X)|^h \right) \\ \leq 2^h (2L_{\alpha+\delta_1})^h \mathbb{E} \left( \sup_{X \in G} \left| \sum_{(i,j)} \zeta_{i,j} \mathbb{1}\{(i,j) \in \Omega\} (\mathbb{1}\{Y_{i,j} = 1\} X_{i,j} - \mathbb{1}\{Y_{i,j} = -1\} X_{i,j}) \right|^h \right) \\ \leq (4L_{\alpha+\delta_1})^h \mathbb{E} \left( \sup_{X \in G} |\langle \Delta_{\Omega} \circ Z \circ Y, X \rangle|^h \right), \end{aligned} \quad (\text{A.6.33})$$

where  $Z$  denotes the matrix with  $(i,j)$ th element being  $\zeta_{i,j}$ ,  $\Delta_{\Omega}$  denotes the indicator matrix for  $\Omega$  such that elements are zero when not in  $\Omega$  and 1 when in  $\Omega$ , and  $\circ$  denotes Hadamard

product. Observing that  $Z \circ Y$  has the same distribution with  $Z$ ,  $(Z, Z \circ Y) \perp\!\!\!\perp \Omega$  and  $\langle A, B \rangle \leq \|A\|_{op} \|B\|_*$ , we have

$$\begin{aligned} \mathbb{E} \left( \sup_{X \in G} |\langle \Delta_\Omega \circ Z \circ Y, X \rangle|^h \right) &= \mathbb{E} \left( \sup_{X \in G} |\langle \Delta_\Omega \circ Z, X \rangle|^h \right) \\ &\leq \mathbb{E} \left( \sup_{X \in G} \|\Delta_\Omega \circ Z\|_{op}^h \|X\|_*^h \right) = \left( \alpha \sqrt{rd_1 d_2} + \delta_2 \right)^h \mathbb{E} \left( \|\Delta_\Omega \circ Z\|_{op}^h \right). \end{aligned} \quad (\text{A.6.34})$$

Observe that  $Z \circ \Delta_\Omega$  is a matrix with i.i.d. symmetric random variables, so according to Theorem 1.1 in Seginer (2000) there is absolute constant  $C$  such that for  $h \leq 2 \log(\max\{d_1, d_2\})$  we have

$$\mathbb{E} \left( \|Z \circ \Delta_\Omega\|^h \right) \leq C \left( \mathbb{E} \left( \max_{1 \leq i \leq d_1} \left( \sum_{j=1}^{d_2} \Delta_{i,j} \right)^{h/2} \right) + \mathbb{E} \left( \max_{1 \leq j \leq d_2} \sum_{i=1}^{d_1} \Delta_{i,j} \right)^{h/2} \right). \quad (\text{A.6.35})$$

Note that  $(\mathbb{E}(|f|^{h/2}))^{2/h}$  is a norm for  $h \geq 2$  and  $(a+b)^{1/h} \leq a^{1/h} + b^{1/h}$ , so we have

$$\begin{aligned} &\left( \|Z \circ \Delta_\Omega\|_{op}^h \right)^{1/h} \\ &\leq C^{1/h} \left( \left( \mathbb{E} \left[ \left( \max_{1 \leq j \leq d_1} \sum_{i=1}^{d_2} \Delta_{i,j} \right)^{h/2} \right] \right)^{1/h} + \left( \mathbb{E} \left[ \left( \max_{1 \leq j \leq d_2} \sum_{i=1}^{d_1} \Delta_{i,j} \right)^{h/2} \right] \right)^{1/h} \right) \\ &\leq C^{1/h} \left( \mathbb{E} \left[ \left( \max_{1 \leq j \leq d_1} \left| \sum_{i=1}^{d_2} \left( \Delta_{i,j} - \frac{n}{d_1 d_2} \right) \right| + \frac{n}{d_1} \right)^{h/2} \right] \right)^{1/h} + \\ &\quad C^{1/h} \left( \mathbb{E} \left[ \left( \max_{1 \leq j \leq d_2} \left| \sum_{i=1}^{d_1} \left( \Delta_{i,j} - \frac{n}{d_1 d_2} \right) \right| + \frac{n}{d_2} \right)^{h/2} \right] \right)^{1/h} \\ &\leq C^{1/h} \left( \sqrt{\frac{n}{d_1}} + \sqrt{\frac{n}{d_2}} \right) + C^{1/h} \left( \mathbb{E} \left[ \left( \max_{1 \leq j \leq d_1} \left| \sum_{i=1}^{d_2} \left( \Delta_{i,j} - \frac{n}{d_1 d_2} \right) \right| \right)^{h/2} \right] \right)^{1/h} + \\ &\quad C^{1/h} \left( \mathbb{E} \left[ \left( \max_{1 \leq j \leq d_2} \left| \sum_{i=1}^{d_1} \left( \Delta_{i,j} - \frac{n}{d_1 d_2} \right) \right| \right)^{h/2} \right] \right)^{1/h}. \end{aligned} \quad (\text{A.6.36})$$

Using Bernstein's inequality, we have for  $t > 0$

$$\mathbb{P} \left( \left| \sum_{j=1}^{d_2} \left( \Delta_{i,j} - \frac{n}{d_1 d_2} \right) \right| > t \right) \leq 2 \exp \left( \frac{-\frac{t^2}{2}}{\frac{n}{d_1} + \frac{t}{3}} \right). \quad (\text{A.6.37})$$

For  $t \geq \frac{6n}{d_1}$ , for each  $i$ , we have

$$\mathbb{P} \left( \left| \sum_{j=1}^{d_2} \left( \Delta_{i,j} - \frac{n}{d_1 d_2} \right) \right| > t \right) \leq 2 \exp(-t) = 2\mathbb{P}(W_i > t), \quad (\text{A.6.38})$$

where  $W_1, \dots, W_{d_1}$  are i.i.d. exponential random variables.

Therefore,

$$\begin{aligned} & \mathbb{E} \left[ \left( \max_{1 \leq j \leq d_2} \left| \sum_{i=1}^{d_1} \left( \Delta_{i,j} - \frac{n}{d_1 d_2} \right) \right| \right)^{h/2} \right] \\ &= \int_0^\infty \mathbb{P} \left( \max_{1 \leq i \leq d_1} \left| \sum_{j=1}^{d_2} \left( \Delta_{i,j} - \frac{n}{d_1 d_2} \right) \right|^h \geq t \right) dt \\ &\leq \left( \frac{6n}{d_1} \right)^h + 2 \int_{\left(\frac{6n}{d_1}\right)^h}^\infty \mathbb{P} \left( \max_{1 \leq i \leq d_1} W_i^h \geq t \right) dt \\ &\leq \left( \frac{6n}{d_1} \right)^h + 2\mathbb{E} \left[ \left( \max_{1 \leq i \leq d_1} W_i \right)^h \right]. \end{aligned} \quad (\text{A.6.39})$$

Note that for i.i.d. exponential random variables  $W_1, \dots, W_{d_1}$  we have

$$\begin{aligned} \mathbb{E} \left[ \left( \max_{1 \leq i \leq d_1} W_i \right)^h \right] &\leq \mathbb{E} \left[ \left( \max_{1 \leq i \leq d_1} W_i^h - \log d_1 \right)_+^h \right] + \log(d_1)^h \\ &\leq 2h! + \log^h(d_1). \end{aligned} \quad (\text{A.6.40})$$

Therefore, we have

$$\begin{aligned}
& \left( \mathbb{E} \left[ \left( \max_{1 \leq j \leq d_2} \left| \sum_{i=1}^{d_1} \left( \Delta_{i,j} - \frac{n}{d_1 d_2} \right) \right| \right)^{h/2} \right] \right)^{1/h} \\
& \leq (1 + \sqrt{6}) \sqrt{\frac{n}{d_1}} + 2^{1/2h} \left( \sqrt{d_1} + 2^{1/2h} \sqrt{h} \right) \\
& \leq (1 + \sqrt{6}) \sqrt{\frac{n}{d_1}} + (2 + \sqrt{2}) \sqrt{\log(d_1 + d_2)},
\end{aligned} \tag{A.6.41}$$

where in the last inequality we use  $h = \log(d_1 + d_2) \geq 1$ . It's easy to check that this choice of  $h$  satisfies the condition required for getting Inequality (A.6.35).

Using similar argument to bound the third term in the right hand side of the last inequality in Inequality (A.6.36), we have

$$\begin{aligned}
\left( \mathbb{E} \left[ \|\Delta_\Omega \circ Z\|_{op}^h \right] \right)^{1/h} & \leq C^{1/h} \left( (1 + \sqrt{6}) \left( \sqrt{\frac{n}{d_1}} + \sqrt{\frac{n}{d_2}} \right) + (4 + 2\sqrt{2}) \sqrt{\log(d_1 + d_2)} \right) \\
& \leq C^{1/h} \sqrt{\frac{n}{d_1} + \frac{n}{d_2} + \log(d_1 + d_2)} \sqrt{(1 + \sqrt{6})^2 + 4 + 2\sqrt{2}} \\
& < 9C^{1/h} \sqrt{\frac{n}{d_1} + \frac{n}{d_2} + \log(d_1 + d_2)}
\end{aligned} \tag{A.6.42}$$

Combing Inequality (A.6.33), (A.6.34), (A.6.42), we have

$$\begin{aligned}
& \left( \mathbb{E} \left( \sup_{X \in G} |\bar{\mathcal{L}}_{\Omega, Y}(X) - \mathbb{E} \bar{\mathcal{L}}_{\Omega, Y}(X)|^h \right) \right)^{1/h} \\
& \leq 4L_{\alpha+\delta_1} \left( \alpha \sqrt{r d_1 d_2} + \delta_2 \right) \times 9C^{1/h} \sqrt{\frac{n}{d_1} + \frac{n}{d_2} + \log(d_1 + d_2)}.
\end{aligned} \tag{A.6.43}$$

Let  $t = 4L_{\alpha+\delta_1} (\alpha \sqrt{r d_1 d_2} + \delta_2) \times 9 \sqrt{\frac{n}{d_1} + \frac{n}{d_2} + \log(d_1 + d_2)} \times e$ . Then we know that

$$\begin{aligned}
& \mathbb{P} \left( \sup_{X \in G} |\bar{\mathcal{L}}_{\Omega, Y}(X) - \mathbb{E} \bar{\mathcal{L}}_{\Omega, Y}(X)| \geq t \right) \\
& \leq C \exp(-h) = \frac{C}{d_1 + d_2}.
\end{aligned} \tag{A.6.44}$$

Set  $\tilde{c}_0 = 4 \times 9 \times e$ ,  $c_1 = C$ , we have the lemma.

#### A.6.5. Proof of Theorem 4.4.1

Denote  $\mathbf{A}_{it}$  to be the matrix with element  $(i, t)$  being 1 and others being 0. Denote  $\varepsilon_{it}$  to be the  $(i, t)$ -th element of  $\boldsymbol{\varepsilon}$ . Let  $\boldsymbol{\mathfrak{E}} = \sum_{(i,t) \in \mathcal{O}} \varepsilon_{it} \mathbf{A}_{it}$ . And  $\|\cdot\|_{op}$  denotes the operator norm (i.e. the largest singular value).

The overall structure of the proof is similar to that in Athey et al. (2021), we have three main lemmas, which we will prove later. The first two lemmas primarily show how the optimization error comes in, and for the third lemma, we do the statistical analysis differently and have improved rate than that in Athey et al. (2021). The three lemmas are as follows.

**Lemma A.6.2.** *For all  $\lambda \geq 3\|\boldsymbol{\mathfrak{E}}\|_{op}/|\mathcal{O}|$ ,*

$$\sum_{(i,t) \in \mathcal{O}} \frac{\langle \mathbf{A}_{it}, \mathbf{L}^* - \tilde{\mathbf{L}} \rangle^2}{|\mathcal{O}|} \leq 10\sqrt{2R}\lambda \|\mathbf{L}^* - \tilde{\mathbf{L}}\|_F + 6\delta. \quad (\text{A.6.45})$$

**Lemma A.6.3.** *With probability at least  $1 - \frac{1}{(N+T)^2}$ , we have*

$$\|\boldsymbol{\mathfrak{E}}\|_{op} \leq 4\sigma \max\{\sqrt{N \log(N+T)}, 8\sqrt{T} \log^{\frac{3}{2}}(N+T)\} + \sigma. \quad (\text{A.6.46})$$

**Lemma A.6.4.** *Suppose  $\lambda \geq 3\|\boldsymbol{\mathfrak{E}}\|_{op}/|\mathcal{O}|$ .*

*Then when  $\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2 \geq 132(L_{max} + \delta_1)^2 \times T \log(N+T) \frac{1}{p_c}$ ,*

$$\begin{aligned} \mathbb{P}_\pi \left( \frac{\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2 p_c}{6} > \sum_{(i,t) \in \mathcal{O}} \langle \mathbf{A}_{it}, \tilde{\mathbf{L}} - \mathbf{L}^* \rangle^2 + 3648 \frac{72R}{p_c} (\sqrt{N} + \sqrt{T})^2 (4(L_{max} + \delta_1)^2) \right. \\ \left. + \frac{432\delta(L_{max} + \delta_1)}{\lambda} (\sqrt{N} + \sqrt{T}) \right) \leq \frac{1}{(N+T)^3} \end{aligned} \quad (\text{A.6.47})$$

Therefore, when  $\lambda \geq \frac{12\sigma \max\{\sqrt{N \log(N+T)}, 8\sqrt{T} \log^{\frac{3}{2}}(N+T)\} + 3\sigma}{|\mathcal{O}|}$ , if  $\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2 \geq 132(L_{max} +$

$\delta_1)^2 \times T \log(N+T) \frac{1}{p_c}$ , then with probability at least  $1 - \frac{2}{(N+T)^2}$ ,

$$\begin{aligned}
\frac{\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2 p_c}{6} &\leq \sum_{(i,t) \in \mathcal{O}} \langle A_{it}, \tilde{\mathbf{L}} - \mathbf{L}^* \rangle^2 + 3648 \frac{72R}{p_c} (\sqrt{N} + \sqrt{T})^2 (4(L_{max} + \delta_1)^2) \\
&\quad + \frac{432\delta(L_{max} + \delta_1)}{\lambda} (\sqrt{N} + \sqrt{T}) \\
&\leq 10\sqrt{2R}(\lambda|\mathcal{O}|) \|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F + 6\delta|\mathcal{O}| + 3648 \frac{72R}{p_c} (\sqrt{N} + \sqrt{T})^2 (4(L_{max} + \delta_1)^2) \\
&\quad + \frac{432\delta(L_{max} + \delta_1)}{\lambda} (\sqrt{N} + \sqrt{T}).
\end{aligned} \tag{A.6.48}$$

Note that

$$10\sqrt{2R}(\lambda|\mathcal{O}|) \|\mathbf{L}^* - \tilde{\mathbf{L}}\|_F \leq \frac{12 \times 200R(\lambda|\mathcal{O}|)^2}{p_c} + \frac{\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2 p_c}{12}, \tag{A.6.49}$$

and  $|\mathcal{O}| \leq NT$ .

