Robust Estimation and Inference under Huber's Contamination Model

by

Peiliang Zhang

B.S. in Statistics, Nanjing University, China, 2017

Submitted to the Graduate Faculty of

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Peiliang Zhang

It was defended on

November 21st 2022

and approved by

Zhao Ren, Department of Statistics, University of Pittsburgh

Kehui Chen, Department of Statistics, University of Pittsburgh

Satish Iyengar, Department of Statistics, University of Pittsburgh

Wen-Xin Zhou, Department of Mathematics, University of California, San Diego

Robust Estimation and Inference under Huber's Contamination Model

Peiliang Zhang, PhD

University of Pittsburgh, 2022

Huber's contamination model is widely used for analyzing distributional robustness when the shape of the real underlying data distribution deviates from the assumed model. Specifically, it models that the observed data are contaminated by some arbitrary unknown distribution with a small fraction. In this dissertation, we study the robust regression and robust density estimation under Huber's contamination model.

In the regression setting, we assume that the noise has a heavy-tailed distribution and may be arbitrarily contaminated with a small fraction under an increasing dimension regime. We show that robust M-estimators can achieve the minimax convergence rate (except for the intercept if the uncontaminated distribution of the noise is asymmetric). We develop a multiplier bootstrap technique to construct confidence intervals for linear functionals of the coefficients. When the contamination proportion is relatively large, we further provide a bias correction procedure to alleviate the bias due to contamination. The robust estimation and inference framework can be extended to a distributed learning setting. Specifically, we demonstrate that a communication-efficient M-estimator can attain the centralized minimax rate (as if one has access to the entire data). Moreover, based on this communication-efficient M-estimator, a distributed multiplier bootstrap method is proposed only on the master machine, which is able to generate confidence intervals with optimal widths. A comprehensive simulation study demonstrates the effectiveness of our proposed procedures.

In the density estimation setting, we aim to robustly estimate a multivariate density function on R^d with L_p loss functions from contaminated data. To investigate the contamination effect on the optimal estimation of the density, we first establish the minimax rate with the assumption that the density is in an anisotropic Nikol'skii class. We then develop a data-driven bandwidth selection procedure for kernel estimators via a robust generalization of the Goldenshluger-Lepski method. We show that the proposed bandwidth selection rule can lead to the estimator being minimax adaptive to either the smoothness parameter or

the contamination proportion. When both of them are unknown, we prove that finding any minimax-rate adaptive method is impossible. Extensions to smooth contamination cases are also discussed.

Keywords: Robust statistics; Huber's contamination model; heavy-tailed distribution; M-estimation; multiplier bootstrap; communication-efficient estimator; distributed inference; minimax rate; adaptive density estimation; the Goldenshluger-Lepski method.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

1.1 Background

In this new era of data science, complex data structure, such as heterogeneity, contamination and heavy-tailedness of distributions, has posed unprecedented challenges to data analysts. Various robust statistical procedures are desired in the hope of accommodating such complexity and extracting useful information from modern datasets. Robust statistics, in its classical form, has been systematically studied since the seminal papers (Tukey, 1960; Huber, 1964; Hampel, 1968); see, for instance, Huber (2004) and Hampel et al. (2011) for a comprehensive introduction and discussion. Recently, notions from classical robust statistics have been revived to study robustness under various modern models. Examples include high-dimensional mean and covariance estimation (Diakonikolas et al., 2019; Lai et al., 2016; Chen et al., 2018), regression (Gao, 2020; Prasad et al., 2020; Pensia et al., 2020), nonparametric estimation (Liu and Gao, 2019; Chen et al., 2016), etc.

In addition to the widely used notions such as breakdown point and influence function, Huber's contamination model, first proposed in Huber (1964), has drawn considerable attention, as it formalizes a common setting where the distribution of the observed data deviates from the assumed model. Formally, it is defined as

$$(1 \quad \epsilon)P_\theta + \epsilon Q.$$

Under this model, data are drawn with probability $1 \quad \epsilon$ from the assumed modeling distribution P_θ (with θ as the parameter of interest), and with probability ϵ to be contaminated by some arbitrary distribution Q. The arbitrary contamination distribution Q can represent gross errors, adversarial corruption, or the outliers due to the underlying heavy-tailed distribution. From the perspective of classical robust statistics, this model depicts a circumstance where the real data distribution deviates from the assumed model P_θ within a radius of ϵ under the total variation distance, and thus desired robustness signifies insensitivity to such small deviations from the assumptions on the data distribution. Moreover, as discussed

in Chen et al. (2018), Huber's contamination model actually provides a unified framework to study the statistical efficiency and robustness simultaneously. Therefore, it gives us a more comprehensive view on the robustness evaluation, compared to the notions like breakdown point. That being said, rigorous theoretical studies of Huber's contamination model have been largely missing in modern topics such as the nonparametric and high-dimensional statistics—mainly due to its difficulty on balancing the trade-off among robustness, statistical efficiency and computation complexity—until some very recent works came up.

In the high-dimensional or increasing-dimensional setting where the dimensionality grows with the sample size, fundamental models like normal mean and covariance estimation were first carefully studied. Diakonikolas et al. (2016) and Lai et al. (2016) were the two pioneering works that proposed polynomial-time robust estimators with dimension-independent error guarantees (regarding the contamination dependence, yet still sub-optimal). Since then, there have been a substantial number of works in theoretical computer science communities trying to improve the results, either in computing time (polynomial or nearly-linear time) or contamination dependence (optimal or near-optimal) (Kothari and Steurer, 2017; Charikar et al., 2017; Diakonikolas et al., 2017, 2018; Cheng et al., 2019). Another line of research is Chen et al. (2018) and Gao (2020), which generalized the idea of Tukey's depth and proved that the depth-based estimators they designed are able to achieve the optimal rate of convergence in many statistical tasks, such as covariance matrix estimation and multivariate regressions, yet with computation intractability issue. With advancements in robust mean and covariance estimation, studies on robust regression began to emerge. Dalalyan and Thompson (2019) and Sasai and Fujisawa (2020) used different techniques to prove that a robust M-estimator with certain ℓ_1 regularization is minimax rate-optimal. Klivans et al. (2018) and Bakshi and Prasad (2020) both designed polynomial-time robust estimators (the latter can provably achieve the optimal convergence rate) via sum-of-squares approaches under a moment condition called hypercontractivity, while Prasad et al. (2020) provided a general class of robust estimators via robust gradient estimation under the same condition. More recently, Pensia et al. (2020) combined classical robust regression methods, such as Huber regression and least trimmed squares, with a covariate filtering technique to obtain optimal estimators under the setting where both covariates and responses are potentially

heavy-tailed and contaminated.

Going through all these literatures, we find that although various optimal or near optimal robust estimators are provided under Huber's contamination model, none of them have discussed how to do statistical inference with the given estimator, which is critical in the decision-making processes. Motivated by the absence of inference under Huber's contamination model, my first part of dissertation will focus on the statistical inference task with contaminated data under an increasing-dimensional regression setting. We show that a robust M-estimator can achieve the optimal convegence rate (except for the intercept when the uncontaminated distribution is asymmetric), and a valid confidence interval can be obtained via a multiplier bootstrap technique. The robust estimation and inference framework is further extended to a distributed context, where the overall data are split across multiple machines, and communication between machines is constrained.

In the nonparametric setting, Du et al. (2018) and Gao (2020) both gave optimal estimation procedures on nonparametric regression tasks, using a local binning median method and a regression depth method, respectively. In the area of nonparametric density estimation, Liu and Gao (2019) investigated the optimal convergence rates and adaptation theory for a density on R under a pointwise loss. Chen et al. (2016) developed a general decision theory and proposed a class of estimators based on Scheffé estimate, which are shown to be optimal when the loss is equivalent to the total variation distance. Noticeably, the results on the robust estimation of a multivariate density function on R^d from contaminated data are still largely unknown. The second part of my dissertation is aimed to fill in this blank, by establishing the optimal adaptive procedures under a general L_p (1 $p <$ 1) loss.