We take  $\lambda = \frac{13\sigma \max\{\sqrt{N \log(N+T)}, 8\sqrt{T} \log^{\frac{3}{2}}(N+T)\}}{|\mathcal{O}|}$ .

Move the  $\frac{\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2 p_c}{12}$  term from the right hand side to the left hand side and then divide both sides with  $\frac{p_c NT}{12}$ , we have there are constants  $q_0, q_1, q_2$ , such that

$$\begin{aligned}
\frac{\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2}{NT} &\leq q_0 \frac{R\sigma^2}{p_c^2} \frac{(N+T) \log^3(N+T)}{NT} + \frac{72}{p_c} \delta + q_1 \frac{\delta(L_{max} + \delta_1)}{\sigma p_c} \frac{1}{NT} \\
&\quad + q_2 \frac{R(L_{max} + \delta_1)^2}{p_c^2} \frac{N+T}{NT}.
\end{aligned} \tag{A.6.50}$$



### Proof of Lemma A.6.2

By the definition of  $\tilde{\mathbf{L}}$ ,  $\hat{\mathbf{L}}$ ,  $\mathbf{L}^*$ , we have

$$\begin{aligned}
& \sum_{(i,t) \in \mathcal{O}} \frac{\langle Y_{it} - \tilde{\mathbf{L}} \rangle^2}{|\mathcal{O}|} + \lambda \|\tilde{\mathbf{L}}\|_* \\
& \leq \sum_{(i,t) \in \mathcal{O}} \frac{\langle Y_{it} - \hat{\mathbf{L}} \rangle^2}{|\mathcal{O}|} + \lambda \|\hat{\mathbf{L}}\|_* + \delta \\
& \leq \sum_{(i,t) \in \mathcal{O}} \frac{\langle Y_{it} - \mathbf{L}^* \rangle^2}{|\mathcal{O}|} + \lambda \|\mathbf{L}^*\|_* + \delta.
\end{aligned} \tag{A.6.51}$$

Therefore, we have

$$\sum_{(i,t) \in \mathcal{O}} \frac{\langle \mathbf{L}^* - \tilde{\mathbf{L}}, \mathbf{A}_{it} \rangle^2}{|\mathcal{O}|} + 2 \sum_{(i,t) \in \mathcal{O}} \frac{\varepsilon_{it} \langle \mathbf{L}^* - \tilde{\mathbf{L}}, \mathbf{A}_{it} \rangle}{|\mathcal{O}|} \leq \lambda \|\mathbf{L}^*\|_* - \lambda \|\tilde{\mathbf{L}}\|_* + \delta. \tag{A.6.52}$$

Denoting  $\Delta = \mathbf{L}^* - \tilde{\mathbf{L}}$ , Inequality (A.6.52) becomes

$$\begin{aligned}
\sum_{(i,t) \in \mathcal{O}} \frac{\langle \Delta, \mathbf{A}_{it} \rangle^2}{|\mathcal{O}|} & \leq -\frac{2}{|\mathcal{O}|} \langle \Delta, \mathfrak{E} \rangle + \lambda \|\mathbf{L}^*\|_* - \lambda \|\tilde{\mathbf{L}}\|_* + \delta \\
& \leq \frac{2}{|\mathcal{O}|} \|\Delta\|_* \|\mathfrak{E}\|_{op} + \lambda \|\mathbf{L}^*\|_* - \lambda \|\tilde{\mathbf{L}}\|_* + \delta \\
& \leq \frac{5}{3} \lambda \|\Delta\|_* + \delta,
\end{aligned} \tag{A.6.53}$$

the inequalities in which are due to the duality of operator norm and nuclear norm, and the range of  $\lambda$ .

Now we state the following lemma, which is proved later in this section.

**Lemma A.6.5.** *Let  $\Delta = \mathbf{L}^* - \tilde{\mathbf{L}}$  for  $\lambda \geq 3\|\mathfrak{E}\|_{op}/|\mathcal{O}|$ . Then there exist a decomposition  $\Delta = \Delta_1 + \Delta_2$  such that*

1.  $\langle \Delta_1, \Delta_2 \rangle = 0$ ,

$$2. \text{rank}(\Delta_1) \leq 2R,$$

$$3. \|\Delta_2\|_* \leq 5\|\Delta_1\|_* + \frac{3\delta}{\lambda}.$$

Now, invoking the decomposition  $\Delta = \Delta_1 + \Delta_2$ , we have

$$\|\Delta\|_* \leq 6\|\Delta_1\|_* + \frac{3\delta}{\lambda} \leq 6\sqrt{2R}\|\Delta_1\|_F + \frac{3\delta}{\lambda} \leq 6\sqrt{2R}\|\Delta\|_F + \frac{3\delta}{\lambda}. \quad (\text{A.6.54})$$

Plugging Inequality (A.6.54) back to Inequality (A.6.53), we have

$$\sum_{(i,t) \in \mathcal{O}} \frac{\langle \Delta, \mathbf{A}_{it} \rangle^2}{|\mathcal{O}|} \leq 10\sqrt{2R}\lambda\|\Delta\|_F + 6\delta. \quad (\text{A.6.55})$$

**Proof of Lemma A.6.5.** Let  $\mathbf{L}^* = \mathbf{U}_{N \times R} \mathbf{S}_{R \times R} (\mathbf{V}_{T \times R})^T$  be the singular value decomposition for the at most rank  $R$  matrix  $\mathbf{L}^*$ . Let  $\mathbf{P}_\mathbf{U} = \mathbf{U}\mathbf{U}^T$ ,  $\mathbf{P}_{\mathbf{U}^\perp} = \mathbf{U}^\perp(\mathbf{U}^\perp)^T$ ,  $\mathbf{P}_\mathbf{V} = \mathbf{V}\mathbf{V}^T$ ,  $\mathbf{P}_{\mathbf{V}^\perp} = \mathbf{V}^\perp(\mathbf{V}^\perp)^T$ . Let  $\Delta_2 = \mathbf{P}_{\mathbf{U}^\perp} \Delta \mathbf{P}_{\mathbf{V}^\perp}$ ,  $\Delta_1 = \Delta - \Delta_2$ .

It's easy to see that  $\mathbf{P}_\mathbf{U} + \mathbf{P}_{\mathbf{U}^\perp} = \mathbf{I}_\mathbf{N}$  and  $\mathbf{P}_\mathbf{V} + \mathbf{P}_{\mathbf{V}^\perp} = \mathbf{I}_\mathbf{T}$ .

Now we check the three claims for Lemma A.6.5.

$$\begin{aligned} \langle \Delta_1, \Delta_2 \rangle &= \langle \Delta - \mathbf{P}_{\mathbf{U}^\perp} \Delta \mathbf{P}_{\mathbf{V}^\perp}, \mathbf{P}_{\mathbf{U}^\perp} \Delta \mathbf{P}_{\mathbf{V}^\perp} \rangle \\ &= \langle \mathbf{P}_\mathbf{U} \Delta + \mathbf{P}_{\mathbf{U}^\perp} \Delta \mathbf{P}_\mathbf{V}, \mathbf{P}_{\mathbf{U}^\perp} \Delta \mathbf{P}_{\mathbf{V}^\perp} \rangle \\ &= 0. \end{aligned} \quad (\text{A.6.56})$$

$$\text{rank}(\Delta_1) = \text{rank}(\mathbf{P}_\mathbf{U} \Delta + \mathbf{P}_{\mathbf{U}^\perp} \Delta \mathbf{P}_\mathbf{V}) \leq \text{rank}(\mathbf{P}_\mathbf{U} \Delta) + \text{rank}(\mathbf{P}_{\mathbf{U}^\perp} \Delta \mathbf{P}_\mathbf{V}) \leq 2R. \quad (\text{A.6.57})$$

For the third one, note that

$$\begin{aligned}\langle \Delta_2, \mathbf{L}^* \rangle &= \langle \mathbf{P}_{\mathbf{U}^\perp} \Delta \mathbf{P}_{\mathbf{V}^\perp}, \mathbf{U}_{N \times R} \mathbf{S}_{R \times R} (\mathbf{V}_{T \times R})^T \rangle \\ &= 0.\end{aligned}\tag{A.6.58}$$

And Inequality (A.6.53) implies that

$$\begin{aligned}\lambda \left( \|\tilde{\mathbf{L}}\|_* - \|\mathbf{L}^*\|_* \right) &\leq \frac{2}{|\mathcal{O}|} \|\Delta\|_* \|\mathfrak{E}\|_{op} + \delta \\ &\leq \frac{2}{3} \lambda \|\Delta\|_* + \delta \leq \frac{2}{3} \lambda (\|\Delta_1\|_* + \|\Delta_2\|_*) + \delta.\end{aligned}\tag{A.6.59}$$

The main part of the left hand sided is lower bound by

$$\begin{aligned}\|\tilde{\mathbf{L}}\|_* - \|\mathbf{L}^*\|_* &= \|\mathbf{L}^* - \Delta_1 - \Delta_2\|_* - \|\mathbf{L}^*\|_* \geq \|\mathbf{L}^* - \Delta_1\|_* - \|\Delta_2\|_* - \|\mathbf{L}^*\|_* \\ &= \|\mathbf{L}^*\|_* + \|\Delta_1\|_* - \|\Delta_2\|_* - \|\mathbf{L}^*\|_* = \|\Delta_1\|_* - \|\Delta_2\|_*.\end{aligned}\tag{A.6.60}$$

Combining Inequality (A.6.59) and (A.6.60), we have

$$\|\Delta_2\|_* \leq 5\|\Delta_1\|_* + \frac{3\delta}{\lambda}.\tag{A.6.61}$$

### Proof of Lemma A.6.3

The proof is very similar to that of lemma 2 in Athey et al. (2021), but our task is to write out the constants explicitly and have the bound as tight as possible.

Although the major parts are very similar, we still write out all the steps for completeness.

The goal is to invoke matrix version Bernstein inequality, a proof of which is in Tropp (2012). Proposition A.6.1 states the matrix version Bernstein inequality.

**Proposition A.6.1** (Matrix Bernstein Inequality). *Let  $\mathbf{Z}_1, \dots, \mathbf{Z}_N$  be independent matrices in  $\mathbb{R}^{d_1 \times d_2}$  such that  $\mathbb{E}[\mathbf{Z}_i] = \mathbf{0}$  and  $\|\mathbf{Z}_i\|_{op} \leq D$  almost surely for all  $i \in [N]$ . Let  $\sigma_Z$  be such*

that

$$\sigma_Z^2 \geq \max \left\{ \left\| \sum_{i=1}^N \mathbb{E}[\mathbf{Z}_i \mathbf{Z}_i^T] \right\|_{op}, \left\| \sum_{i=1}^N \mathbb{E}[\mathbf{Z}_i^T \mathbf{Z}_i] \right\|_{op} \right\}.$$

Then, for any  $\alpha \geq 0$ ,

$$\mathbb{P} \left\{ \left\| \sum_{i=1}^N \mathbf{Z}_i \right\|_{op} \geq \alpha \right\} \leq (d_1 + d_2) \exp \left[ \frac{-\alpha^2}{2\sigma_Z^2 + (2D\alpha)/3} \right]. \quad (\text{A.6.62})$$

Same as the notations in Athey et al. (2021), define independent random matrices  $\mathbf{B}_1, \dots, \mathbf{B}_N$  as follows. For  $1 \leq i \leq N$ , define

$$\mathbf{B}_i = \sum_{t=1}^{t_i} \varepsilon_{it} \mathbf{A}_{it}.$$

Then,  $\mathfrak{E} = \sum_{i=1}^N \mathbf{B}_i$  and  $\mathbb{E}[\mathbf{B}_i] = 0$ . Define the bound  $D = C_2 \sigma \sqrt{\log(N+T)}$  for a constant  $C_2$  that we will specify later. For each  $(i, t) \in \mathcal{O}$ , let  $\bar{\varepsilon}_{it} = \varepsilon_{it} \mathbb{1}\{|\varepsilon_{it}| \leq D\}$ . For  $1 \leq i \leq N$ , let  $\bar{\mathbf{B}}_i = \sum_{t=1}^{t_i} \bar{\varepsilon}_{it} \mathbf{A}_{it}$ .

The  $\sigma$ -sub-Gaussian implies

$$\begin{aligned} \mathbb{P}(|\varepsilon_{it}| \geq t) &= 2 \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \\ &\leq \frac{2\sigma}{\sqrt{2\pi}} \int_{\frac{t^2}{2\sigma^2}}^\infty \exp(-x) dx = \frac{2\sigma}{\sqrt{2\pi}t} \exp\left(-\frac{t^2}{2\sigma^2}\right). \end{aligned} \quad (\text{A.6.63})$$

Therefore, for  $\alpha > 0$ ,

$$\begin{aligned} \mathbb{P}\{\|\mathfrak{E}\|_{op} \geq \alpha\} &\leq \mathbb{P}\left\{ \left\| \sum_{i=1}^B \bar{\mathbf{B}}_i \right\|_{op} \geq \alpha \right\} + \sum_{(i,t) \in \mathcal{O}} \mathbb{P}(|\varepsilon_{it}| \geq D) \\ &\leq \mathbb{P}\left\{ \left\| \sum_{i=1}^B \bar{\mathbf{B}}_i \right\|_{op} \geq \alpha \right\} + |\mathcal{O}| \times \frac{2\sigma}{\sqrt{2\pi}D} \exp\left(-\frac{D^2}{2\sigma^2}\right) \\ &\leq \mathbb{P}\left\{ \left\| \sum_{i=1}^B \bar{\mathbf{B}}_i \right\|_{op} \geq \alpha \right\} + \sqrt{\frac{2}{\pi}} \frac{NT}{C_2 \sqrt{\log(N+T)}} (N+T)^{-\frac{C_2^2}{2}}. \end{aligned} \quad (\text{A.6.64})$$

For  $1 \leq i \leq N$ , define  $\mathbf{Z}_i = \bar{\mathbf{B}}_i - \mathbb{E}[\bar{\mathbf{B}}_i]$ . Then,

$$\begin{aligned} \left\| \sum_{i=1}^N \bar{\mathbf{B}}_i \right\|_{op} &\leq \left\| \sum_{i=1}^N \mathbf{Z}_i \right\|_{op} + \left\| \mathbb{E} \left[ \sum_{i=1}^N \bar{\mathbf{B}}_i \right] \right\|_{op} \\ &\leq \left\| \sum_{i=1}^N \mathbf{Z}_i \right\|_{op} + \left\| \mathbb{E} \left[ \sum_{i=1}^N \bar{\mathbf{B}}_i \right] \right\|_F \leq \left\| \sum_{i=1}^N \mathbf{Z}_i \right\|_{op} + \sqrt{NT} \left\| \mathbb{E} \left[ \sum_{i=1}^N \bar{\mathbf{B}}_i \right] \right\|_{\infty}. \end{aligned} \quad (\text{A.6.65})$$

Further,

$$\begin{aligned} |\mathbb{E}[\bar{\varepsilon}_{it}]| &= |\mathbb{E}[\varepsilon_{it} \mathbb{1}\{|\varepsilon_{it}| \leq D\}]| = |\mathbb{E}[\varepsilon_{it} \mathbb{1}\{|\varepsilon_{it}| \geq D\}]| \leq \sqrt{\mathbb{E}[\varepsilon_{it}^2] \mathbb{P}(|\varepsilon_{it}| \geq D)} \\ &\leq \sigma \sqrt{\sqrt{\frac{2}{\pi}} \frac{1}{C_2 \sqrt{\log(N+T)}} (N+T)^{-\frac{c_2^2}{2}}}. \end{aligned} \quad (\text{A.6.66})$$

Therefore,

$$\sqrt{NT} \left\| \mathbb{E} \left[ \sum_{i=1}^N \bar{\mathbf{B}}_i \right] \right\|_{\infty} \leq \sigma \sqrt{\sqrt{\frac{2}{\pi}} \frac{NT}{C_2 \sqrt{\log(N+T)}} (N+T)^{-\frac{c_2^2}{2}}}. \quad (\text{A.6.67})$$

Note that  $\|\mathbf{Z}_i\|_{op} \leq 2D\sqrt{T}$  for all  $1 \leq i \leq N$ . The only step left for invoking Proposition A.6.1 is to calculate  $\sigma_Z$  in there.