1.2 Overview of the Dissertation

This dissertation studies the robust linear regression inference and robust density estimation problem under Huber's contamination model.

In Chapter 2, we study the robust estimation and inference problem for linear models in that the increasing dimension regime. Given random design, we consider the conditional

distributions of error terms are contaminated by some arbitrary distribution with proportion ϵ but otherwise can also be heavy-tailed and asymmetric. Under the setting of Huber's contamination model, we prove that simple robust M-estimators like Huber's estimator can still achieve the optimal minimax rate of convergence except for the intercept. In addition, we generate confidence intervals for linear functionals of the coefficients by a multiplier bootstrap technique. The non-asymptotic theoretical guarantee is established when the necessary condition on contamination proportion $\epsilon = o(1/\sqrt{n})$ holds, where n is the sample size. For a larger ϵ, we further propose a debiasing procedure to reduce the potential bias caused by contamination, and prove the validity of the debiased confidence interval as long as $\epsilon = o(1)$.

In Chapter 3, we extend the above ideas to the distributed estimation and inference setting. Specifically, we demonstrate that a communication-efficient M-estimator can attain the centralized minimax rate (as if one has access to the entire data) with the distributed contaminated data. Moreover, based on the proposed communication-efficient M-estimator, a distributed multiplier bootstrap procedure is further developed. It can be conducted only on the master machine, and thus incurs no extra communication costs. While this procedure only bootstraps the local sample data stored on the master machine, it leverages the information aggregated from other machines in the distributed estimation phase. Therefore, the resulting confidence interval has the optimal width. We establish its theoretical validity under weak assumptions on the local sample size and the contamination proportion. A comprehensive simulation study demonstrates the effectiveness of our proposed procedures.

In Chapter 4, we address the problem of density function estimation in $\mathbb{R}^d$ with L_p losses ($1 \leq p < \infty$) under Huber's contamination model. We investigate the effects of contamination proportion ϵ among other key quantities on the minimax rates of convergence for both structured and unstructured contamination over a scale of the anisotropic Nikol'skii classes. The corresponding adaptation theory is further studied by establishing L_p risk oracle inequalities via a novel generalization of the Goldenshluger-Lepski method. Specifically, we develop a data-driven bandwidth selection procedure for kernel estimators which can lead to the estimator being minimax adaptive to either the smoothness parameter or the contamination proportion. When both of them are unknown, we prove that it is impossible

to find any minimax-rate adaptive method. This illustrates the contamination effect on the adaptation theory of density estimation.

In this dissertation, notations may differ from chapter to chapter, and will be introduced in the first section of each chapter. Each chapter's supplement has the same notations as the main chapter.

2.0 Robust Regression under Huber's Contamination Model

2.1 Preliminaries

In this chapter, we consider a regression setting and assume that the conditional distribution of the response variable $y \in \mathbb{R}$ given covariates $\boldsymbol{x} \in \mathbb{R}^d$ follows Huber's contamination model. That is,

$$y = \beta_0 + \boldsymbol{x}^\intercal \boldsymbol{\beta} + \varepsilon, \quad \varepsilon | \boldsymbol{x} \sim (1-\epsilon)F + \epsilon G_{\boldsymbol{x}}. \tag{2.1.1}$$

Besides the independence on $\boldsymbol{x}$, we make very few weak assumptions on the distribution F (see Condition 2.2.2) such that F can be heavy-tailed. Under our model, the contamination distribution $G_{\boldsymbol{x}}$ is arbitrary and may depend on the covariates $\boldsymbol{x}$, which can be used to model the adversarial manipulations of the response based on its features. We consider an increasing dimension scheme: the dimensionality of features d can grow with the sample size (but at a slower rate). Under this framework, the goal of this work is two folds: (i) robustly estimate the coefficients $\boldsymbol{\beta}$ with the optimal estimation error; (ii) build valid confidence intervals for linear functionals of $\boldsymbol{\beta}$ with the optimal width. We present the robust estimation and inference results in Section 2.2 and Section 2.3, respectively. In Section 2.4, we propose a debiased confidence interval method for cases with a large contamination proportion. Section 2.5 reports our experimental results for both estimation and confidence interval coverage and width. The proofs for all the theoretical results are given in Appendix A.

Notations. In this and the following chapter, we use bold letters like $\boldsymbol{u}$, $\boldsymbol{A}$ to represent vectors and matrices. Given any two vectors $\boldsymbol{u} = (u_1, \ldots, u_k)^\intercal, \boldsymbol{v} = (v_1, \ldots, v_k)^\intercal \in \mathbb{R}^k$, we define their inner product by $\boldsymbol{u}^\intercal \boldsymbol{v} = \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^k u_i v_i$. We use the notation $\|\boldsymbol{u}\|_p, 1 \le p < \infty$ for the ℓ_p-norms of vectors in $\mathbb{R}^k$: $\|\boldsymbol{u}\|_p = (\sum_{i=1}^k |u_i|^p)^{1/p}$. For $k \ge 2$, $\mathbb{S}^{k-1} = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 = 1\}$ denotes the unit sphere in $\mathbb{R}^k$. For a positive semi-definite matrix $\boldsymbol{A} \in \mathbb{R}^{k \times k}$, $\|\cdot\|_{\boldsymbol{A}}$ denotes the norm induced by $\boldsymbol{A}$, that is, $\|\boldsymbol{u}\|_{\boldsymbol{A}} = \|\boldsymbol{A}^{1/2}\boldsymbol{u}\|_2, \boldsymbol{u} \in \mathbb{R}^k$. For any two real number $a, b \in \mathbb{R}$, we write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We use $\lfloor a \rfloor$ to denote the largest integer strictly less than a. For two positive sequences $\{a_n\}$ and

$f b_n g$, we write $a_n \ . \ b_n$ or $a_n = O(b_n)$ if $a_n \quad Cb_n$ for all n and some positive constant C independent of n. We write $a_n \ \& \ b_n$ or $a_n = \Omega(b_n)$ if $b_n \ . \ a_n$, and write $a_n \quad b_n$ if $a_n \ . \ b_n$ and $b_n \ . \ a_n$. We use $a_n = o(b_n)$ or $b_n \quad a_n$ to denote $a_n/b_n \ ! \ 0$ when $n \ ! \ 1$.

2.2 Robust M-estimation

Assume we have access to n observations $f(\boldsymbol{x}_i, y_i) g_{i=1}^n$ *i.i.d.* from $(\boldsymbol{x}, y)$ that follows the linear model with contamination (2.1.1), where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\intercal$ represents the unknown regression coefficients, $\boldsymbol{x} = (x_1, \ldots, x_d)^\intercal$ is the covariate vector, and ε is the regression noise. In the following, we denote $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\intercal)^\intercal$, $\bar{\boldsymbol{x}} = (1, \boldsymbol{x}^\intercal)^\intercal$ and write the linear model as $y = \bar{\boldsymbol{x}}^\intercal \boldsymbol{\theta} + \varepsilon$. Under Huber's ϵ-contamination model (2.1.1), the response y is contaminated with probability ϵ by some arbitrary distribution $G_{\boldsymbol{x}}$ that could depend on $\boldsymbol{x}$. Moreover, we do not impose any restrictive moment conditions on the "true" underlying distribution F so that it can model the heavy-tailed distribution. To robustly estimate the coefficients with the contaminated and heavy-tailed errors, we consider M-estimators Ronchetti and Huber (2009), which are defined as

$$\widehat{\boldsymbol{\theta}}_\tau = (\widehat{\beta}_{0,\tau}, \widehat{\boldsymbol{\beta}}_\tau^\intercal)^\intercal \ 2 \ \underset{\boldsymbol{\theta} 2 \mathsf{R}^{d+1}}{\operatorname{argmin}} L_\tau(\boldsymbol{\theta}), \text{ with } L_\tau(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \ell_\tau(y_i \quad \bar{\boldsymbol{x}}_i^\intercal \boldsymbol{\theta}), \tag{2.2.1}$$

for a given loss function $\ell(\ )$ and a tuning parameter τ, where $\ell_\tau(\ ) = \tau^2 \ell(\ /\tau)$. The population minimizer of the above loss under the distribution F is denoted by

$$\boldsymbol{\theta}_\tau = (\beta_{0,\tau}, \boldsymbol{\beta}_\tau^\intercal)^\intercal \ 2 \ \underset{\boldsymbol{\theta} 2 \mathsf{R}^{d+1}}{\operatorname{argmin}} \mathsf{E}_{\boldsymbol{x},\varepsilon \ F} f \ell_\tau(y \quad \bar{\boldsymbol{x}}^\intercal \boldsymbol{\theta}) g. \tag{2.2.2}$$