Recall that  $\mathbb{E}[(\bar{\varepsilon}_{it} - \mathbb{E}[\bar{\varepsilon}_{it}])^2] \leq \sigma^2$ .

We have

$$\begin{aligned} \left\| \sum_{i=1}^N \mathbb{E}[\mathbf{Z}_i \mathbf{Z}_i^T] \right\|_{op} &\leq \max_{1 \leq i \leq N} \left( \mathbb{E} \left( \sum_{t:(i,t) \in \mathcal{O}} \mathbb{E}[(\bar{\varepsilon}_{it} - \mathbb{E}[\bar{\varepsilon}_{it}])^2] \right) \right) \\ &\leq \sigma^2 T, \end{aligned} \quad (\text{A.6.68})$$

and

$$\begin{aligned} \left\| \sum_{i=1}^N \mathbb{E}[\mathbf{Z}_i^T \mathbf{Z}_i] \right\|_{op} &\leq \sigma^2 \max_{1 \leq i \leq T} \sum_{j=1}^N \mathbb{P}((j, i) \in \mathcal{O}) \\ &\leq \sigma^2 N. \end{aligned} \quad (\text{A.6.69})$$

The first inequality in Inequality (A.6.69) is due to  $\mathbb{E} \left\{ (\bar{\varepsilon}_{it} - \mathbb{E}[\bar{\varepsilon}_{it}])(\bar{\varepsilon}_{js} - \mathbb{E}[\bar{\varepsilon}_{js}]) \middle| \mathcal{O} \right\} = 0$  for  $(i, t) \neq (j, s)$ .

Therefore  $\sigma_Z^2 = \sigma^2 \max\{N, T\}$  is a possible choice. Invoking Proposition A.6.1, we have

$$\mathbb{P} \left\{ \left\| \sum_{i=1}^N \mathbf{Z}_i \right\|_{op} \geq \alpha \right\} \leq (N + T) \exp \left[ \frac{-\alpha^2}{2\sigma^2 \max\{N, T\} + (4C_2\sigma\sqrt{\log(N+T)T\alpha})/3} \right]. \quad (\text{A.6.70})$$

Taking  $C_2 = 3$ ,  $\alpha = \max\{4\sigma\sqrt{\max\{N, T\}}\sqrt{\log(N+T)}, 32T^{\frac{1}{2}}(\log(N+T))^{\frac{3}{2}}\sigma\}$ .

Combing Inequalities (A.6.64), (A.6.65), (A.6.67), (A.6.70), we have with probability at least  $1 - \frac{1}{2(N+T)^2} - \frac{1}{2(N+T)^3}$

$$\|\mathfrak{E}\|_{op} \leq 4\sigma \max\{\sqrt{\max\{N, T\}}\sqrt{\log N + T}, 8T^{\frac{1}{2}}(\log(N+T))^{\frac{3}{2}}\} + \sigma. \quad (\text{A.6.71})$$

#### Proof of Lemma A.6.4

We define some additional notation here, which are similar to the additional notation in Athey et al. (2021). Given observation set  $\mathcal{O}$ , for every  $N$  by  $T$  matrix  $\mathbf{M}$ , define  $\mathcal{X}_{\mathcal{O}}(\mathbf{M})$  and  $\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})$  as follows.

$$\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M}) = [\langle \mathbf{A}_{i1}, \mathbf{M} \rangle, \dots, \langle \mathbf{A}_{iT}, \mathbf{M} \rangle]^T, \quad (\text{A.6.72})$$

$$\mathcal{X}_{\mathcal{O}}(\mathbf{M}) = \begin{bmatrix} \mathcal{X}_{\mathcal{O}}^{(1)}(M) \\ \cdot \\ \cdot \\ \cdot \\ \mathcal{X}_{\mathcal{O}}^{(N)}(M) \end{bmatrix}. \quad (\text{A.6.73})$$

Define a  $L_{(\text{II})}^2$  norm of  $\mathbf{M}$  as

$$\|\mathbf{M}\|_{L_{(\text{II})}^2} = \sqrt{\mathbb{E}_{\pi} (\|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|_2^2)}, \quad (\text{A.6.74})$$

where  $\mathbb{E}_{\pi}$  is taking expectation with respect to the distribution of  $\mathcal{O}$ .

Define the constraint set as

$$\mathcal{C}(\theta, \eta) = \left\{ \mathbf{M} \in \mathbb{R}^{N \times T} \mid \|\mathbf{M}\|_{\infty} \leq 1, \|\mathbf{M}\|_{L_{(\text{II})}^2}^2 \geq \theta, \|\mathbf{M}\|_* \leq \sqrt{\eta} \|\mathbf{M}\|_F + \frac{3\delta}{2\lambda(L_{\max} + \delta_1)} \right\}. \quad (\text{A.6.75})$$

Then according to Lemma A.6.3, we know that either

$$\frac{\tilde{\mathbf{L}} - \mathbf{L}^*}{2(L_{\max} + \delta_1)} \in \mathcal{C}(\theta, (6\sqrt{2R})^2)$$

or

$$\left\| \frac{\tilde{\mathbf{L}} - \mathbf{L}^*}{2(L_{\max} + \delta_1)} \right\|_{L_{(\text{II})}^2}^2 \leq \theta.$$

Observe that  $\left\| \frac{\tilde{\mathbf{L}} - \mathbf{L}^*}{2(L_{\max} + \delta_1)} \right\|_{L_{(\text{II})}^2}^2 \leq \theta$  implies  $\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2 \leq \frac{4(L_{\max} + \delta_1)^2 \theta}{p_c}$ .

We set  $\theta = 33T \log(N + T)$ .

Let  $\xi > 1$  be a number that we will specify later. Define

$$\mathcal{C}(\theta, \eta, \rho) = \left\{ \mathbf{M} \in \mathcal{C}(\theta, \eta) \mid \rho \leq \|\mathbf{M}\|_{L_{(\text{II})}^2}^2 \leq \rho \xi \right\}. \quad (\text{A.6.76})$$

We state a lemma that we will prove later in this section.

**Lemma A.6.6.** *Suppose  $\xi > 1$ . Let*

$$Z_\rho = \frac{1}{T} \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \{ \|\mathbf{M}\|_{L^2_{(\Pi)}}^2 - \|\mathcal{X}_\mathcal{O}(\mathbf{M})\|^2 \}, \quad (\text{A.6.77})$$

then for  $t > 0$ ,

$$\begin{aligned} P \left( Z_\rho \geq \frac{48}{T} \left( \sqrt{\frac{\eta\rho\xi}{p_c}} + \frac{3\delta}{2\lambda(L_{max} + \delta_1)} \right) (\sqrt{N} + \sqrt{T}) + t \right) \leq \\ \exp \left( -\frac{t}{4} \log \left( 1 + 2 \log \left( 1 + \frac{t}{\frac{96}{T} \left( \sqrt{\frac{\eta\rho\xi}{p_c}} + \frac{3\delta}{2\lambda(L_{max} + \delta_1)} \right) (\sqrt{N} + \sqrt{T}) + \frac{\rho\xi}{T}} \right) \right) \right). \end{aligned} \quad (\text{A.6.78})$$

According to Lemma A.6.6, if we set

$$\begin{aligned} t_0 &= \frac{1}{4T} \left( 96 \left( \sqrt{\frac{\eta\rho\xi}{p_c}} + \frac{3\delta}{2\lambda(L_{max} + \delta_1)} \right) (\sqrt{N} + \sqrt{T}) + \rho\xi \right), \\ t &= \frac{1}{T} \left( \frac{\rho\xi}{4} + \frac{\rho}{4} + \frac{4 * 144\eta\xi}{p_c} (\sqrt{N} + \sqrt{T})^2 + \frac{72\delta}{2\lambda(L_{max} + \delta_1)} (\sqrt{N} + \sqrt{T}) \right), \end{aligned} \quad (\text{A.6.79})$$

then we know that  $t_0 \leq t$ , so we have

$$\begin{aligned} \mathbb{P} \left( \mathbf{M} \in \mathcal{C}(\theta, \eta, \rho), \|\mathbf{M}\|_{L^2_{(\Pi)}}^2 \geq \|\mathcal{X}_\mathcal{O}(\mathbf{M})\|^2 + 48\|\mathbf{M}\|_{L^2_{(\Pi)}} \sqrt{\frac{\eta\xi}{p_c}} (\sqrt{N} + \sqrt{T}) \right. \\ \left. + \frac{144\delta}{2\lambda(L_{max} + \delta_1)} (\sqrt{N} + \sqrt{T}) + \right. \\ \left. \frac{\|\mathbf{M}\|_{L^2_{(\Pi)}}^2}{4} + \frac{\|\mathbf{M}\|_{L^2_{(\Pi)}}^2 \xi}{4} + \frac{576\xi\eta}{p_c} (\sqrt{N} + \sqrt{T})^2 + \frac{72\delta}{2\lambda(L_{max} + \delta_1)} (\sqrt{N} + \sqrt{T}) \right) \\ \leq \exp \left( -\frac{1}{22T} \rho(\xi + 1) - \frac{10\eta\xi}{p_c} \right). \end{aligned} \quad (\text{A.6.80})$$

Given that

$$\bigcup_{i=0}^{\infty} \mathcal{C}(\theta, \eta, \theta\xi^i) = \mathcal{C}(\theta, \eta) \quad (\text{A.6.81})$$



we have

$$\begin{aligned}
& \mathbb{P}\left(\mathbf{M} \in \mathcal{C}(\theta, \eta), \|\mathbf{M}\|_{L^2_{(\text{II})}}^2 \geq \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2 + 48\|\mathbf{M}\|_{L^2_{(\text{II})}} \sqrt{\frac{\eta\xi}{p_c}}(\sqrt{N} + \sqrt{T}) \right. \\
& \quad \left. + \frac{144\delta}{2\lambda(L_{\max} + \delta_1)}(\sqrt{N} + \sqrt{T}) + \right. \\
& \quad \left. \frac{\|\mathbf{M}\|_{L^2_{(\text{II})}}^2}{4} + \frac{\|\mathbf{M}\|_{L^2_{(\text{II})}}^2 \xi}{4} + \frac{576\xi\eta}{p_c}(\sqrt{N} + \sqrt{T})^2 + \frac{72\delta}{2\lambda(L_{\max} + \delta_1)}(\sqrt{N} + \sqrt{T}) \right) \\
& \leq \exp\left(-\frac{\theta}{11T} - 10\eta\xi\right) \frac{1}{1 - \exp\left(-\frac{\theta(\xi-1)}{22T}\right)}.
\end{aligned} \tag{A.6.82}$$

Note that  $48\|\mathbf{M}\|_{L^2_{(\text{II})}} \sqrt{\frac{\eta\xi}{p_c}}(\sqrt{N} + \sqrt{T}) \leq \frac{\|\mathbf{M}\|_{L^2_{(\text{II})}}^2}{4} + 2304\frac{\eta}{p_c}(\sqrt{N} + \sqrt{T})^2$ , and  $\|\mathbf{M}\|_{L^2_{(\text{II})}}^2 \geq p_c\|\mathbf{M}\|_F^2$ , if we set  $\xi = \frac{4}{3}$ , we have

$$\begin{aligned}
& \mathbb{P}\left(\frac{p_c\|\mathbf{M}\|_F^2}{6} \geq \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2 + 3648\frac{\eta}{p_c}(\sqrt{N} + \sqrt{T})^2 + \frac{216\delta}{2\lambda(L_{\max} + \delta_1)}(\sqrt{N} + \sqrt{T}) \right) \\
& \leq P\left(\frac{\|\mathbf{M}\|_{L^2_{(\text{II})}}^2}{6} \geq \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2 + 3648\frac{\eta}{p_c}(\sqrt{N} + \sqrt{T})^2 + \frac{216\delta}{2\lambda(L_{\max} + \delta_1)}(\sqrt{N} + \sqrt{T}) \right) \\
& \leq \exp\left(-\frac{\theta}{11T}\right) \frac{\exp(-10\eta)}{1 - \exp\left(-\frac{\theta}{66T}\right)}.
\end{aligned} \tag{A.6.83}$$

Note that we set  $\theta = 33T \log(N + T)$  and we have  $\frac{\tilde{\mathbf{L}} - \mathbf{L}^*}{2(L_{\max} + \delta_1)} \in \mathcal{C}(\theta, \eta)$  for  $\eta = 72R$  according to Lemma A.6.2, so we have the Lemma A.6.4.