M-estimators were first proposed by Huber in his seminal work Huber (1964), to study robustness against distributional contamination. Recent research Chen and Zhou (2020); Sun et al. (2020); Zhou et al. (2018) has demonstrated that M-estimators, like Huber estimators, are also robust to heavy-tailed data in the sense that a sub-Gaussian type deviation bound can be obtained, while only assuming the existence of a finite second moment. Specifically,

they prove that when there is no contamination ($\epsilon = 0$), under standard regularity conditions, it holds

$$\|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}\|_2 \leq \underbrace{\|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau\|_2}_{\text{Estimation Error (Variance)}} + \underbrace{\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}\|_2}_{\text{Approximation Error (Bias)}} \cdot \left(\sigma\sqrt{\frac{d}{n}} + \tau\frac{d}{n}\right) + \frac{\sigma^2}{\tau},$$

where σ^2 is the variance of F, and the loss function is Huber loss

$$\ell_\tau(x) = \begin{cases} x^2/2 & \text{if } |x| \leq \tau, \\ \tau|x| - \tau^2/2 & \text{if } |x| > \tau. \end{cases} \tag{2.2.3}$$

Notably, a bias, resulting from the discrepancy between the population minimizer and the true parameter, arises when the heavy-tailed distribution F is asymmetric. Under this case, an adaptive choice of τ at the order $\sigma\sqrt{n/d}$ is suggested to achieve the optimal convergence rate $O(\sqrt{d/n})$.

With the presence of contamination ($\epsilon > 0$), however, the estimation error $\|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau\|_2$ would also involve bias since $\boldsymbol{\theta}_\tau$ is no longer the population minimizer with contamination distribution $G_{\boldsymbol{x}}$. In other words, instead of the above error decomposition, now we have

$$\|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}\|_2 \leq \underbrace{\|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau\|_2}_{\text{Estimation Error (Variance+Contamination Bias)}} + \underbrace{\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}\|_2}_{\text{Approximation Error (Bias)}} \cdot \left(\sigma\sqrt{\frac{d}{n}} + \tau\frac{d}{n} + \epsilon(\tau \wedge 1)\right) + \frac{\sigma^2}{\tau},$$

where the new bias term $\epsilon(\tau \wedge 1)$ origins from the fact that $\|\mathbb{E}_{\boldsymbol{x},\varepsilon\sim(1-\epsilon)F+\epsilon G_{\boldsymbol{x}}} \nabla L_\tau(\boldsymbol{\beta}_F)\|$ is non-zero when $\epsilon > 0$, but has a bounded norm of order $\epsilon\tau$, assuming that the loss function $\ell(\,)$ has a bounded derivative. With $\tau \asymp \sigma(d/n + \epsilon)^{1/2}$, we get the optimized total error $O(\sqrt{d/n + \epsilon})$, which is still slower than the optimal minimax error rate $O(\sqrt{d/n} + \epsilon)$ as shown in Theorem 2.2.2 below. This exhibits the contamination effect on the robustness theory of M-estimators.

However, if we are primarily interested in estimating the coefficients $\boldsymbol{\beta}$, this is not the end of the story. The following Proposition 4.3.1 reveals that under mild conditions on the loss function $\ell(\,)$ and the distribution F, the population minimizer $\boldsymbol{\theta}_\tau$ is unique and its coefficient part $\boldsymbol{\beta}_\tau$ coincides with $\boldsymbol{\beta}$. In other words, for the estimation of $\boldsymbol{\beta}$, the bias from the approximation error $\|\boldsymbol{\beta}_\tau - \boldsymbol{\beta}\|_2$ vanishes, allowing the M-estimator $\widehat{\boldsymbol{\beta}}_\tau$ to attain an improved rate of convergence.

Condition 2.2.1 (Globally Lipschitz and locally quadratic loss function)**.** Let ℓ : R 7! [0, 1) be a convex function that satisfies: (i) $\ell'(0) = 0$ and $j\ell'(x)j$ c_1 for all x 2 R, and (ii) $\ell''(0) = 1$ and $\ell''(x)$ c_2 for all jxj c_3, where c_1–c_3 are positive constants.

Proposition 2.2.1. Assume that $\mathbf{S} = \mathsf{E}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathsf{T}})$ is positive definite. For a given loss function ℓ satisfying Condition 2.2.1 and $\tau > 0$, assume that the function α 7! $\mathsf{E}_{\varepsilon \ F}f\ell_\tau(\varepsilon \quad \alpha)g$ has a unique minimizer, denoted by α_τ, which satisfies

$$\mathsf{P}_{\varepsilon \ F}(j\varepsilon \quad \alpha_\tau j \quad c_3\tau) > 0$$

where c_3 is given in Condition 2.2.1. Then for $\boldsymbol{\theta}_\tau$ in (2.2.2) and $\boldsymbol{\theta}$, it holds

$$\beta_{0,\tau} = \beta_0 + \alpha_\tau \quad \text{and} \quad \boldsymbol{\beta}_\tau = \boldsymbol{\beta} \ .$$

This proposition illustrates that with intercept added, the bias α_τ (i.e. the approximation error for the population minimizer $\boldsymbol{\theta}_\tau$) is retained only at the intercept term. It further implies that for the M-estimator $\widehat{\boldsymbol{\beta}}_\tau$ (with intercept added), we have

$$k\widehat{\boldsymbol{\beta}}_\tau \quad \boldsymbol{\beta} \ k_2 \qquad \underbrace{k\widehat{\boldsymbol{\beta}}_\tau \quad \boldsymbol{\beta}_\tau k_2}_{\text{Estimation Error (Variance+Contamination Bias)}} + \underbrace{k\boldsymbol{\beta}_\tau \quad \boldsymbol{\beta} \ k_2}_{\text{Approximation Error (= 0)}}$$
$$. \ \sigma\sqrt{\frac{d}{n}} + \tau\frac{d}{n} + \epsilon(\tau \_ 1),$$

and thus the minimax rate $\mathsf{O}(\sqrt{d/n} + \epsilon)$ can be achieved now with τ 1.

Table 1: M-Estimator (e.g. Huber Estimator) Analysis

Model	Bias	Variance	Optimal Rate
$\boldsymbol{\theta}$: Heavy-tail	σ^2/τ	$\sigma\sqrt{d/N} + \tau d/N$	$\sqrt{d/N}$
$\boldsymbol{\theta}$: Heavy-tail+Contamination	$\sigma^2/\tau + \epsilon(\tau \_ 1)$	$\sigma\sqrt{d/N} + \tau d/N$	$\sqrt{d/N + \epsilon}$
$\boldsymbol{\beta}$: Heavy-tail+Contamination	$\epsilon(\tau \_ 1)$	$\sigma\sqrt{d/N} + \tau d/N$	$\sqrt{d/N} + \epsilon$

The theoretical properties of M-estimators are formally presented in Theorem 2.2.1, under mild conditions on $\boldsymbol{x}$ and F, as summarized in Condition 2.2.2.

Condition 2.2.2. (i) $\mathbf{S} = \mathsf{E}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathsf{I}})$ is positive definite. $\boldsymbol{z} = \mathbf{S}^{1/2}\bar{\boldsymbol{x}}$ is a sub-Gaussian random vector with parameter $\nu_{\boldsymbol{x}}$. That is,

$$\nu_{\boldsymbol{x}} = \sup_{\boldsymbol{u} \in \mathbb{S}^d} \inf\{t > 0 : \mathsf{E}\exp\{\langle \boldsymbol{u}, \boldsymbol{z}\rangle^2/t^2\} \quad 2\}.$$

(ii) The noise variable $\varepsilon | \boldsymbol{x}$ $(1 \quad \epsilon)F + \epsilon G_{\boldsymbol{x}}$. F is independent of $\boldsymbol{x}$. (iii) Assume that the function $\alpha \mapsto \mathsf{E}_{\varepsilon \sim F}\{\ell_\tau(\varepsilon \quad \alpha)\}$ has a unique minimizer α_τ, and $\kappa_\tau := \mathsf{P}_{\varepsilon \sim F}(|\varepsilon \quad \alpha_\tau| \quad c_3\tau/2) > 0$, where c_3 is given in Condition 2.2.1.

Theorem 2.2.1. Assume Condition 2.2.1 and Condition 2.2.2 hold. Then, for any $t > 0$, with probability at least $1 \quad 2e^{-t}$, we have

$$\|\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau\|_{\mathbf{S}} \quad C\frac{c_1\tau\nu_{\boldsymbol{x}}}{c_2(1 \quad \epsilon)\kappa_\tau}\Big(\sqrt{\frac{d+t}{n}} + \epsilon\Big) \tag{2.2.4}$$

as long as $\epsilon < c\kappa_\tau$ and $n \quad C'\kappa_\tau^{-2}(d+t)$, where $c, C' > 0$ are two constants depending only on $(\nu_{\boldsymbol{x}}, c_1, c_2, c_3)$, and C is some absolute constant.

Remark 2.2.1. (i) In this theorem, we consider the prediction loss $\|\cdot\|_{\mathbf{S}}$, which is equivalent to ℓ_2 loss when $\mathbf{S}$ has bounded largest and smallest eigenvalues. Under such cases, Theorem 2.2.1 guarantees that the M-estimator satisfies

$$\|\widehat{\boldsymbol{\beta}}_\tau \quad \boldsymbol{\beta}\|_2 \ . \ \sqrt{\frac{d}{n}} + \epsilon$$

with a sub-Gaussian type concentration bound.

(ii) (About Condition 2.2.1) Examples of the loss function satisfying Condition 2.2.1 include Huber's loss (2.2.3) (with $c_1 = c_2 = c_3 = 1$) and various smoothed Huber loss and pseudo Huber loss functions; see Section A.1 in the supplement.

(iii) (About Condition 2.2.2) It is easy to check that the condition that the function $\alpha \mapsto \mathsf{E}_{\varepsilon \sim F}\{\ell_\tau(\varepsilon \quad \alpha)\}$ has a unique minimizer is satisfied in many cases, due to the globally convex and locally strong convex property of the loss function ℓ_τ.

(iv) (About the choice of τ) Theorem 2.2.1 indicates that we need $\kappa_\tau \gtrsim \epsilon + \sqrt{(d+t)/n}$ and $\tau/\kappa_\tau \to 1$ to reach the optimal rate $O(\sqrt{d/n} + \epsilon)$. To fulfill the requirements, one can select τ at the scale of a finite moment of the distribution F (if it exists). For example, if F is mean zero and has a finite variance σ^2, then with $\tau = C\sigma$, we have

$$\kappa_\tau = 1 - \mathsf{P}_{\varepsilon \sim F}(|\varepsilon - \alpha_\tau| > c_3\tau/2) \geq 1 - \frac{\mathsf{E}_{\varepsilon \sim F}|\varepsilon - \alpha_\tau|^2}{(c_3\tau/2)^2} \geq 1 - \frac{\sigma^2 + \alpha_\tau^2}{(c_3\tau/2)^2} \geq c > 0,$$

for some absolute constants $C, c > 0$. Here we use the fact that $|\alpha_\tau| \leq \sigma^2/\{\tau(1 - \sigma^2/\tau^2)\}$, which can be shown under Huber's loss.

Our next theorem shows that the above estimation error rate is optimal, by providing a minimax lower bound under Huber's contamination model.

Theorem 2.2.2. Let $\mathcal{P} = \mathcal{P}(\boldsymbol{\theta}(\cdot), \epsilon, F, G_{\boldsymbol{x}})$ to denote the collection of all the joint distributions of $(\boldsymbol{x}, y)$ that follow model (2.1.1) and satisfy Condition 2.2.2, then we have

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \sup_{P(\boldsymbol{x},y) \in \mathcal{P}} \mathsf{E}_{P(\boldsymbol{x},y)} \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\mathbf{S}} \gtrsim \sqrt{\frac{d}{n}} + \epsilon.$$

Moreover, denote $\widetilde{\mathcal{P}} = \{P(\boldsymbol{x}, y) \in \mathcal{P} : c_{\mathbf{S}} \leq \lambda_{\min}(\mathbf{S}) \leq \lambda_{\max}(\mathbf{S}) \leq C_{\mathbf{S}}\}$ with constants $c_{\mathbf{S}}, C_{\mathbf{S}} > 0$, then we have

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta} \in \mathbb{R}^d} \sup_{P(\boldsymbol{x},y) \in \widetilde{\mathcal{P}}} \mathsf{E}_{P(\boldsymbol{x},y)} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \gtrsim \sqrt{\frac{d}{n}} + \epsilon.$$

Remark 2.2.2. The above information-theoretical lower bound can be seen as an application of Theorem 5.1 in Chen et al. (2018) to the conditional distribution of $y|\boldsymbol{x}$. The main idea is based on the observation that Huber's ϵ-contamination model of the form $(1-\epsilon)P_\theta + \epsilon Q$ can be viewed as a perturbation of the true distribution P_θ under the total variation distance at the order of ϵ. In fact, one can check that any two distributions $P_{\theta_1}, P_{\theta_2}$ with total variation bounded by $\epsilon/(1-\epsilon)$ cannot be distinguished under Huber's contamination model, and thus a price of $L(\theta_1, \theta_2)$ has to be paid for estimating θ under a given loss $L(\cdot, \cdot)$. This insight is characterized by the notion of modulus of continuity Donoho and Liu (1991); Donoho (1994); see Section 5 of Chen et al. (2018) for more details.

2.3 Robust Inference via Multiplier Bootstrap

2.3.1 Bahadur Representation

We start the inference with presenting a non-asymptotic Bahadur representation for $\widehat{\boldsymbol{\theta}}_\tau$. To this end, we need to assume that the loss function $\ell(\ )$'s second derivative is Lipschitz, which is satisfied by a large number of smooth loss functions. Yet for Huber's loss, this is not the case and we instead assume an anti-concentration property for the distribution F.

Condition 2.3.1. For the loss function ℓ, we assume that either of the following condition holds. (i) ℓ'' is L-Lipschitz and bounded: $|\ell''(u)| \leq c_4$ for any $u \in \mathbb{R}$. (ii) The loss ℓ is Huber's loss. In this case, we assume that $\mathbb{P}_{\varepsilon\sim F}(a \leq |\varepsilon| \leq b) \geq C_F(b-a)$ for any $b > a > 0$ with some contant C_F.

Theorem 2.3.1. Assume Condition 2.2.1, 2.2.2 and 2.3.1 hold. Denote $c_F = \mathbb{E}_{\varepsilon\sim F}\ell_\tau''(|\varepsilon-\alpha_\tau|)$. Then for any $t \geq 1/2$, with probability at least $1-3e^{-t}$, we have

$$\left\| c_F \mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau) - \frac{1}{n}\sum_{i=1}^n \ell_\tau'(\varepsilon_i - \alpha_\tau)\boldsymbol{z}_i \right\|_2 \leq C_2\Big(\frac{d+t}{n} + \epsilon^2\Big) \tag{2.3.1}$$

as long as $\epsilon < c$ and $n \geq C(d+t)$, where $\boldsymbol{z}_i = \mathbf{S}^{-1/2}\bar{\boldsymbol{x}}_i$ and c, C, C_2 are constants independent of (d, t, n, ϵ).