**Proof of Lemma A.6.6** The goal here is to invoke theorem 12.9 of Boucheron et al. (2013).

Note that  $\|\mathbf{M}\|_{L^2_{(\text{II})}}^2 - \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2$  has its rows independent and

$$\mathbb{E}_{\pi}\left(\mathbb{E}_{\pi}(\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|^2) - \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|^2\right) = 0$$

for all  $1 \leq i \leq N$ . Although theorem 12.9 in Boucheron et al. (2013) requires countability

of the index set, given that  $\mathcal{C}(\theta, \eta, \rho)$  is bounded, compact, and  $\|\mathbf{M}\|_{L^2(\Pi)}^2 - \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2$  is uniformly continuous for all  $\mathcal{O}$ , theorem 12.9 is applicable to our setting. The next steps are to find a bound for  $\mathbb{E}(Z_\rho)$  and

$$\sigma^2 = \frac{1}{T^2} \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \sum_{i=1}^N \text{Var}(\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|^2).$$

For  $\sigma^2$ , we have

$$\begin{aligned} \sigma^2 &\leq \frac{1}{T^2} \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \sum_{i=1}^N \mathbb{E}(\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^4) \\ &\leq \frac{1}{T} \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \sum_{i=1}^N \mathbb{E}(\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2) \leq \frac{\rho\xi}{T}. \end{aligned} \tag{A.6.84}$$

For  $\mathbb{E}(Z_\rho)$ , suppose  $\zeta_i$  ( $i = 1, \dots, N$ ) are i.i.d. Rademacher variable, then we have, for any

$\tau$

$$\begin{aligned} \mathbb{E}(Z_\rho) &\stackrel{(i)}{\leq} \frac{1}{T} \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \left\{ \left| \|\mathbf{M}\|_{L^2(\Pi)}^2 - \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2 \right| \right\} \\ &\stackrel{(ii)}{\leq} \frac{2}{T} \mathbb{E} \left[ \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \left| \sum_{i=1}^N \zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2 \right| \right] \\ &\stackrel{(iii)}{\leq} \frac{4}{T} \mathbb{E} \left[ \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \sum_{i=1}^N \zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2 \right] \\ &\stackrel{(iv)}{\leq} \frac{4}{T} \left( 2\tau^2 + 2 \log N(\tau, \theta, \eta, \rho) + 2 \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \mathbb{E} \left( \sum_{i=1}^N \zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2 \right) \right) \\ &= \frac{8}{T} (\tau^2 + \log N(\tau, \theta, \eta, \rho)), \end{aligned} \tag{A.6.85}$$

where Inequality ii is due to lemma 6.3 of Ledoux and Talagrand (1991), Inequality iii is due to

$$\sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \sum_{i=1}^N \zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2 \geq 0 \tag{A.6.86}$$

and

$$\begin{aligned} \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \left| \sum_{i=1}^N \zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2 \right| &= \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \left| \sum_{i=1}^N -\zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2 \right| \\ &= - \sum_{i=1}^N \zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2. \end{aligned} \quad (\text{A.6.87})$$

$N(\tau, \theta, \eta, \rho)$  in Inequality iv is the  $\tau$  covering number (Wainwright, 2019) of  $\mathcal{C}(\theta, \eta, \rho)$ , and Inequality iv is due to typical arguments bounding empirical process that we list as follows. Let  $\mathfrak{N} = N(\tau, \theta, \eta, \rho)$ . Suppose  $\mathbf{M}_1, \dots, \mathbf{M}_{\mathfrak{N}}$  is the  $\tau$ -cover. Then we have

$$\begin{aligned} &\mathbb{E} \left( \sup_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \sum_{i=1}^N \zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2 \right) \\ &\leq \mathbb{E} \left( 2 \sup_{1 \leq j \leq \mathfrak{N}} \sum_{i=1}^N \zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M}_j)\|_2^2 + 2 \sup_{1 \leq j \leq \mathfrak{N}} \inf_{\mathbf{M} \in \mathcal{C}(\theta, \eta, \rho)} \|\mathbf{M}_j - \mathbf{M}\|_2^2 \right) \\ &= 2 \log \left( \exp \left( \mathbb{E} \left( \sup_{1 \leq j \leq \mathfrak{N}} \sum_{i=1}^N \zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M}_j)\|_2^2 \right) \right) \right) + 2\tau^2 \\ &\leq 2 \log \left( \sum_{j=1}^{\mathfrak{N}} \exp \left\{ \mathbb{E} \left( \sum_{i=1}^N \zeta_i \|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M}_j)\|_2^2 \right) \right\} \right) + 2\tau^2 \\ &= 2 \log \mathfrak{N} + 2\tau^2. \end{aligned} \quad (\text{A.6.88})$$

Readers interested in more details on covering number can take Wainwright (2019) as a reference.

Now we proceed with Inequality (A.6.85) with bounding  $\log N(\tau, \theta, \eta, \rho)$ .

Suppose  $G$  is a  $\mathbb{R}^{N \times T}$  matrix with i.i.d.  $N(0, 1)$  entries. Let  $B_1(R) = \{\Delta \in \mathbb{R}^{N \times T} \mid \|\Delta\|_* \leq R\}$ . Then  $\mathcal{C}(\theta, \eta, \rho) \subset B_1(\sqrt{\frac{\eta \rho \xi}{p_c}} + \frac{3\delta}{2\lambda(L_{\max} + \delta_1)})$ . Let  $\tilde{N}(\tau, R)$  be the  $\tau$ -covering number of  $B_1(R)$ . By Sudakov minoration (Theorem 5.20 in Wainwright (2019)), and the fact that

packing number is no smaller than covering number, we have

$$\begin{aligned}
\sqrt{\log N(\tau, \theta, \eta, \rho)} &\leq \sqrt{\tilde{N}(\tau, \sqrt{\frac{\eta\rho\xi}{p_c}} + \frac{3\delta}{2\lambda(L_{max} + \delta_1)})} \\
&\leq \frac{3}{\tau} \mathbb{E} \left[ \sup_{\|\Delta\|_* \leq \sqrt{\frac{\eta\rho\xi}{p_c}} + \frac{3\delta}{2\lambda(L_{max} + \delta_1)}} \langle G, \Delta \rangle \right] \\
&\leq \frac{3(\sqrt{\frac{\eta\rho\xi}{p_c}} + \frac{3\delta}{2\lambda(L_{max} + \delta_1)})}{\tau} \mathbb{E}(\|G\|_{op}).
\end{aligned} \tag{A.6.89}$$

By (4.2.5) in Tropp (2015), we have

$$\mathbb{E}(\|G\|_{op}) \leq \sqrt{N} + \sqrt{T}. \tag{A.6.90}$$

Therefore, taking  $\tau = \sqrt{\frac{3(\sqrt{\frac{\eta\rho\xi}{p_c}} + \frac{3\delta}{2\lambda(L_{max} + \delta_1)})}{\tau}(\sqrt{N} + \sqrt{T})}$ , we have

$$\mathbb{E}(Z_\rho) \leq \frac{48}{T} \left( \sqrt{\frac{\eta\rho\xi}{p_c}} + \frac{3\delta}{2\lambda(L_{max} + \delta_1)} \right) (\sqrt{N} + \sqrt{T}). \tag{A.6.91}$$

Now invoking theorem 12.9 of Boucheron et al. (2013) with Inequalities (A.6.91) and (A.6.84), we have, for  $t > 0$ ,

$$\begin{aligned}
P \left( Z_\rho \geq \frac{48}{T} \left( \sqrt{\frac{\eta\rho\xi}{p_c}} + \frac{3\delta}{2\lambda(L_{max} + \delta_1)} \right) (\sqrt{N} + \sqrt{T}) + t \right) &\leq \\
\exp \left( -\frac{t}{4} \log \left( 1 + 2 \log \left( 1 + \frac{t}{\frac{96}{T} \left( \sqrt{\frac{\eta\rho\xi}{p_c}} + \frac{3\delta}{2\lambda(L_{max} + \delta_1)} \right) (\sqrt{N} + \sqrt{T}) + \frac{\rho\xi}{T}} \right) \right) \right).
\end{aligned} \tag{A.6.92}$$

### A.6.6. Proof of Theorem 4.2.1

Write the  $F$  in Equation (4.2.3) in the following form

$$F(X) = f(X) + g(X) + \mathfrak{T}\{X \in C_1 \cap C_2 \cap \cdots \cap C_J\}. \quad (\text{A.6.93})$$

For ease of notation, denote  $\mathcal{C} = C_1 \cap C_2 \cap \cdots \cap C_J$ .

Recalling that

$$\begin{aligned} X_{k+0.5} &= X_k - \eta \nabla f(X_k) \\ X_{k+1} &= \widetilde{\text{Prox}}_{\eta(g(X) + \mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5}), \end{aligned} \quad (\text{A.6.94})$$

we denote

$$\begin{aligned} G(X_k) &= \frac{X_k - \text{Prox}_{\eta(g(X) + \mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5})}{\eta} \\ \tilde{G}(X_k) &= \frac{X_k - \widetilde{\text{Prox}}_{\eta(g(X) + \mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5})}{\eta}. \end{aligned} \quad (\text{A.6.95})$$

Then it's clear that

$$\begin{aligned} X_{k+1} &= X_k - \eta \tilde{G}(X_k) \\ \text{Prox}_{\eta(g(X) + \mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5}) &= X_k - \eta G(X_k). \end{aligned} \quad (\text{A.6.96})$$

Recalling the definition of  $\text{Prox}_{\eta(g(X) + \mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5})$ ,

$$\text{Prox}_{\eta(g(X) + \mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5}) = \arg \min_X \left\{ \frac{1}{2\eta} \|X - X_{k+0.5}\|^2 + g(X) + \mathfrak{T}\{X \in \mathcal{C}\} \right\}, \quad (\text{A.6.97})$$

we know that

$$\mathbf{0} \in X - X_{k+0.5} + \eta \partial g(X) + \eta \partial \mathfrak{T}\{X \in \mathcal{C}\} \Big|_{X=X_k - \eta G(X_k)}. \quad (\text{A.6.98})$$

In the later part of this proof, we choose  $\partial g(X_k - \eta G(X_k))$  and  $\partial \mathfrak{T}\{X_k - \eta G(X_k) \in \mathcal{C}\}$  such that

$$\partial g(X_k - \eta G(X_k)) + \partial \mathfrak{T}\{X_k - \eta G(X_k) \in \mathcal{C}\} + \nabla f(X_k) - G(X_k) = \mathbf{0}.$$

We have

$$\begin{aligned} & f(X_k - \eta \tilde{G}(X_k)) + g(X_k - \eta \tilde{G}(X_k)) \\ & \leq f(X_k - \eta G(X_k)) + \langle \nabla f(X_k - \eta G(X_k)), (X_k - \eta \tilde{G}(X_k)) - (X_k - \eta G(X_k)) \rangle + \\ & \quad \frac{L}{2} \|\eta \tilde{G}(X_k) - \eta G(X_k)\|^2 + g(X_k - \eta G(X_k)) + \langle \partial g(X_k - \eta \tilde{G}(X_k)), \eta G(X_k) - \eta \tilde{G}(X_k) \rangle \\ & \leq f(X_k - \eta G(X_k)) + g(X_k - \eta G(X_k)) + L_f \delta_0 + L_g \delta_0 + \frac{L}{2} \delta_0^2. \end{aligned} \quad (\text{A.6.99})$$

To further bound the first two terms in the right hand side, we have for any  $y \in \mathbb{R}^{n \times m}$ ,

$$\begin{aligned} & f(X_k - \eta G(X_k)) + g(X_k - \eta G(X_k)) + \mathfrak{T}\{X_k - \eta G(X_k) \in \mathcal{C}\} \\ & \leq f(X_k) + \langle \nabla f(X_k), -\eta G(X_k) \rangle + \frac{L}{2} \|\eta G(X_k)\|^2 + \\ & \quad g(y) + \langle \partial g(X_k - \eta G(X_k)), X_k - \eta G(X_k) - y \rangle \\ & \quad + \mathfrak{T}\{y \in \mathcal{C}\} + \langle \partial I X_k - \eta G(X_k), X_k - \eta G(X_k) - y \rangle \quad (\text{A.6.100}) \\ & \leq f(y) + \langle \nabla f(X_k), X_k - y - \eta G(X_k) \rangle + \frac{L}{2} \|\eta G(X_k)\|^2 + g(y) + \mathfrak{T}\{y \in \mathcal{C}\} + \\ & \quad \langle \partial g(X_k - \eta G(X_k)) + \partial \mathfrak{T}\{X_k - \eta G(X_k) \in \mathcal{C}\}, X_k - \eta G(X_k) - y \rangle \\ & = f(y) + \langle G(X_k), X_k - y - \eta G(X_k) \rangle + \frac{L}{2} \|\eta G(X_k)\|^2 + g(y) + I(y), \end{aligned}$$

where the last equality is due to (A.6.98).

If we further let  $y = X^*$ , we have

$$\begin{aligned}
& f(X_k - \eta G(X_k)) + g(X_k - \eta G(X_k)) + \mathfrak{T}\{X_k - \eta G(X_k) \in \mathcal{C}\} \\
& \leq f(X^*) + g(X^*) + \mathfrak{T}\{X^* \in \mathcal{C}\} + \langle G(X_k), X_k - X^* - \frac{\eta G(X_k)}{2} \rangle \\
& \quad + \left( \frac{L}{2} \eta^2 - \frac{\eta}{2} \right) \|G(X_k)\|^2 \\
& = f(X^*) + g(X^*) + \mathfrak{T}\{X^* \in \mathcal{C}\} + \frac{1}{2\eta} (\|X_k - X^*\|^2 - \|X_k - \eta G(X_k) - X^*\|^2) \quad (\text{A.6.101}) \\
& \quad + \frac{\eta}{2} (L\eta - 1) \|G(X_k)\|^2 \\
& \leq f(X^*) + g(X^*) + \mathfrak{T}\{X^* \in \mathcal{C}\} + \frac{1}{2\eta} (\|X_k - X^*\|^2 - \|X_k - \eta \tilde{G}(X_k) - X^*\|^2) \\
& \quad + \frac{\delta_0^2}{2\eta} + \frac{\delta_0 D}{\eta} + \frac{\eta}{2} (L\eta - 1) \|G(X_k)\|^2,
\end{aligned}$$

where  $D$  is the diameter of  $\mathcal{C}$ , and the last Inequality is due to

$$\begin{aligned}
& \|X_k - \eta \tilde{G}(X_k) - X^*\|^2 - \|X_k - \eta G(X_k) - X^*\|^2 \\
& = \|X_k - \eta \tilde{G}(X_k) - X^* - (X_k - \eta G(X_k) - X^*)\|^2 + \\
& \quad 2\langle (X_k - \eta \tilde{G}(X_k)) - (X_k - \eta G(X_k)), X_k - \eta G(X_k) - X^* \rangle \quad (\text{A.6.102}) \\
& \leq \delta_0^2 + 2\delta_0 D.
\end{aligned}$$

If we further let  $\eta \leq \frac{1}{L}$  in Inequality (A.6.101), combining with Inequality (A.6.99), and noting that  $X_k - \eta G(X_k), X^* \in \mathcal{C}$ , we have

$$\begin{aligned}
f(X_{k+1}) + g(X_{k+1}) & \leq f(X^*) + g(X^*) + \frac{1}{2\eta} (\|X_k - X^*\|^2 - \|X_{k+1} - X^*\|^2) \\
& \quad + \frac{\delta_0^2}{2\eta} + \frac{\delta_0 D}{\eta} + \frac{L}{2} \delta_0^2 + (L_f + L_g) \delta_0. \quad (\text{A.6.103})
\end{aligned}$$

Adding up  $k = 0 \cdots K - 1$  for Inequality (A.6.103), we have

$$\begin{aligned}
\frac{1}{K} \sum_{j=1}^K (f(X_j) + g(X_j)) & \leq f(X^*) + g(X^*) + \frac{1}{2\eta} \|X_0 - X^*\|^2 + \frac{\delta_0^2}{2\eta} + \frac{\delta_0 D}{\eta} + \frac{L}{2} \delta_0^2 + (L_f + L_g) \delta_0. \\
& \quad (\text{A.6.104})
\end{aligned}$$

This proves the theorem. But now, we also give a variant of the theorem. Suppose  $\bar{X}^K = \frac{1}{K} \sum_{j=1}^K X_j$ , then the convexity of  $f$  and  $g$  implies that the left hand side of Inequality (A.6.104) is larger equal to  $f(\bar{X}^K) + g(\bar{X}^K)$ .