The above Bahadur representation suggests that for the inference of a linear functional of $\boldsymbol{\beta}: \boldsymbol{\mu}^\intercal\boldsymbol{\beta}$ with some given $\boldsymbol{\mu} \in \mathbb{R}^d$, we may consider the following distribution approximation

$$\sqrt{n}\boldsymbol{\mu}^\intercal(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}) = \sqrt{n}\boldsymbol{\lambda}^\intercal(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i, \quad \text{with}$$
$$\boldsymbol{\lambda} = (0, \boldsymbol{\mu}^\intercal)^\intercal, \quad U_i := c_F^{-1}\ell_\tau'(\varepsilon_i - \alpha_\tau)\boldsymbol{\lambda}^\intercal\mathbf{S}^{-1/2}\boldsymbol{z}_i,$$

and utilize the asymptotic normality of $n^{-1/2}\sum_{i=1}^n U_i$. However, there are two problems with this strategy:

(i) (Contamination Bias) U_i is generally not mean zero due to contamination. This is because under the contamination distribution $G_{\boldsymbol{x}}$, ε_i may depend on $\boldsymbol{z}_i$. It thus requires the contamination proportion $\epsilon = o(1/\sqrt{n})$ to make the bias $\sqrt{n}\mathbb{E}U_i \to 0$.

(ii) (Variance Estimation) To estimate the variance of U_i, one needs consistent estimators for $c_F = \mathsf{E}_{\varepsilon \sim F}\ell'_\tau(\varepsilon - \alpha_\tau)$, $\mathsf{E}_{\varepsilon \sim F}[f\ell'_\tau(\varepsilon_i - \alpha_\tau)g^2$ and $\mathbf{S}^{-1}$. Consequently, it can be quite challenging to generate a trustworthy variance approximation in the presence of contamination and in a regime of increasing dimensionality.

In light of these concerns, we first propose a multiplier bootstrap procedure, which can implicitly estimate the variance of U_i and thus tends to be more impervious to contamination and more accurate with finite sample and growing dimensionality, compared to the approaches based on asymptotic normality. See Section 3.5.2 for comparisons (under distributed setting) in numeric experiments. To address the contamination bias issue, we develop a debiasing technique for confidence intervals, as described in Section 2.4. The rest of this section will be used to present the methodology and theoretical guarantee of the multiplier bootstrap procedure.

2.3.2 Methodology of Multiplier Bootstrap

Suppose that we generate $i.i.d.$ scalar random variables $w_1, \ldots, w_n$ that are independent of the observed data $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and satisfy $w_i \geq 0, \mathsf{E}(w_i) = 1$, $\mathrm{var}(w_i) = 1$. Consider $\{w_i\}_{i=1}^n$ as random weights to each sample observation $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and define the bootstrap loss and bootstrap estimator as

$$L^\flat_\tau(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^n w_i \ell_\tau(y_i - \bar{\boldsymbol{x}}_i^\intercal \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \mathsf{R}^{d+1}, \quad \text{and}$$

$$\widehat{\boldsymbol{\theta}}^\flat_\tau = (\widehat{\beta}^\flat_0, \widehat{\beta}^\flat_1, \ldots, \widehat{\beta}^\flat_d)^\intercal \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathsf{R}^{d+1}} L^\flat_\tau(\boldsymbol{\theta}),$$

respectively. Notice that conditional on the sample data $\mathcal{D}_n$, the expectation of the bootstrap loss $L^\flat_\tau(\boldsymbol{\theta})$ is the empirical loss $L_\tau(\boldsymbol{\theta})$. This implies that the robust estimator $\widehat{\boldsymbol{\theta}}_\tau$ in the $\mathcal{D}_n$-world is a target parameter in the bootstrap world, i.e.,

$$\widehat{\boldsymbol{\theta}}_\tau \in \operatorname*{argmin}_{\boldsymbol{\theta}} L_\tau(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathsf{E}[L^\flat_\tau(\boldsymbol{\theta}) | \mathcal{D}_n].$$

This motivates us to consider using the bootstrap estimation residual $\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau$ to mimic the empirical estimation residual $\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau$. Take the inference for β_j (for $j = 1, \ldots, d$) as an example. To estimate the q-quantile of $\widehat{\beta}_j - \beta_j$

$$c_j(q) = \inf\{t \in \mathbb{R} : \mathbb{P}(\widehat{\beta}_j - \beta_j \le t) \ge q\},$$

we consider the conditional q-quantile of $\widehat{\beta}_j^\flat - \widehat{\beta}_j$ given $\boldsymbol{D}_n$

$$c_j^\flat(q) = \inf\{t \in \mathbb{R} : \mathbb{P}(\widehat{\beta}_j^\flat - \widehat{\beta}_j \le t | \boldsymbol{D}_n) \ge q\}.$$

Therefore, a $(1 - \alpha)$ confidence interval for β_j can be given by

$$I_j^\flat = [\widehat{\beta}_j - c_j^\flat(1 - \alpha/2), \widehat{\beta}_j - c_j^\flat(\alpha/2)], \ j = 1, \ldots, d. \tag{2.3.2}$$

This type of multiplier bootstrap technique was studied in Spokoiny et al. (2015); Chen and Zhou (2020); Pan and Zhou (2021) with applications to ordinary least squares (OLS) regression, Huber's regression (for heavy-tailed data) and quantile regression, respectively. In this paper, we apply it to M-estimators for contaminated data and generalize it to a distributed context in a communication-efficient way, as demonstrated in Section 3.3.

2.3.3 Theoretical Results for Bootstrap Inference

In this section, we first derive the estimation error bound and non-asymptotic Bahadur representation for the bootstrap estimator $\widehat{\boldsymbol{\theta}}_\tau^\flat$. Built on that, Theorem 2.3.3 and 2.3.4 provide the theoretical guarantee of our bootstrap procedure.

Condition 2.3.2. $w_1, \ldots, w_n$ are i.i.d random varibles satisfying that $w_i \ge 0, \mathbb{E}(w_i) = 1$, $\mathrm{var}(w_i) = 1$. Let $e_i = w_i - 1$. Assume that e_i is sub-Gaussian with parameter ν_e.

Theorem 2.3.2. Denote $\mathbb{P}^\flat(\cdot) := \mathbb{P}(\cdot | \boldsymbol{D}_n)$. Under Conditions 2.2.1-2.3.2, for any $t \ge 1/2$, the bootstrap estimator $\widehat{\boldsymbol{\theta}}_\tau^\flat$ satisfies that

(i) with probability (over $\boldsymbol{D}_n$) at least $1 - 4e^{-t}$,

$$\mathbb{P}^\flat\left(\|\widehat{\boldsymbol{\theta}}_\tau^\flat - \boldsymbol{\theta}_\tau\|_{\mathbf{S}} \le C_1^\flat\left(\sqrt{\frac{d+t}{n}} + \epsilon\right)\right) \ge 1 - 2e^{-t}$$

(ii) with probability (over D_n) at least $1 - 8e^{-t}$,

$$P\left(\left\|c_F \mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau) - \frac{1}{n}\sum_{i=1}^n \ell_\tau'(\varepsilon_i^\flat - \alpha_\tau) e_i \boldsymbol{z}_i\right\|_2 \le C_2^\flat\left(\frac{d+t}{n} + \epsilon^2\right)\right) \ge 1 - 8e^{-t} \tag{2.3.3}$$

as long as $\epsilon < c$ and $n \ge C'(d+t)^2$, where $C_1^\flat, C_2^\flat, c, C' > 0$ are constants independent of (d, t, n, ϵ) .

By comparing the Bahadur representations of the robust estimator (2.3.1) and the bootstrap estimator (2.3.3), we have the following approximation results:

$$\boldsymbol{\mu}'(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*) = \boldsymbol{\lambda}'(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau^*) \approx S_n := \frac{1}{n}\sum_{i=1}^n U_i \approx N(\mathsf{E}U_1, \frac{1}{n}\operatorname{Var} U_1),$$

$$\boldsymbol{\mu}'(\widehat{\boldsymbol{\beta}}_\tau^\flat - \widehat{\boldsymbol{\beta}}_\tau) = \boldsymbol{\lambda}'(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau) \approx S_n^\flat := \frac{1}{n}\sum_{i=1}^n e_i U_i \overset{|D_n}{\approx} N(0, \frac{1}{n^2}\sum_{i=1}^n U_i^2),$$

When the bias $\mathsf{E}U_1$ is negligible to its variance $n^{-1}\operatorname{Var} U_1$ i.e., $\epsilon = o(1/\sqrt{n})$, the two normal distributions in the above formulas are close to each other, and thus validates the distribution approximation of the bootstrap estimator. The following theorem justifies this intuition in the form of the Kolmogorov distance.

Theorem 2.3.3. Under Conditions 2.2.1-2.3.2, for any $\boldsymbol{\mu} \in \mathsf{R}^d$, and any $t \le 1/2$, we have

$$\sup_{x \in \mathsf{R}} \left|P\left(\boldsymbol{\mu}'(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*) \le x\right) - P\left(\boldsymbol{\mu}'(\widehat{\boldsymbol{\beta}}_\tau^\flat - \widehat{\boldsymbol{\beta}}_\tau) \le x\right)\right| \le C\left(\frac{d+t}{n} + \sqrt{n}\epsilon\right) + 11e^{-t}$$

with probability (over D_n) at least $1 - 14e^{-t}$, respectively, as long as $\epsilon < c$ and $n \ge C'(d+t)^2$, for some constants $c, C, C' > 0$ independent of $(d, t, n, \epsilon, \boldsymbol{\lambda}, x)$.

Theorem 2.3.4 (Validity of bootstrap confidence intervals)**.** Under the same conditions of Theorem 2.3.3, for any $j = 1, \ldots, d$, we have

$$\sup_{q \in (0,1)} \left|P\left(\widehat{\beta}_j - \beta_j^* \le c_j^\flat(q)\right) - q\right| \le C\left(\frac{d+t}{n} + \sqrt{n}\epsilon\right) + 28e^{-t},$$

In particular, if $\epsilon = o(1/\sqrt{n})$, and $d = o(\sqrt{n})$, then we establish the validity of the bootstrap confidence intervals for β_j^* in (2.3.2):

$$\sup_{\alpha \in (0,1)} \left|P\left(\beta_j^* \in I_j^\flat\right) - (1 - \alpha)\right| = o(1)$$

as $n \to \infty$.

2.4 Debiased Confidence Intervals

Theorem 2.3.4 indicates that the bootstrap confidence intervals (2.3.2) are asymptotically valid only when the contamination proportion is small, compared to the order $1/\sqrt{n}$. This originates from the fact that a larger contamination proportion would make the bias of our estimator exceeds its variance and thus can no longer be covered by the bootstrap confidence intervals like (2.3.2). This motivates us to consider adding a debiasing term to the confidence intervals to boost its coverage with larger contamination proportions.

The Bahadur representation (3.2.4) shows that the bias of our distributed estimator $\boldsymbol{\mu}^\intercal \widehat{\boldsymbol{\beta}}_\tau = \boldsymbol{\lambda}^\intercal \widehat{\boldsymbol{\theta}}_\tau$ can be characterized by $\|U_1\|$ with $U_1 = c_F^{-1}\ell_\tau'(\varepsilon_1 - \alpha_\tau)\boldsymbol{\lambda}^\intercal \mathbf{S}^{-1/2}\boldsymbol{z}_1$. It is easy to check that $\|U_1\| \le \epsilon c_1 \tau c_F^{-1}\|\boldsymbol{\lambda}^\intercal \mathbf{S}^{-1/2}\|_2$. Therefore, we consider estimating c_F and $\mathbf{S}$ to obtain an estimate of the upper bound of $\|U_1\|$. Let

$$\widehat{c}_F = \frac{1}{n}\sum_{i=1}^n \ell_\tau''(\widehat{\varepsilon}_i), \quad \widehat{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^n \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\intercal, \quad \widetilde{D}_j := \epsilon c_1 \tau \left\| \frac{1}{\widehat{c}_F} \boldsymbol{\lambda}^\intercal \widehat{\mathbf{S}}^{-1/2} \right\|_2$$

where $\widehat{\varepsilon}_i = y_i - \bar{\boldsymbol{x}}_i^\intercal \widehat{\boldsymbol{\theta}}_\tau$ is the residual for our M-estimator $\widehat{\boldsymbol{\theta}}_\tau$, and $\boldsymbol{\lambda} = (0, \ldots, 0, 1, 0, \ldots, 0)$ is a vector in $\mathbb{R}^{d+1}$ with only the $(j+1)$-th component being 1, for $1 \le j \le d$. Then $\widetilde{D}_j$ serves as an estimate of an upper bound of the contamination bias for estimating the j-th coefficient β_j. This implies that we may define a debiased bootstrap confidence interval for β_j as following:

$$\widehat{I}_{D,j}^\flat = \left[\widehat{\beta}_j - C_D\left(c_j^\flat(1-\alpha/2) + \widetilde{D}_j\right), \widehat{\beta}_j - C_D\left(c_j^\flat(\alpha/2) - \widetilde{D}_j\right)\right], \tag{2.4.1}$$

where $C_D \ge 1$ is an universal constant to be specified.

Theorem 2.4.1. Assume Conditions 2.2.1-2.3.2 hold. Further assume that the largest and smallest eigenvalue of $\mathbf{S}$ are both bounded: $c_{\mathbf{S}} \le \lambda_{\min}(\mathbf{S}) \le \lambda_{\max}(\mathbf{S}) \le C_{\mathbf{S}}$ for some constants $C_{\mathbf{S}}, c_{\mathbf{S}}$. Then for any $j = 1, \ldots, d$, any $\alpha \in (0,1)$, any $C_D > 1$, we have

$$\mathbb{P}\left(\beta_j \in \widehat{I}_{D,j}^\flat\right) \ge (1-\alpha) - C\left(\sqrt{\frac{\log n}{n}} + \epsilon^2\right)$$

as long as $\epsilon \le c$ and $n \ge C' d^2$ for some large enough C' and small enough c. Here, c, C, C' are independent of (n, d, ϵ).

Remark 2.4.1. (i) The above theorem shows that the debiased bootstrap confidence intervals (3.4.1) can achieve the desired coverage probability asymptotically when $\epsilon = o(1)$ and $d = o(\sqrt{p}/\sqrt{n})$, which relaxes the constraint on the contamination proportion, compared with the confidence intervals without a debiasing step.

(ii) Although we construct the debiased bootstrap confidence intervals in a conservative way by including an estimate of the upper bound of the bias, this procedure does not sacrifice too much efficiency in the sense that the length of the resulting confidence interval is of order $O(1/\sqrt{p}/\sqrt{n}+\epsilon)$, which matches the centralized minimax rate of convergence, as shown by Theorem 2.2.2. (Although Theorem 2.2.2 only presents a minimax lower bound for the norm $\|\cdot\|_S$, one can prove a lower bound $O(1/\sqrt{p}/\sqrt{n} + \epsilon)$ for the sup-norm following the same proof.)

(iii) Our numerical experiments show that selecting $C_D = 1$ in (2.4.1) is sufficient to lead to satisfactory coverage proportion.

As Theorem 2.4.1 can be viewed as a special case of Theorem 3.4.1 with $m = 1$, we only provide a proof for Theorem 3.4.1 in Appendix B and omit the proof for Theorem 2.4.1.

2.5 Numerical Studies

2.5.1 Robust Estimation

In this section, we investigate the numerical performance of the proposed M-estimators with contrast to the following robust estimation methods: (i) ordinary least squares (OLS); (ii) least median squares (LMedian) (Rousseeuw, 1984); (iii) TORRENT (Bhatia et al., 2015); (iv) RANSAC (Fischler and Bolles, 1981). TORRENT is an iterative hard-thresholding algorithm similar to the least trimmed square (Rousseeuw, 1984; Rousseeuw and Van Driessen, 2006). As discussed in Bhatia et al. (2015) and Prasad et al. (2020), TORRENT is shown to outperform various regularized ℓ_1 algorithms for robust regression and thus we do not consider these methods in this comparison. RANSAC is a family of algorithms widely used in the image processing field, which use random sampling of the data to separate the inliers

from the outliers.