#### A.6.7. Proof of Proposition 4.2.1

Define the following averages:

$$\bar{W}^t = \frac{1}{t-1} \sum_{i=1}^t W^i, \bar{Z}^t = \frac{1}{t-1} \sum_{i=1}^t Z^i, \bar{P}^t = \frac{1}{t-1} \sum_{i=1}^t P^i. \quad (\text{A.6.105})$$

Writing the constraints of optimization problem (4.2.8) in matrix form, we have

$$\begin{pmatrix} \mathbf{0} & -\mathbf{I}_{nm} & \mathbf{I}_{nm} \\ -\mathbf{I}_{nm} & \mathbf{0} & \mathbf{I}_{nm} \end{pmatrix} \begin{pmatrix} \text{vec}(W) \\ \text{vec}(Z) \\ \text{vec}(P) \end{pmatrix} = \mathbf{0}. \quad (\text{A.6.106})$$

Note that the coefficient matrix blocks corresponding to  $\text{vec}(Z)$  and  $\text{vec}(P)$  in the linear constraint (A.6.106) are full column rank matrices. It suffices the conditions of Theorem 4.1 in Cai et al. (2017). Applying Inequality (4.3) in Cai et al. (2017) to our setting with  $\theta_1(x) = h_1(x)$ ,  $\theta_2(x) = h_2(x)$ ,  $\theta_3(x) = \|x - P_0\|^2$ ,  $x'_1 = W^*$ ,  $x'_2 = Z^*$ ,  $x'_3 = P^*$ , we have, for  $\beta \leq \frac{6}{17}$ ,

$$\begin{aligned} & 2\beta t \left\{ \left[ h_1(\bar{W}^t) + h_2(\bar{Z}^t) + \|\bar{P}^t - P_0\|^2 + \langle \Lambda_1^*, (\bar{W}^t - \bar{P}^t) \rangle + \langle \Lambda_2^*, (\bar{Z}^t - \bar{P}^t) \rangle \right] \right. \\ & \quad \left. - \left[ h_1(W^*) + h_2(Z^*) + \|P^* - P_0\|^2 + \langle \Lambda_1^*, (W^* - P^*) \rangle + \langle \Lambda_2^*, (Z^* - P^*) \rangle \right] \right\} \\ & \leq \beta^2 \|Z^1 - Z^*\|^2 + 2\beta^2 \|P^1 - P^*\|^2 + \|\Lambda^1 - \Lambda^*\|^2 + \frac{10}{3} \beta^2 * 2 \|P^1 - P^0\|^2. \end{aligned} \quad (\text{A.6.107})$$



For the left hand side, we define a function

$$U(W, Z, P) = h_1(W) + h_2(Z) + \|P - P_0\|_F^2 + \langle \Lambda_1^*, W \rangle + \langle \Lambda_2^*, Z \rangle - \langle (\Lambda_1^* + \Lambda_2^*), P \rangle. \quad (\text{A.6.108})$$

Given that  $(W^*, Z^*, P^*), (\Lambda_1^*, \Lambda_2^*)$  is a solution to

$$\max_{\Lambda_1, \Lambda_2} \min_{W, Z, P} h_1(W) + h_2(Z) + \|P - P_0\|_F^2 + \langle \Lambda_1, W \rangle + \langle \Lambda_2, Z \rangle - \langle (\Lambda_1 + \Lambda_2), P \rangle,$$

we have

$$0 = \left. \frac{\partial U(W, Z, P)}{\partial P} \right|_{W=W^*, Z=Z^*, P=P^*} = 2(P^* - P_0) - (\Lambda_1^* + \Lambda_2^*). \quad (\text{A.6.109})$$

Further, since  $U(W, Z, P)$  is separable with respect to  $W, Z, P$ , we have

$$\begin{aligned} & U(W, Z, P) - U(W^*, Z^*, P^*) \\ & \geq U(W^*, Z^*, P) - U(W^*, Z^*, P^*) \\ & = \|P - P_0\|^2 - \|P^* - P_0\|^2 - (\Lambda_1^{*T} + \Lambda_2^{*T})(P - P^*) \\ & = \|P - P^*\|^2 + \langle P - P^*, 2(P^* - P_0) - \Lambda_1^* - \Lambda_2^* \rangle \\ & = \|P - P^*\|^2. \end{aligned} \quad (\text{A.6.110})$$

Combining Equation (A.6.109) and (A.6.110), we have

$$\|\bar{P}^t - P^*\|^2 \leq \frac{1}{2\beta t} \left( \beta^2 \|Z^1 - Z^*\|^2 + 2\beta^2 \|P^1 - P^*\|^2 + \|\Lambda^1 - \Lambda^*\|^2 + \frac{20}{3}\beta^2 \|P^1 - P^0\|^2 \right). \quad (\text{A.6.111})$$

#### A.6.8. Proof of Lemma 4.2.1

We begin with bounding  $C(C_1, C_2)$ .

$$C(C_1, C_2) = \frac{1}{2 \cos^2\left(\frac{\theta(C_1, C_2)}{2}\right)} = \frac{1}{\cos(\theta(C_1, C_2)) + 1}. \quad (\text{A.6.112})$$

Observe that  $B_d(x) \subset C_1 \cap C_2$ , we have

$$\begin{aligned}
\cos(\theta(C_1, C_2)) &= \inf_{P \in \partial(C_1 \cap C_2)} \cos\left(\sup_{\lambda_1 \in N_{C_1}(P), \lambda_2 \in N_{C_2}(P)} \arccos(\langle \lambda_1, \lambda_2 \rangle)\right) \\
&\geq \inf_{P \in \partial(C_1 \cap C_2)} \cos\left(\sup_{\lambda_1 \in N_{B_d(x)}(P), \lambda_2 \in N_{B_d(x)}(P)} \arccos(\langle \lambda_1, \lambda_2 \rangle)\right) \\
&= \inf_{P \in \partial(C_1 \cap C_2)} -\left(2 \frac{\|P - x\|^2 - d^2}{\|P - x\|^2} - 1\right) \\
&\geq -1 + \frac{2d^2}{\tilde{D}^2},
\end{aligned} \tag{A.6.113}$$

where  $\tilde{D} = \sup_{P \in \partial(C_1 \cap C_2)} \|P - x\|_F$ .

Therefore,

$$C(C_1, C_2) \leq \frac{\tilde{D}^2}{2d^2} \leq \frac{D^2}{2d^2}, \tag{A.6.114}$$

where  $D = \sup_{P_1, P_2 \in \partial(C_1 \cap C_2)} \|P_1 - P_2\|_F$ .

Now we continue with bounding dual variable  $\Lambda^*$  in the case that  $h_1(X) = \mathfrak{T}\{X \in C_1\}, h_2(X) = \mathfrak{T}\{X \in C_2\}$ .

From Equation (A.6.109), we know that

$$\begin{aligned}
4\|P^* - P_0\|^2 &= \|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2 + 2\langle \Lambda_1^*, \Lambda_2^* \rangle \\
&\geq \|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2 + 2\cos(\theta(C_1, C_2))\|\Lambda_1^*\|\|\Lambda_2^*\| \\
&\geq \|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2 + \min\{0, 2\cos(\theta(C_1, C_2))\} \frac{\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2}{2} \\
&\geq \min\{1, \frac{1}{C(C_1, C_2)}\}(\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2).
\end{aligned} \tag{A.6.115}$$

Therefore, we have

$$\|\Lambda^*\|_F^2 \leq \max\{4, 4C(C_1, C_2)\}\|P^* - P_0\|^2. \tag{A.6.116}$$

### A.6.9. Proof of Proposition 4.3.1

To apply Proposition 4.2.1 to 1 bit completion matrix problem, we only need to find the  $L, L_f, L_g, D$  and a bound for  $\|X_0 - X^*\|$  in Proposition 4.2.1 in 1 bit matrix completion setting and bound.

Since  $g = 0$  in this case, we have  $L_g = 0$ . Since  $C_1 = [-\alpha, \alpha]^{d_1 \times d_2}$ , we have  $D \leq 2\alpha\sqrt{d_1 d_2}$ .

Easy calculation also shows  $\sup_{|x| \leq \alpha + \delta_0} \frac{|l'(x)|}{l(x)(1-l(x))}$  is the Lipschitz constant for the smooth objective function  $-\mathcal{L}_{\Omega, Y}(X)$ .

Easy calculation also show that

$$\sup_{|x| \leq \alpha + \delta_0} \left\{ \frac{|l''(x)l(x) - (l'(x))^2|}{l(x)^2}, \frac{|l''(x)(1-l(x)) + (l'(x))^2|}{(1-l(x))^2} \right\}$$

is the smoothness parameter for the smooth objective function  $-\mathcal{L}_{\Omega, Y}(X)$ .

Also, given that  $X_0 = \mathbf{0}$  and  $X^* \in [-\alpha, \alpha]^{d_1 \times d_2}$ , we have  $\|X_0 - X^*\|^2 \leq \alpha^2 d_1 d_2$ .

With the step size set to be the inverse of smoothness parameter, we completes the proof of the Proposition.

### A.6.10. Proof of Proposition 4.3.2

Note that when  $X \in \mathbb{R}^{d_1 \times d_2}$  satisfies  $\|X\|_F \leq \alpha$ , we have  $\|X\|_* \leq \sqrt{\text{rank}(X)}\|X\|_F \leq \sqrt{\min\{d_1, d_2\}}\alpha \leq \alpha\sqrt{r d_1 d_2}$ , and  $\|X\|_\infty \leq \alpha$ . Therefore, we have  $d \geq \alpha$ .

Note that when  $X \in [-\alpha, \alpha]^{d_1 \times d_2}$ , we have  $\|X\|_F \leq \alpha\sqrt{d_1 d_2}$ . Therefore,  $\tilde{D} \leq \alpha\sqrt{d_1 d_2}$ , where  $\tilde{D}$  is defined after Inequality (A.6.113).

According to the proof of Lemma 4.2.1, when we take  $x$  in the  $B_d(x)$  there to be  $\mathbf{0}$ , we have  $C(C_1, C_2) \leq \frac{d_1 d_2}{2}$ .

We continue with bounding the terms in right hand side of Inequality (4.2.13) in Proposition 4.2.1.

Recall the steps we take in Algorithm 4.3.2, then we have

$$\begin{aligned}
\|Z^1 - Z^*\| &= \|\text{Proj}_{C_2}(P_0) - P^*\| \leq \|P_0 - P^*\|, \\
\|P^1 - P^*\| &= \left\| \frac{\beta}{2(\beta+1)} (\text{Proj}_{C_1}(P_0) - P^* + \text{Proj}_{C_2}(P_0) - P^*) \right. \\
&\quad \left. + \frac{1}{\beta+1} (P_0 - P^*) \right\| \leq \|P_0 - P^*\|, \\
\|\Lambda^1 - \Lambda^*\|^2 &\leq 2\|\Lambda^1\|^2 + 2\|\Lambda^*\|^2 \\
&\leq 2\beta^2 \left\| \frac{1}{\beta+1} \left( P_0 + \frac{\beta}{2} \text{Proj}_{C_2}(P_0) - (1 + \frac{\beta}{2}) \text{Proj}_{C_1}(P_0) \right) \right\|^2 \\
&\quad + 2\beta^2 \left\| \frac{1}{\beta+1} \left( P_0 + \frac{\beta}{2} \text{Proj}_{C_1}(P_0) - (1 + \frac{\beta}{2}) \text{Proj}_{C_2}(P_0) \right) \right\|^2 \\
&\quad + \max\{4, 8C(C_1, C_2)\} \|P_0 - P^*\|^2 \\
&\leq 4\beta^2 \|P_0 - P^*\|^2 + \max\{4, 8C(C_1, C_2)\} \|P_0 - P^*\|^2, \\
\|P^1 - P_0\| &\leq \frac{\beta}{2(\beta+1)} \|\text{Proj}_{C_1}(P_0) - P_0 + \text{Proj}_{C_2}(P_0) - P_0\| \leq \frac{\beta}{\beta+1} \|P_0 - P^*\|.
\end{aligned} \tag{A.6.117}$$

Some of the inequalities in Inequality (A.6.117) are due to  $\|P_0 - \text{Proj}_{C_i}(P_0)\| \leq \|P_0 - P^*\|$ ,  $\|\text{Proj}_{C_1}(P_0) - \text{Proj}_{C_2}(P_0)\| \leq \sum_{i=1}^2 \|P_0 - \text{Proj}_{C_i}(P_0)\|$ .

Plugging Inequality A.6.117 back to Proposition 4.2.1, we have

$$\begin{aligned}
\|\bar{P}^t - P^*\|^2 &\leq \frac{1}{2\beta t} (7\beta^2 + \max\{4, 8C(C_1, C_2)\} + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2}) \|P_0 - P^*\|^2 \\
&\leq \frac{1}{2\beta t} (7\beta^2 + 4d_1d_2 + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2}) \|P_0 - P^*\|^2.
\end{aligned} \tag{A.6.118}$$

### A.6.11. Proof of Theorem 4.3.2

First, we will show that for  $t \geq t_0$ ,  $\delta_0 \leq \min\{u_0, 1\}$ .

We prove this by mathematical induction. For  $X_0, X_0 \in C_1 \cap C_2$ , therefore  $\delta_0 \leq u_0$  holds for  $k=0$ . One thing to note is that  $L_{\alpha+u_0} \leq 2L_\alpha, \tilde{L}_{\alpha+u_0} \leq 2\tilde{L}_\alpha$ . Also, recall that  $\eta = \frac{1}{2\tilde{L}_\alpha}$ . Suppose  $\delta_{k-1} \leq \min\{u_0, 1\}$  holds for  $k \leq H$ , then for  $k = H + 1$ , we have

$$\begin{aligned} \|X_k - \eta \nabla f(X_k) - \text{Prox}_{C_1 \cap C_2}(X_k)\| &\leq \|X_k - \text{Prox}_{C_1 \cap C_2}(X_k)\| + |\eta \nabla f(X_k)| \\ &\leq u_0 + \frac{1}{2\tilde{L}_\alpha} 2L_\alpha. \end{aligned} \quad (\text{A.6.119})$$

Therefore,

$$\|X_k - \eta \nabla f(X_k) - \text{Prox}_{C_1 \cap C_2}(X_k - \eta \nabla f(X_k))\| \leq u_0 + \frac{L_\alpha}{\tilde{L}_\alpha}. \quad (\text{A.6.120})$$

According to Proposition 4.3.2, for

$$t \geq \frac{1}{2\beta} \left( 7\beta^2 + 4d_1d_2 + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2} \right) \left( 1 + \frac{L_\alpha}{u_0\tilde{L}_\alpha} + \frac{L_\alpha}{\tilde{L}_\alpha} \right)^2,$$

we have

$$\|X_{k+1} - \text{Prox}_{C_1 \cap C_2}(X_{k+0.5})\|^2 \leq \min\{u_0^2, 1\}. \quad (\text{A.6.121})$$

So  $\delta_0 \leq \{u_0, 1\}$  also holds for  $k = H + 1$ .

Therefore,  $\delta_0 \leq \{u_0, 1\}$  for all  $k$ . So the Lipschitz constant  $L_f \leq 2L_\alpha$ , and the smooth parameter  $L \leq 2\tilde{L}_\alpha$  for the objective function on  $u_0$  neighbor of  $C_1 \cap C_2$ .

Further, we have,

$$\delta_0 \leq \sqrt{\frac{1}{t}} \sqrt{\frac{1}{2\beta} \left( 7\beta^2 + 4d_1d_2 + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2} \right)}. \quad (\text{A.6.122})$$

According to Proposition 4.3.1, we have

$$\begin{aligned} \delta \leq & \frac{\alpha^2 \tilde{L}_\alpha d_1 d_2}{T} + 4\alpha \tilde{L}_\alpha \sqrt{d_1 d_2} \sqrt{\frac{1}{t}} \sqrt{\frac{1}{2\beta} \left( 7\beta^2 + 4d_1 d_2 + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2} \right)} \\ & + 2L_\alpha \sqrt{\frac{1}{t}} \sqrt{\frac{1}{2\beta} \left( 7\beta^2 + 4d_1 d_2 + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2} \right)} + 2\tilde{L}_\alpha \frac{1}{t} \frac{1}{2\beta} \left( 7\beta^2 + 4d_1 d_2 + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2} \right). \end{aligned} \quad (\text{A.6.123})$$

#### A.6.12. Proof of Proposition 4.4.1

To apply Proposition 4.2.1 to causal inference for panel data, we only need to find the  $L, L_f, L_g, D$  and a bound for  $\|X_0 - X^*\|$  in Proposition 4.2.1 in causal inference for panel data.

Since  $C_1 = [-L_{\max}, L_{\max}]^{N \times T}$ , we have  $D = 2L_{\max} \sqrt{NT}$ .

Since  $g(\mathbf{L}) = \frac{\lambda|\mathcal{O}|}{2} |\mathbf{L}|$ , we have  $\|\partial g\| \leq \frac{\lambda|\mathcal{O}|}{2} \sqrt{\min\{N, T\}}$ .

Since  $f(\mathbf{L}) = \frac{1}{2} \|\mathbf{P}_{\mathcal{O}}(Y_{\mathbf{L}})\|_F^2$ , we have the smooth parameter  $L \leq 1$ , the Lipschitz constant  $L_f \leq \max_{L \in C_1} \|Y - L\|_F$ .

Also, we have  $\|\mathbf{L}_0 - \hat{\mathbf{L}}\| \leq L_{\max} \sqrt{NT}$ . Recall that  $\eta = 1$ .

Plugging in the quantities into Proposition 4.2.1, we have

$$\begin{aligned} \min_{0 \leq k \leq K} \frac{1}{2} \|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L}_k)\|_F^2 + \frac{\lambda|\mathcal{O}|}{2} \|\mathbf{L}_k\|_* & \leq \frac{1}{2} \|\mathbf{P}_{\mathcal{O}}(Y - \hat{\mathbf{L}})\|_F^2 + \frac{\lambda|\mathcal{O}|}{2} \|\hat{\mathbf{L}}\|_* \\ & + \frac{1}{2K} \|\mathbf{L}_0 - \hat{\mathbf{L}}\|^2 + \left( \frac{\lambda|\mathcal{O}|}{2} \sqrt{\min\{N, T\}} + \max_{L \in C_1} \|\mathbf{P}_{\mathcal{O}}(Y - L)\|_F \right) \delta_0 \\ & + \delta_0^2 + 2L_{\max} \sqrt{NT} \delta_0. \end{aligned} \quad (\text{A.6.124})$$

### A.6.13. Proof of Proposition 4.4.2

We continue with bounding the terms in right hand side of Inequality (4.2.13) in Proposition

4.2.1. Recall the steps we take in Algorithm 4.4.11, we have

$$\begin{aligned}
\|Z^1 - Z^*\| &= \|\text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - P^*\| \leq \|P_0 - P^*\| + \frac{\lambda|\mathcal{O}|}{\beta} \sqrt{\min\{N, T\}}, \\
\|P^1 - P^*\| &\leq \|P_0 - P^*\| + \|P^1 - P_0\|, \\
\|\Lambda^1 - \Lambda^*\|^2 &\leq 2(\|\Lambda_1^1\|^2 + \|\Lambda_2^1\|^2 + \|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2) \\
&\leq 2\left(\frac{\beta}{1+\beta}\right)^2 \left( \left\| P_0 - \text{Proj}_{C_1}(P_0) + \frac{\beta}{2} \left( \text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \text{Proj}_{C_1}(P_0) \right) \right\|^2 \right. \\
&\quad \left. + \left\| P_0 - \text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \frac{\beta}{2} \left( \text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \text{Proj}_{C_1}(P_0) \right) \right\|^2 \right) \\
&\quad + 2\|\Lambda_1^*\| + 2\|\Lambda_2^*\|, \\
\|P^1 - P_0\| &= \left\| \frac{\beta}{2(\beta+1)} \left( \text{Proj}_{C_1}(P_0) - P_0 + \text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - P_0 \right) \right\| \\
&\leq \frac{\beta}{2(\beta+1)} \|\text{Proj}_{C_1}(P_0) - P_0\| + \frac{\beta}{2(\beta+1)} \frac{\lambda|\mathcal{O}|}{\beta} \sqrt{\min\{N, T\}}.
\end{aligned} \tag{A.6.125}$$

We continue with bounding the two terms in the right hand side for  $\|\Lambda^1 - \Lambda^*\|^2$ . We start with the first term.

$$\begin{aligned}
& \left\| P_0 - \text{Proj}_{C_1}(P_0) + \frac{\beta}{2} \left( \text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \text{Proj}_{C_1}(P_0) \right) \right\|^2 \\
& + \left\| P_0 - \text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \frac{\beta}{2} \left( \text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \text{Proj}_{C_1}(P_0) \right) \right\|^2 \\
& = \|P_0 - \text{Proj}_{C_1}(P_0)\|^2 + \|P_0 - \text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta})\|^2 \\
& \quad + (\beta + \frac{\beta^2}{2}) \|\text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \text{Proj}_{C_1}(P_0)\|^2 \\
& \leq (1 + \beta)^2 \left( \|P_0 - \text{Proj}_{C_1}(P_0)\|^2 + \|P_0 - \text{thresh}(P_0, \frac{\lambda|\mathcal{O}|}{\beta})\|^2 \right) \\
& \leq (1 + \beta)^2 \left( \|P_0 - \text{Proj}_{C_1}(P_0)\|^2 + \min\{N, T\} (\frac{\lambda|\mathcal{O}|}{\beta})^2 \right).
\end{aligned} \tag{A.6.126}$$

We proceed with bounding  $\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2$ .

According to Equation (A.6.109), we have

$$\begin{aligned}
\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2 &= \|2(P^* - P_0) - \Lambda_2^*\|^2 + \|\Lambda_2^*\|^2 \\
&\leq 8\|P^* - P_0\|^2 + 3\|\Lambda_2^*\|^2.
\end{aligned} \tag{A.6.127}$$

Taking derivative with respect to  $Z$  for function  $U(W, Z, P)$  at point  $(W^*, Z^*, P^*)$ , we have

$$\mathbf{0} = \frac{\partial U(W, Z, P)}{\partial Z} \Big|_{W=W^*, Z=Z^*, P=P^*} = \partial h_2(Z^*) + \Lambda_2^*. \tag{A.6.128}$$

Observe that  $\partial h_2(Z^*) \leq \lambda|\mathcal{O}|\sqrt{\min\{N, T\}}$ , continuing with Inequality (A.6.127), we have

$$\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2 \leq 8\|P^* - P_0\|^2 + 3(\lambda|\mathcal{O}|)^2 \min\{N, T\}. \tag{A.6.129}$$

Putting together Inequalities (A.6.125), (A.6.126), (A.6.129), together with Proposition



4.2.1, we have

$$\begin{aligned}
& \|\bar{P}^k - P^*\|^2 \\
& \leq \frac{1}{2\beta k} \left( 2\beta^2 \|P_0 - P^*\|^2 + 2(\lambda|\mathcal{O}|)^2 \min\{N, T\} + 4\beta^2 \|P_0 - P^*\|^2 + \right. \\
& \quad 4\beta^2 \|P^1 - P_0\|^2 + 2\beta^2 \|P_0 - \text{Proj}_{C_1}(P_0)\|^2 + 2\min\{N, T\} (\lambda|\mathcal{O}|)^2 \\
& \quad \left. + 16\|P^* - P_0\|^2 + 6(\lambda|\mathcal{O}|)^2 \min\{N, T\} + \frac{20}{3}\beta^2 \|P^1 - P_0\|^2 \right) \\
& \leq \frac{1}{2\beta k} \left( (6\beta^2 + 16)\|P_0 - P^*\|^2 + \left( 10 + (2 + \frac{10}{3})(\frac{\beta}{1+\beta})^2 \right) (\lambda|\mathcal{O}|)^2 \min\{N, T\} \right. \quad (\text{A.6.130}) \\
& \quad \left. + \left( 2\beta^2 + (2 + \frac{10}{3}) \left( \frac{\beta^2}{1+\beta} \right)^2 \right) \|P_0 - \text{Proj}_{C_1}(P_0)\|^2 \right) \\
& = \frac{1}{\beta k} \left( (3\beta^2 + 8)\|P_0 - P^*\|^2 + \left( 5 + \frac{8}{3}(\frac{\beta}{1+\beta})^2 \right) (\lambda|\mathcal{O}|)^2 \min\{N, T\} \right. \\
& \quad \left. + \left( \beta^2 + \frac{8}{3}(\frac{\beta^2}{1+\beta})^2 \right) \|P_0 - \text{Proj}_{C_1}(P_0)\|^2 \right).
\end{aligned}$$

#### A.6.14. Proof of Theorem 4.4.2

Suppose  $\inf_{\mathbf{L} \in C_1} \|\mathbf{L}_j - \mathbf{L}\| \leq \delta_0$  for  $j \leq k$ , where  $k \geq 0$ .

Recall that

$$\mathbf{L}_{k+0.5} = \mathbf{L}_k + \mathbf{P}_{\mathcal{O}}(Y - \mathbf{L}_k), \quad (\text{A.6.131})$$

we have

$$\begin{aligned}
\|\text{Proj}_{C_1}(\mathbf{L}_{k+0.5}) - \mathbf{L}_{k+0.5}\|^2 & \leq \|\text{Proj}_{C_1}(\mathbf{L}_k) - \mathbf{L}_{k+0.5}\|^2 \leq (C(Y) + \delta_0)^2 \leq 2C(Y)^2 + 2\delta_0^2. \\
& \quad (\text{A.6.132})
\end{aligned}$$

Recalling that  $\text{Prox}_{\frac{\lambda|\mathcal{O}|}{2}\|\mathbf{L}\|_* + \mathfrak{T}\{\mathbf{L} \in C_1\}}(\mathbf{L}_{k+0.5})$  is defined as

$$\arg \min_{\mathbf{L}} \|\mathbf{L} - \mathbf{L}_{k+0.5}\|^2 + \lambda|\mathcal{O}|\|\mathbf{L}\|_* + \mathfrak{T}\{\mathbf{L} \in C_1\}, \quad (\text{A.6.133})$$

we have

$$\begin{aligned} & \|\text{Prox}(\mathbf{L}_{k+0.5}) - \mathbf{L}_{k+0.5}\|^2 + \lambda|\mathcal{O}|\|\text{Prox}(\mathbf{L}_{k+0.5})\|_* + \mathfrak{T}\{\text{Prox}(\mathbf{L}_{k+0.5}) \in C_1\} \\ & \leq \|\mathbf{0} - \mathbf{L}_{k+0.5}\|^2 + \lambda|\mathcal{O}|\|\mathbf{0}\|_* + \mathfrak{T}\{\mathbf{0} \in C_1\} = \|\mathbf{L}_{k+0.5}\|^2 \\ & \leq \|Y\|_2^2 + (\sqrt{NT - |\mathcal{O}|}L_{\max} + \delta_0)^2. \end{aligned} \quad (\text{A.6.134})$$

Combing Proposition 4.4.1 and Proposition 4.4.2, we have for  $\beta \leq \frac{6}{17}$ , then

$$\begin{aligned} \|\mathbf{L}_{k+1} - \text{Prox}(\mathbf{L}_{k+0.5})\|^2 & \leq \frac{1}{\beta k} \left( (3\beta^2 + 8) \left( \|Y\|_2^2 + (\sqrt{NT - |\mathcal{O}|}L_{\max} + \delta_0)^2 \right) \right. \\ & \quad + \left( 5 + \frac{8}{3} \left( \frac{\beta}{1+\beta} \right)^2 \right) (\lambda|\mathcal{O}|)^2 \min\{N, T\} \\ & \quad \left. + \left( \beta^2 + \frac{8}{3} \left( \frac{\beta^2}{1+\beta} \right)^2 \right) (2C(Y)^2 + 2\delta_0^2) \right) \\ & \leq \frac{1}{k} \left( \delta_0^2 \left( \frac{1}{\beta} \left( 6\beta^2 + 16 + 2\beta^2 + \frac{16}{3} \left( \frac{\beta^2}{1+\beta} \right)^2 \right) \right) + \right. \\ & \quad \frac{1}{\beta} (3\beta^2 + 8) (\|Y\|^2 + 2(NT - |\mathcal{O}|)L_{\max}^2) + \\ & \quad \frac{1}{\beta} \left( 5 + \frac{8}{3} \left( \frac{\beta}{1+\beta} \right)^2 \right) (\lambda|\mathcal{O}|)^2 \min\{N, T\} + \\ & \quad \left. \beta \left( 2 + \frac{16}{3} \left( \frac{\beta}{1+\beta} \right)^2 \right) C(Y)^2 \right). \end{aligned} \quad (\text{A.6.135})$$

Let

$$\begin{aligned}
q_0(\beta) &= \left( \frac{1}{\beta} \left( 6\beta^2 + 16 + 2\beta^2 + \frac{16}{3} \left( \frac{\beta^2}{1+\beta} \right)^2 \right) \right), \\
q_1(\beta) &= \frac{1}{\beta} \left( 5 + \frac{8}{3} \left( \frac{\beta}{1+\beta} \right)^2 \right), \\
q_2(\beta) &= \beta \left( 2 + \frac{16}{3} \left( \frac{\beta}{1+\beta} \right)^2 \right), \\
q_3(\beta) &= \frac{1}{\beta} (3\beta^2 + 8), \\
\delta(k) &= \sqrt{\frac{q_1(\beta)(\lambda|\mathcal{O}|)^2 \min\{N, T\} + q_2(\beta)C(Y)^2 + q_3(\beta)(\|Y\|^2 + 2(NT - |\mathcal{O}|)L_{\max}^2)}{k - q_0(\beta)}}
\end{aligned} \tag{A.6.136}$$

We show next that when  $k \geq q_0(\beta)$ ,  $\inf_{\mathbf{L} \in C_1} \|\mathbf{L}_k - \mathbf{L}\| \leq \delta(k)$  and  $\|\mathbf{L}_{k+1} - \text{Prox}(\mathbf{L}_{k+0.5})\| \leq \delta(k)$  for all  $k \geq 0$ . For  $k = 0$ ,  $\mathbf{L}_0 \in C_1$ , the first part claim holds. Suppose the first part of claim holds for  $k \leq k_0$ , where  $k_0 \geq 0$ , then for  $k = k_0 + 1$ ,

$$\begin{aligned}
&\|\mathbf{L}_{k_0+1} - \text{Prox}(\mathbf{L}_{k_0+0.5})\|^2 \\
&\leq \frac{1}{k} \left( \delta(k)^2 q_0(\beta) + q_1(\beta)(\lambda|\mathcal{O}|)^2 \min\{N, T\} + q_2(\beta)C(Y)^2 \right. \\
&\quad \left. + q_3(\beta)(\|Y\|^2 + 2(NT - |\mathcal{O}|)L_{\max}^2) \right) \\
&= \delta(k)^2.
\end{aligned} \tag{A.6.137}$$

Since  $\text{Prox}(\mathbf{L}_{k_0+0.5}) \in C_1$ , the first part of claim holds for  $k = k_0 + 1$ . So the first part of the holds for all  $k \geq 0$ . Since Inequality A.6.137 is based on  $\|\mathbf{L}_{k_0} - \text{Proj}_{C_1}(\mathbf{L}_{k_0})\| \leq \delta(k)$ , it holds for all  $k_0 \geq 0$ .