For our proposed M-estimator, we employ Huber's loss (2.2.3) and select the tuning parameter τ via the following heuristic two-step way: (i) initially set $\tau = 1$ and obtain the residuals using the available data; (ii) calculate the median absolute deviation (MAD) of the residuals and use it as the final value of τ. Ideally, this results in τ having the same scale as the variance of the noise ε under distribution F. For TORRENT, we set the thresholding parameter to be the contamination proportion ϵ plus 0.05 (as we find this is more robust and performs better than using just ϵ). We also try setting it 50% for a conservative usage of TORRENT when we don't know the oracle ϵ. These two are denoted by TORRENT* and TORRENT50 (in the figures below). For RANSAC, we set the minimal sample size to be 5, and the maximum distance for residuals to be the 95% quantile of F distribution. For least median squares (LMedian), we use a Monte Carlo type technique to draw 1000 random subsamples of 5 different points to seek an approximate solution (Rousseeuw and Leroy, 2005).

To generate contaminated data $(\boldsymbol{x}_i, y_i)$ following model (2.1.1), we consider $\beta_0 = 0$ and $\boldsymbol{\beta} = (5, 2, 1, \quad 1, \quad 3)^\intercal$ ($d = 5$), and simulate $\boldsymbol{x}$ from standard multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ ($\boldsymbol{I}_d$ is a $d \quad d$ identity matrix). Regarding the distribution of the noise ε, we consider the following three heavy-tailed distributions for F: (F1) t-distribution with 1.5 degrees of freedom, denoted as $t(1.5)$; (F2) Lognormal distribution with $\mu = 0$ and $\sigma = 1$, i.e., the logarithm of the standard normal distribution; (F3) Pareto distribution with shape parameter 1.5 and scale parameter 1. We investigate the following four types of contamination distribution G: (G1) $Unif(\ \ j y_0 j, j y_0 j)$, where $y_0 = \beta_0 + \boldsymbol{x}^\intercal \boldsymbol{\beta} \ + \varepsilon$ with $\varepsilon \quad F$, which represents the response variable without contamination; (G2) $sign(y_0) \quad F$; (G3.1) G is a point mass at x_1, where x_1 is the first coordinate of $\boldsymbol{x}$; (G3.2) G is a point mass at $100x_1$. The first two contamination distributions represent contamination based on the initial response without contamination, and the last two are about contamination based on the covariates $\boldsymbol{x}$. To ensure a fair comparison for the distributed M-estimator without intercept, all of the aforementioned distributions for F and G are centered.

We fix the sample size $n = 250$ and explore the contamination proportion ϵ in the range of $[0, 25\%]$. We report the estimation error of the coefficient $\boldsymbol{\beta}$ under the ℓ_2-norm,

i.e., $k\widehat{\boldsymbol{\beta}} \quad \boldsymbol{\beta}\, k_2$. Figure 1 draws the estimation error versus the contamination proportion ϵ for the following combinations of F and G: (F1,G1), (F2,G2), (F3,G3.1) and (F3,G3.2), averaged over 100 random trials. We can see that the M-estimators (denoted as Huber) consistently have the minimal errors across almost all the settings, except for the case where the contamination proportion is very large (over 20%) under the distributions (F3,G3.2). For that case, TORRENT* performs better. This is not surprising as in this setting, the contamination distribution is $100x_1$, making the outliers very easy to be identified and deleted from the estimation procedures of TORRENT, while Huber includes all the data in the estimation. Based on the empirical findings, we may conclude that when the contamination proportion is not very large or the contamination distribution G is not so distinguishable from the original noise distribution F, M-estimators typically lead to higher estimation accuracy than the methods based on detecting and excluding outliers (like TORRENT).

2.5.2 Robust Inference

This section presents the empirical performance of inference procedures with contaminated data. We compare the following methods: (i) OLS-norm: standard asymptotic normal-based confidence intervals; (ii) OLS-boot: the multiplier bootstrap confidence intervals based on OLS estimators; (iii) M-boot: the multiplier bootstrap confidence intervals based on M-estimators (Algorithm 1). For the method (iii), we consider five robust loss functions: Huber loss, two smooth-Huber losses and two pseudo-Huber losses, as presented in Appendix A.1. We use the same way to select tuning parameter τ as in Section 2.5.1.

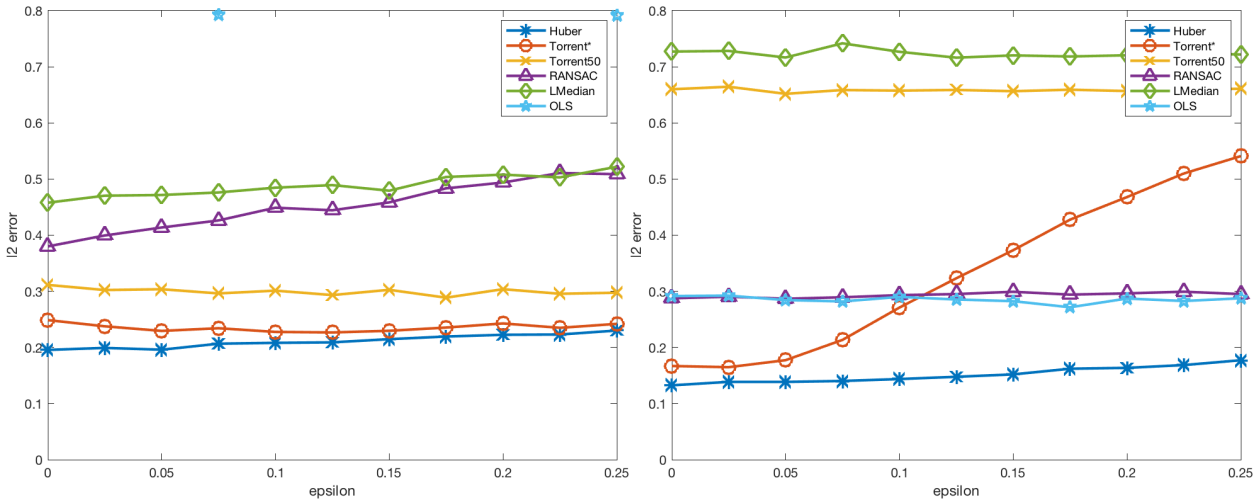
Following the data generation process in Section 2.5.1, we set $\boldsymbol{\beta} = (5, 2, 1, \quad 1, \quad 3)^{\mid}$ and $\beta_0 = 0$ with $d = 5$ and $n = 250$, and simulate $\boldsymbol{x} \qquad \boldsymbol{N}(\mathbf{0}, \boldsymbol{I}_d)$. To this end, we fix the contamination proportion $\epsilon = 0.1/\sqrt{\boldsymbol{p}}\overline{n}$ to satisfy the condition that $\epsilon = o(1/\sqrt{\boldsymbol{p}}\overline{n})$, as required by Theorem 2.3.4. We choose the contamination distribution as a point mass at the first coordinate of $\boldsymbol{x}$, i.e. $G_{\boldsymbol{x}} = 100x_1$. Under this contamination distribution, the data would act as though the value of the first component of $\boldsymbol{\beta}$ (i.e. β_1) being boosted by 100 times. Therefore, we focus on constructing a confidence interval for β_1. Regarding the distribution F, we consider the following various distributions: (1) Standard normal $\boldsymbol{N}(0, 1)$: $t)(1.5)$;(2)

t-distribution with degrees of freedom 1.5; (3) Pareto distribution with scale parameter 1 and shape parameter 1.5; (4) Lognormal distribution-a logarithm of a standard normal: Lognormal$(0, 1)$; (5) Gamma distribution with shape parameter 5 and scale parameter 1; (6) Weibull distribution with scale parameter 1 and shape parameter 1; (7) Pareto-t mixture: 0.5Pareto$(1, 1.5)$+0.5$t(1.5)$; (8) Normal-Lognormal mixture: $0.5\boldsymbol{N}(0, 1)$+0.5Lognormal$(0, 1)$.