Therefore, for  $k \geq q_0(\beta)$ , we have  $\delta_0 \leq \delta(k)$ . Therefore, we know that  $\delta_1 \leq \delta(k)$ .

Now we proceed with bounding  $\delta$ . According to Proposition 4.4.1, we have

$$\begin{aligned}\delta &\leq \frac{2}{|\mathcal{O}|} \left( \frac{1}{2K} \|\mathbf{L}_0 - \hat{\mathbf{L}}\|^2 + \delta(k)^2 + \left( 2L_{\max} \sqrt{NT} + C(Y) + \min\{\sqrt{N}, \sqrt{T}\} \frac{\lambda|\mathcal{O}|}{2} \right) \delta(k) \right) \\ &\leq \frac{NTL_{\max}^2}{K|\mathcal{O}|} + \frac{2\delta(k)^2}{|\mathcal{O}|} + \left( \frac{4L_{\max} \sqrt{NT}}{|\mathcal{O}|} + \frac{2C(Y)}{|\mathcal{O}|} + \min\{\sqrt{N}, \sqrt{T}\} \lambda \right) \delta(k).\end{aligned}\tag{A.6.138}$$

This finishes the proof.

#### A.6.15. Proof of Theorem 4.5.1

Recall that we use  $\rho^2(\Sigma)$  to denote the maximum diagonal entry of the covariance matrix  $\Sigma$ .

It suffices to prove the following two results

**Proposition A.6.2.** *Under the linear regression model (4.5.1), for any sparse index set  $S$  such that the cardinal of  $S$ ,  $|S| = s$ , denote  $\theta_{S^c}^*$  to be the vector keeping elements not in  $S$  the same and setting those in  $S$  to be 0. Suppose  $c_1\kappa \geq 64s \cdot c_2\rho^2(\Sigma) \frac{\log d}{n}$ , where  $c_1, c_2$  are constants and can be taken as  $c_1 = 1/8, c_2 = 50$ , and  $\kappa$  is the smallest singular value of  $\Sigma$ . For  $\lambda_n \geq \frac{2\|\mathbf{X}^T w\|_\infty}{n}$ ,  $\tilde{\theta}$  satisfying (4.5.3) has the following property*

$$P(\|\Delta\|_2 < \frac{\delta}{2\lambda_n\sqrt{s}} + \frac{\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa}) \geq 1 - \frac{\exp(-n/32)}{1 - \exp(-n/32)}.\tag{A.6.139}$$

**Lemma A.6.7.** *For the random matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$ , in which each row  $x_i$  is drawn i.i.d. from a  $N(0, \Sigma)$  distribution, its columns  $\tilde{x}_k$  satisfies the following with probability at least  $1 - \exp(-\frac{n}{4\rho^2(\Sigma)}\epsilon)$ ,*

$$\max_{1 \leq k \leq d} \frac{\|\tilde{x}_k\|_2^2}{n} \leq 2 \log 2 \cdot \rho^2(\Sigma) + \frac{4\rho^2(\Sigma)}{n} \log d + \epsilon.\tag{A.6.140}$$

For  $w$  with  $w_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$  and  $w$  independent with  $\mathbf{X}$ , we have that

$$P_{\mathbf{X}, w} \left( \left\| \frac{\mathbf{X}^T w}{n} \right\|_\infty < 2\rho(\Sigma) \sqrt{\left(\frac{\log d}{n} + 1\right)\sigma} \sqrt{\frac{2 \log(2d)}{n} + \mu} \right) \geq 1 - \exp\left(-\frac{n}{2}\right) - \exp\left(-\frac{n\mu}{2}\right). \quad (\text{A.6.141})$$

### Proof of Proposition A.6.2

From Inequality (4.5.3), we have

$$\|y - \mathbf{X}\tilde{\theta}\|_2^2 + \lambda_n \|\tilde{\theta}\|_1 \leq \|y - \mathbf{X}\hat{\theta}\|_2^2 + \lambda_n \|\hat{\theta}\|_1 + \delta \leq \|y - \mathbf{X}\theta^*\|_2^2 + \lambda_n \|\theta^*\|_1 + \delta. \quad (\text{A.6.142})$$

Denote  $\Delta = \tilde{\theta} - \theta^*$ .

Therefore, we have that

$$\begin{aligned} 0 &\leq \frac{1}{2n} \|\mathbf{X}\Delta\|^2 \leq \left\| \frac{\mathbf{X}^T w}{n} \right\|_\infty \|\Delta\|_1 + \lambda_n \left( \|\theta^*\|_1 - \|\tilde{\theta}\|_1 \right) + \delta \\ &\leq \frac{\lambda_n}{2} \left( \|\Delta\|_1 + 2\|\theta^*\|_1 - 2\|\tilde{\theta}\|_1 \right) + \delta \\ &\leq \frac{\lambda_n}{2} (3\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 + 2\|\theta_{S^c}^*\|_1 - 2\|\theta_{S^c}^* + \Delta_{S^c}\|_1) + \delta \\ &\leq \frac{\lambda_n}{2} (3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1 + 4\|\theta_{S^c}^*\|_1) + \delta. \end{aligned} \quad (\text{A.6.143})$$

Therefore, we have

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \leq 4\|\Delta_S\|_1 + 4\|\theta_{S^c}^*\|_1 + \frac{2\delta}{\lambda_n} \leq 4\sqrt{s}\|\Delta\|_2 + 4\|\theta_{S^c}^*\|_1 + \frac{2\delta}{\lambda_n}. \quad (\text{A.6.144})$$

On the other hand, according Theorem 7.16 in Wainwright (2019), we have that with probability at  $1 - \frac{\exp(-n/32)}{1 - \exp(-n/32)}$ ,

$$\frac{\|\mathbf{X}\Delta\|_2^2}{n} \geq c_1 \|\sqrt{\Sigma}\Delta\|_2^2 - c_2 \rho^2(\Sigma) \frac{\log d}{n} \|\Delta\|_1^2, \quad (\text{A.6.145})$$

where  $c_1, c_2$  are absolute constants and can be taken as  $c_1 = 1/8, c_2 = 50$ .

Note that  $\|\sqrt{\Sigma}\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2$ , going back to Inequality (A.6.143), we have

$$\begin{aligned}
c_1\kappa\|\Delta\|_2^2 &\leq c_2\rho^2(\Sigma)\frac{\log d}{n}\|\Delta\|_1^2 + \lambda_n(3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1 + 4\|\theta_{S^c}^*\|_1) + 2\delta \\
&\leq c_2\rho^2(\Sigma)\frac{\log d}{n}\left(4\sqrt{s}\|\Delta\|_2 + 4\|\theta_{S^c}^*\|_1 + \frac{2\delta}{\lambda_n}\right)^2 \\
&\quad + \lambda_n(3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1 + 4\|\theta_{S^c}^*\|_1) + 2\delta \\
&\leq c_1\kappa\left(\frac{\|\Delta\|_2}{2} + \frac{\delta}{4\lambda_n\sqrt{s}} + \frac{\|\theta_{S^c}^*\|_1}{2\sqrt{s}}\right)^2 + \lambda_n(3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1 + 4\|\theta_{S^c}^*\|_1) + 2\delta \\
&\leq c_1\kappa\left(\frac{\|\Delta\|_2}{2} + \frac{\delta}{4\lambda_n\sqrt{s}} + \frac{\|\theta_{S^c}^*\|_1}{2\sqrt{s}}\right)^2 + \lambda_n(3\sqrt{s}\|\Delta\|_2 + 4\|\theta_{S^c}^*\|_1) + 2\delta.
\end{aligned} \tag{A.6.146}$$

Solving the Inequality for  $\|\Delta\|_2$ , we have

$$\|\Delta\|_2 < \frac{\delta}{2\lambda_n\sqrt{s}} + \frac{\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}. \tag{A.6.147}$$

### Proof of Lemma A.6.7

Denote  $\nu_k = \frac{\|\tilde{x}_k\|_2^2}{n}$ .

For  $\frac{n}{2\rho^2(\Sigma)} > \lambda > 0$ ,

$$\mathbb{E}(\exp(\lambda \max\{\nu_k : 1 \leq k \leq d\})) \leq \sum_{k=1}^d \mathbb{E}(\exp(\nu_k \lambda)) \leq d\left(\frac{1}{1 - \frac{2\lambda\rho^2(\Sigma)}{n}}\right)^{\frac{n}{2}}. \tag{A.6.148}$$

Therefore, for  $\Delta > 0$

$$P(\max\{\nu_k : 1 \leq k \leq d\} > \Delta) \leq d\left(\frac{1}{1 - \frac{2\lambda\rho^2(\Sigma)}{n}}\right)^{\frac{n}{2}} \exp(-\lambda\Delta). \tag{A.6.149}$$

Take  $\lambda = \frac{n}{4\rho^2(\Sigma)}$ , and  $\Delta = 2\rho^2(\Sigma) \log 2 + \frac{4\rho^2(\Sigma)}{n} \log d + \epsilon$ , we have

$$P(\max\{\nu_k : 1 \leq k \leq d\} > \Delta) \leq \exp\left(-\frac{n}{4\rho^2(\Sigma)}\epsilon\right). \quad (\text{A.6.150})$$

Therefore, the proof of the first statement is concluded.

For the second statement, suppose  $\max\{\nu_k : 1 \leq k \leq d\} \leq C_\nu$ . Then we have for  $\lambda > 0$ ,

$$\mathbb{E}\left(\exp\left(\lambda \max\left\{\left|\frac{\tilde{X}_k^T w}{n}\right| : 1 \leq k \leq d\right\}\right)\right) \leq 2d \exp\left(\frac{\lambda^2}{n} C_\nu^2 \sigma^2 / 2\right). \quad (\text{A.6.151})$$

Therefore, for  $\Delta > 0$ ,

$$P\left(\left\|\frac{\mathbf{X}w}{n}\right\|_\infty > \Delta\right) \leq \exp\left(\log(2d) + \frac{\lambda^2}{n} C_\nu^2 \sigma^2 / 2 - \lambda\Delta\right). \quad (\text{A.6.152})$$

Take  $\lambda = \frac{n\Delta}{C_\nu^2 \sigma^2}$ , we have

$$P\left(\left\|\frac{\mathbf{X}w}{n}\right\|_\infty > \Delta\right) \leq \exp\left(\log(2d) - \frac{n\Delta^2}{2C_\nu^2 \sigma^2}\right). \quad (\text{A.6.153})$$

Setting

$$\Delta = C_\nu \sigma \sqrt{\frac{2 \log(2d)}{n}} + \mu, \quad (\text{A.6.154})$$

and note that  $C_\nu \leq \sqrt{4\rho^2(\Sigma) + 4\rho^2(\Sigma) \frac{\log d}{n}}$  with probability at least  $1 - \exp(-\frac{n}{2})$ , we have the statement of second inequality of the lemma.

#### A.6.16. Proof of Theorem 4.5.2

It's easy to check that

$$\frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 \quad (\text{A.6.155})$$

is  $\left\|\frac{\mathbf{X}^T \mathbf{X}}{n}\right\|_s$ -smooth.

By Theorem 4.2.1, and take  $\delta_0 = 0$  gives the result.

### A.6.17. Proof of Theorem 4.5.3

It's easy to see that  $\frac{1}{2n}\|\mathbf{X}\theta\|_2^2$  is  $\frac{\|\mathbf{X}^T\mathbf{X}\|_s}{n}$ -smooth, where  $\|\cdot\|_s$  denotes the spectral norm.

Denote  $L = \frac{\|\mathbf{X}^T\mathbf{X}\|_s}{n}$ .

Note that we have an alternative expression for  $\theta_{k+1}$  for  $k \geq 0$ :

$$\theta_{k+1} = \arg \min_{\theta} \frac{1}{2n}\|\mathbf{X}\theta_k\|_2^2 + \left\langle \frac{\mathbf{X}^T\mathbf{X}\theta_k}{n}, \theta - \theta_k \right\rangle + \frac{L}{2}\|\theta - \theta_k\|_2^2 + \lambda_n\|\theta\|_1. \quad (\text{A.6.156})$$

For simplicity we define

$$\phi_k(\theta) = \frac{1}{2n}\|\mathbf{X}\theta_k\|_2^2 + \left\langle \frac{\mathbf{X}^T\mathbf{X}\theta_k}{n}, \theta - \theta_k \right\rangle + \frac{L}{2}\|\theta - \theta_k\|_2^2 + \lambda_n\|\theta\|_1. \quad (\text{A.6.157})$$

Theorem 10.16 in Beck (2017) gives that

$$F(\theta) - F(\theta_{k+1}) \geq \frac{L}{2}\|\theta - \theta_{k+1}\|_2^2 - \frac{L}{2}\|\theta - \theta_k\|_2^2 + D(\theta, \theta_k), \quad (\text{A.6.158})$$

where

$$D(\theta, \theta_k) = \frac{1}{2n}\|\mathbf{X}\theta\|_2^2 - \frac{1}{2n}\|\mathbf{X}\theta_k\|_2^2 - \left\langle \frac{\mathbf{X}^T\mathbf{X}\theta_k}{n}, \theta - \theta_k \right\rangle. \quad (\text{A.6.159})$$

Taking  $\theta = \theta_k$  gives

$$F(\theta_k) \geq F(\theta_{k+1}) + \frac{L}{2}\|\theta_k - \theta_{k+1}\|_2^2. \quad (\text{A.6.160})$$

Taking  $\theta = \theta^*$  gives

$$F(\theta^*) - F(\theta_{k+1}) \geq \frac{L}{2}\|\theta^* - \theta_{k+1}\|_2^2 - \frac{L}{2}\|\theta^* - \theta_k\|_2^2 + D(\theta^*, \theta_k). \quad (\text{A.6.161})$$



Adding up the inequality from 1 to  $k + 1$  gives

$$\frac{L}{2} \|\theta^*\|_2^2 \geq \sum_{j=1}^{k+1} F(\theta_j) - F(\theta^*) \geq (k+1)(F(\theta_{k+1}) - F(\theta^*)). \quad (\text{A.6.162})$$

Taking  $\theta = \hat{\theta}$ , gives

$$F(\hat{\theta}) - F(\theta_{k+1}) \geq \frac{L}{2} \|\hat{\theta} - \theta_{k+1}\|_2^2 - \frac{L}{2} \|\hat{\theta} - \theta_k\|_2^2 + D(\hat{\theta}, \theta_k). \quad (\text{A.6.163})$$

Adding up the inequality from 1 to  $k + 1$  gives

$$F(\theta_{k+1}) - F(\hat{\theta}) \leq \frac{1}{k+1} \frac{L}{2} \|\hat{\theta}\|_2^2. \quad (\text{A.6.164})$$

This gives the second statement of the theorem.

Recalling Inequality (A.6.143), we have that

$$0 \leq 3 \|(\theta_k - \theta^*)_S\|_1 - \|(\theta_k - \theta^*)_{S^c}\|_1 + 4 \|\theta_{S^c}^*\|_1 + \frac{2(F(\theta_k) - F(\theta^*))}{\lambda_n}. \quad (\text{A.6.165})$$

This gives

$$\|\theta_k - \theta^*\|_1 \leq 4\sqrt{s} \|\theta_k - \theta^*\|_2 + 4 \|\theta_{S^c}^*\|_1 + \frac{2(F(\theta_k) - F(\theta^*))}{\lambda_n}. \quad (\text{A.6.166})$$

Therefore

$$\|\hat{\theta} - \theta_k\|_1 \leq \|\theta_k - \theta^*\|_1 + \|\hat{\theta} - \theta^*\|_1 \leq 4\sqrt{s} \|\hat{\theta} - \theta_k\|_2 + 8\sqrt{s} \|\hat{\theta} - \theta^*\|_2 + 8 \|\theta_{S^c}^*\|_1 + \frac{2(F(\theta_k) - F(\theta^*))}{\lambda_n}. \quad (\text{A.6.167})$$

Recall the definition of  $\phi_k(\theta)$  in Equation (A.6.157). For  $0 < \alpha < 1$ , we have

$$\begin{aligned}
F(\theta_{k+1}) &\leq \phi_k(\theta_{k+1}) \leq \phi_k(\alpha\hat{\theta} + (1-\alpha)\theta_k) \\
&\leq \frac{1}{2n}\|\mathbf{X}\theta_k\|_2^2 + \alpha\left\langle \frac{\mathbf{X}^T\mathbf{X}\theta_k}{n}, \hat{\theta} - \theta_k \right\rangle + \frac{L\alpha^2}{2}\|\hat{\theta} - \theta_k\|_2^2 + \alpha\lambda_n\|\hat{\theta}\|_1 + (1-\alpha)\lambda_n\|\theta_k\|_1 \\
&\leq \alpha F(\hat{\theta}) + (1-\alpha)F(\theta_k) + \frac{L\alpha^2}{2}\|\theta_k - \hat{\theta}\|_2^2.
\end{aligned} \tag{A.6.168}$$

Now we will bound  $\|\theta_k - \hat{\theta}\|_2^2$ .

Note that  $\hat{\theta}$  is the minimizer of  $F(\theta)$ , we have

$$\begin{aligned}
F(\theta_k) - F(\hat{\theta}) &= F(\theta_k) - F(\hat{\theta}) - \langle \partial F(\hat{\theta}), \theta_k - \hat{\theta} \rangle \geq D(\theta_k, \hat{\theta}) \geq \frac{a_1}{2}\|\hat{\theta} - \theta_k\|_2^2 - \frac{a_2}{2}\|\hat{\theta} - \theta_k\|_1^2 \\
&\geq \frac{a_1}{2}\|\hat{\theta} - \theta_k\|_2^2 - \frac{a_2}{2}\left(4\sqrt{s}\|\theta_k - \hat{\theta}\|_2 + 8\sqrt{s}\|\hat{\theta} - \theta^*\|_2 + 8\|\theta_{S^c}^*\|_1 + \frac{2(F(\theta_k) - F(\theta^*))}{\lambda_n}\right)_+^2
\end{aligned} \tag{A.6.169}$$

Since  $a_1 \geq 64s \cdot a_2$ , we have

$$\frac{a_1}{4}\|\hat{\theta} - \theta_k\|_2^2 \leq F(\theta_k) - F(\hat{\theta}) + a_2\left(8\sqrt{s}\|\hat{\theta} - \theta^*\|_2 + 8\|\theta_{S^c}^*\|_1 + \frac{2(F(\theta_k) - F(\theta^*))}{\lambda_n}\right)_+^2. \tag{A.6.170}$$

Therefore

$$\begin{aligned}
\|\hat{\theta} - \theta_k\|_2^2 &\leq \frac{4}{a_1}\left(F(\theta_k) - F(\hat{\theta})\right) + \frac{4a_2}{a_1} \cdot 128\left(\sqrt{s}\|\hat{\theta} - \theta^*\|_2 + \|\theta_{S^c}^*\|_1\right)^2 \\
&\quad + \frac{32a_2}{a_1}\left(\frac{F(\theta_k) - F(\theta^*)}{\lambda_n}\right)_+^2.
\end{aligned} \tag{A.6.171}$$

Let  $\alpha = \frac{a_1}{4L}$  in Inequality (A.6.168), we have that

$$\begin{aligned}
F(\theta_{k+1}) - F(\hat{\theta}) &\leq \left(1 - \frac{a_1}{8L}\right) \left(F(\theta_k) - F(\hat{\theta})\right) + \\
&\quad \frac{a_1}{4L} \cdot 64a_2s \cdot \left(\|\hat{\theta} - \theta^*\|_2 + \frac{\|\theta_{S^c}^*\|_1}{\sqrt{s}}\right)^2 + \frac{a_1 \cdot 64a_2s}{64L \cdot s} \left(\frac{F(\theta_k) - F(\theta^*)}{\lambda_n}\right)_+^2.
\end{aligned} \tag{A.6.172}$$

From Theorem 4.5.1 we have that

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{\|\theta_{S^c}^*\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1\kappa}. \tag{A.6.173}$$

Plug in Inequality (A.6.173) into Inequality (A.6.172) and note that  $F(\theta_k) - F(\theta^*) \leq F(\theta_k) - F(\hat{\theta}) \leq F(\theta_K) - F(\hat{\theta})$  for  $K \leq k$  gives the statement.

## BIBLIOGRAPHY

- A. Agarwal, D. P. Foster, D. J. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *Advances in Neural Information Processing Systems*, 24, 2011.
- S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116:1716–1730, 2021.
- P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In N. H. Bshouty and C. Gentile, editors, *Learning Theory*, pages 454–468, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-72927-3.
- A. Beck. *First-order methods in optimization*. SIAM, 2017.
- E. Belitser, S. Ghosal, H. van Zanten, et al. Optimal two-stage procedures for estimating location and size of the maximum of a multivariate regression function. *The Annals of Statistics*, 40(6):2850–2876, 2012.
- E. Belitser, S. Ghosal, and H. van Zanten. Correction note: Optimal two-stage procedures for estimating location and size of the maximum of a multivariate regression function. *Ann. Statist.* 40 (2012) 2850–2876. *The Annals of Statistics*, 49(1):612 – 613, 2021. doi: 10.1214/20-AOS1993. URL <https://doi.org/10.1214/20-AOS1993>.
- Q. Berthet and P. Rigollet. Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828*, 2013.
- L. Birge. The Grenader estimator: A nonasymptotic approach. *The Annals of Statistics*, 17(4):1532–1549, 1989. doi: 10.1214/aos/1176347380. URL <https://doi.org/10.1214/aos/1176347380>.
- J. R. Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, 25(4):737–744, 1954. doi: 10.1214/aoms/1177728659. URL <https://doi.org/10.1214/aoms/1177728659>.
- L. Bottou and O. Bousquet. 13 the tradeoffs of large-scale learning. *Optimization for machine learning*, page 351, 2011.
- S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- T. T. Cai and Z. Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.
- T. T. Cai and M. G. Low. A framework for estimation of convex functions. *Statistica Sinica*, pages 423–456, 2015.

- T. T. Cai, M. G. Low, and Y. Xia. Adaptive confidence intervals for regression functions under shape constraints. *The Annals of Statistics*, 41(2):722–750, 2013.
- X. Cai, D. Han, and X. Yuan. On the convergence of the direct extension of admm for three-block separable convex minimization models with one strongly convex function. *Computational Optimization and Applications*, 66(1):39–73, 2017.
- V. Chandrasekaran and M. I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.
- S. Chatterjee, J. C. Duchi, J. Lafferty, and Y. Zhu. Local minimax complexity of stochastic convex optimization. *Advances in Neural Information Processing Systems*, 29, 2016. URL <https://proceedings.neurips.cc/paper/2016/file/b9f94c77652c9a76fc8a442748cd54bd-Paper.pdf>.
- C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1):57–79, 2016.
- H. Chen. Lower rate of convergence for locating a maximum of a function. *The Annals of Statistics*, 16(3):1330–1334, 1988. doi: 10.1214/aos/1176350965. URL <https://doi.org/10.1214/aos/1176350965>.
- H. Chen, M.-N. L. Huang, and W.-J. Huang. Estimation of the location of the maximum of a regression function using extreme order statistics. *Journal of Multivariate Analysis*, 57(2):191–214, 1996.
- Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.
- Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.
- M. A. Davenport, Y. Plan, E. Van Den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.
- H. Deng, Q. Han, and B. Sen. Inference for local parameters in convexity constrained models. *arXiv preprint arXiv:2006.10264*, 2020.
- J. Dippon. Accelerated randomized stochastic optimization. *The Annals of Statistics*, 31(4):1260–1281, 2003.
- L. Dumbgen. New goodness-of-fit tests and their application to nonparametric confidence sets. *The Annals of Statistics*, 26(1):288–314, 1998. ISSN 00905364. URL <http://www.jstor.org/stable/119988>.
- M. R. Facer and H.-G. Müller. Nonparametric estimation of the location of a maximum in a response surface. *Journal of Multivariate Analysis*, 87(1):191–217, 2003. ISSN 0047-259X.

- doi: [https://doi.org/10.1016/S0047-259X\(03\)00030-7](https://doi.org/10.1016/S0047-259X(03)00030-7). URL <https://www.sciencedirect.com/science/article/pii/S0047259X03000307>.
- P. Ghosal and B. Sen. On univariate convex regression. *Sankhya A*, 79(2):215–253, Aug 2017. ISSN 0976-8378. doi: 10.1007/s13171-017-0104-8. URL <https://doi.org/10.1007/s13171-017-0104-8>.
- D. J. Griffiths and D. F. Schroeter. *Introduction to quantum mechanics*. Cambridge University Press, 2018.
- A. Guntuboyina, B. Sen, et al. Nonparametric shape-restricted regression. *Statistical Science*, 33(4):568–594, 2018.
- N. W. Hengartner and P. B. Stark. Finite-sample confidence envelopes for shape-restricted densities. *The Annals of Statistics*, 23(2):525–550, 1995. doi: 10.1214/aos/1176324534. URL <https://doi.org/10.1214/aos/1176324534>.
- M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199, 2017.
- I. Horev, B. Nadler, E. Arias-Castro, M. Galun, and R. Basri. Detection of long edges on a computational budget: A sublinear approach. *SIAM Journal on Imaging Sciences*, 8(1):458–483, 2015.
- P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.
- K. Jiang, D. Sun, and K.-C. Toh. An inexact accelerated proximal gradient method for large scale linearly constrained convex sdp. *SIAM Journal on Optimization*, 22(3):1042–1064, 2012.
- J. Kiefer. Optimum rates for non-parametric density and regression estimates under order restrictions. *Statistics and Probability: Essays in honor of CR Rao*, 419:428, 1982.
- J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.
- R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Proceedings of the 17th International Conference on Neural Information Processing Systems, NIPS’04*, pages 697–704, Cambridge, MA, USA, 2004. MIT Press.
- R. Kleinberg, A. Slivkins, and E. Upfal. Bandits and experts in metric spaces. *J. ACM*, 66(4), May 2019. ISSN 0004-5411. doi: 10.1145/3299873. URL <https://doi.org/10.1145/3299873>.
- S. Kpotufe and N. Verma. Time-accuracy tradeoffs in kernel prediction: controlling prediction quality. *The Journal of Machine Learning Research*, 18(1):1443–1471, 2017.

- M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.
- T. Lin, S. Ma, and S. Zhang. Iteration complexity analysis of multi-block admm for a family of convex minimization without strong convexity. *Journal of Scientific Computing*, 69(1): 52–81, 2016.
- T. Lin, S. Ma, and S. Zhang. Global convergence of unmodified 3-block admm for a class of convex minimization problems. *Journal of Scientific Computing*, 76(1):69–88, 2018.
- P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.
- A. Mokkadem and M. Pelletier. A companion for the Kiefer–Wolfowitz–Blum stochastic approximation algorithm. *The Annals of Statistics*, 35(4):1749–1772, 2007.
- H.-G. Muller. Adaptive nonparametric peak estimation. *The Annals of Statistics*, pages 1053–1069, 1989.
- B. T. Polyak and A. B. Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, 26(2):45–53, 1990.
- M. Schmidt, N. L. Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *arXiv preprint arXiv:1109.2415*, 2011.
- Y. Seginer. The expected norm of random matrices. *Combinatorics, Probability and Computing*, 9(2):149–166, 2000.
- D. Shender and J. Lafferty. Computation-risk tradeoffs for covariance-thresholded regression. In *International Conference on Machine Learning*, pages 756–764. PMLR, 2013.
- J.-M. Shoung, C.-H. Zhang, et al. Least squares estimators of the mode of a unimodal regression function. *The Annals of Statistics*, 29(3):648–665, 2001.
- A. Simonetto. Smooth strongly convex regression. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 2130–2134. IEEE, 2021.
- D. L. Sussman, A. Volfovsky, and E. M. Airoldi. Analyzing statistical and computational tradeoffs of estimation procedures. *arXiv preprint arXiv:1506.07925*, 2015.
- R. J. Tibshirani. Dykstra’s algorithm, admm, and coordinate descent: Connections, insights, and extensions. *arXiv preprint arXiv:1705.04768*, 2017.
- J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

- J. A. Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.
- M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- L. Wang, X. Zhang, and Q. Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. In *Artificial Intelligence and Statistics*, pages 981–990. PMLR, 2017.
- T. Wang, Q. Berthet, and R. J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896–1930, 2016.