Table 2 and 3 report the coverage and width of the 95% bootstrap confidence intervals for β_1, based on 500 random trials. From these tables, it can be seen that compared to OLS-based methods, our robust methods generally achieve the nominal confidence level 95% under the various heavy-tailed distributions F, with significantly smaller widths at the same time. This indicates our proposed inference procedures enjoy the robustness and effectiveness simultaneously.

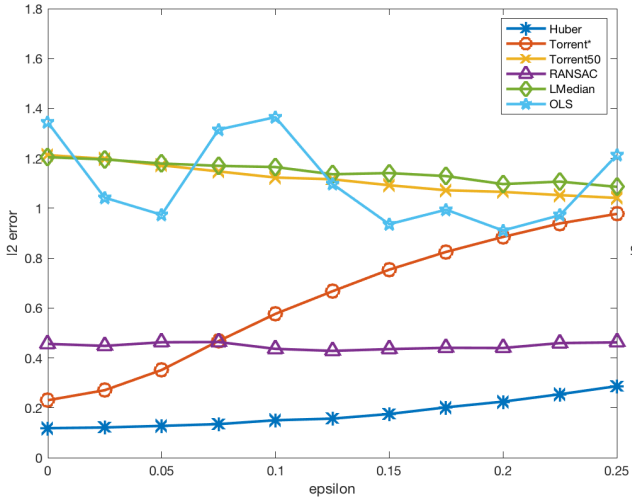
We then consider increase the contamination proportion to $\epsilon = 1/\sqrt{n}$ and evaluate the performance of the debiased confidence intervals (2.4.1) (with $C_D = 1$). For simplicity, we only consider Huber's loss for our M-estimator. (As we notice that the performances of pseduo-Huber loss and smooth-Huber loss are similar to Huber loss, as shown in the previous section.)

Table 4 and 5 show the coverage and mean widths of the confidence intervals based on 500 random trials. In each table, we use OLS-norm and OLS-boot as baselines. Huber and Debiased-Huber represent the bootstrap confidence intervals (2.3.2) and the debiased bootstrap confidence intervals (2.4.1), respectively, with both based on Huber's estimator. From the coverage table, we can see that when contamination proportion ϵ increases to the scale $1/\sqrt{n}$, the coverage proportions of the bootstrap confidence intervals drop below 95%, while the debiased the debiased bootstrap confidence intervals constantly have a coverage probability higher than 95%. From the width table, we observe that the widths of the debiased confidence intervals are roughly twice as large as those wihout debiasing. This is consistent with our theory that the width of the bootstrap confidence interval is $\boldsymbol{O}(1/\sqrt{n})$ and the width of the debiased the bootstrap confidence interval is $\boldsymbol{O}(1/\sqrt{n} + \epsilon)$ (optimal width).

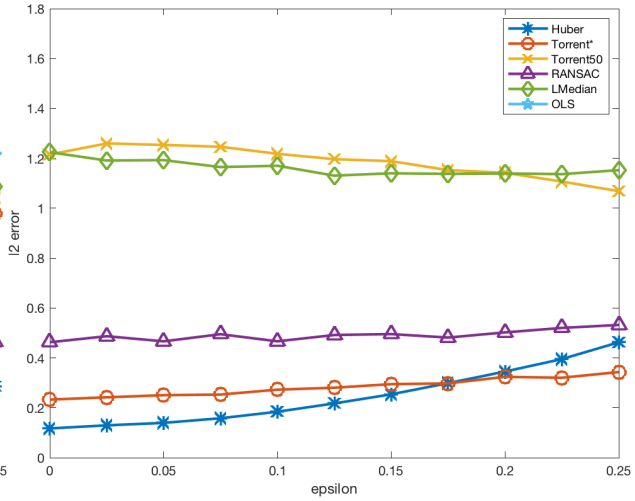


(a) (F1,G1)

(b) (F2,G2)



(c) (F3,G3.1)



(d) (F3,G3.2)

Figure 1: ℓ_2 estimation error versus the contamination proportion ϵ under various settings for distribution F and G.

Table 2: Coverage of the 95% confidence intervals for β_1 under various settings for distribution F when $\epsilon = 0.1/\sqrt{p}\overline{n}$. The results are based on 500 random trials.

	OLS-norm	OLS-boot	Huber	PseHub I	PseHub II	SmthHub I	SmthHub II
norm	0.884	0.978	0.95	0.954	0.956	0.952	0.96
t	0.88	0.97	0.946	0.942	0.954	0.954	0.946
pareto	0.932	0.988	0.99	0.98	0.984	0.974	0.986
lognormal	0.854	0.972	0.946	0.924	0.93	0.934	0.938
Gamma	0.852	0.972	0.952	0.946	0.95	0.948	0.938
Weibull	0.848	0.972	0.958	0.952	0.952	0.954	0.956
Pareto-t	0.906	0.976	0.966	0.968	0.972	0.966	0.964
norm-lognormal	0.868	0.98	0.952	0.958	0.962	0.954	0.96

Table 3: Mean widths of the 95% confidence intervals for β_1 under various settings for distribution F when $\epsilon = 0.1/\sqrt{p}\overline{n}$. The results are based on 500 random trials.

	OLS-norm	OLS-boot	Huber	PseHub I	PseHub II	SmthHub I	SmthHub II
norm	1.508	1.513	0.2778	0.2738	0.274	0.2894	0.2761
t	2.708	2.618	0.3924	0.4033	0.4011	0.3821	0.394
pareto	62.59	42.74	0.6497	0.7582	0.7213	0.6758	0.6716
lognormal	1.74	1.88	0.2687	0.2901	0.283	0.2809	0.2731
Gamma	1.67	1.848	0.5845	0.5835	0.5787	0.6123	0.5818
Weibull	1.585	1.634	0.2034	0.2157	0.2122	0.2182	0.2051
Pareto-t	26.67	17.55	0.4967	0.5339	0.5224	0.473	0.5008
norm-lognormal	1.632	1.694	0.295	0.2971	0.2961	0.2854	0.2945

Table 4: Coverage of 95% confidence intervals for β_1 under various settings for distribution F when $\epsilon = 1/\sqrt{n}$. The results are based on 500 random trials.

	OLS-norm	OLS-boot	Huber	Debiased-Huber
norm	0.008	0.068	0.846	0.994
t	0.048	0.106	0.888	0.998
pareto	0.506	0.542	0.95	0.998
lognormal	0.014	0.048	0.906	0.996
Gamma	0.02	0.086	0.856	0.992
Weibull	0.018	0.048	0.876	0.994
Pareto-t	0.292	0.372	0.868	0.996
norm-lognormal	0.016	0.092	0.884	0.996

Table 5: Mean widths of the 95% bootstrap confidence intervals for β_1 under various settings for distribution F when $\epsilon = 1/\sqrt{n}$. The results are based on 500 random trials.

	OLS-norm	OLS-boot	Huber	Debiased-Huber
norm	5.97	9.44	0.308	0.503
t	6.31	9.54	0.444	0.719
pareto	292	105	0.746	1.14
lognormal	5.97	9.3	0.299	0.476
Gamma	5.97	9.31	0.643	1.05
Weibull	5.92	9.33	0.225	0.362
Pareto-t	22.3	19.3	0.573	0.913
norm-lognormal	5.9	9.1	0.33	0.537

3.0 Distributed Robust Regression Inference under Huber's Contamination Model

3.1 Introduction

With the advent of modern data collection techniques, statistical estimation and inference now face unprecedented challenges. The first challenge comes with data storage and communication. In many applications like web-search (Corbett et al., 2013), data sets are often too vast to be stored on a single machine, and thus have to be be split among numerous machines. However, machine to machine communication can be costly, time-consuming, or power-intensive (Fuller and Millett, 2011). Additionally, data collection from multiple sites is common in fields like medicine (Sidransky et al., 2009; Cheng et al., 2017), where direct data exchange is constrained due to privacy protection for individual-level information. For such distributed data sets with communication constraints, various communication-efficient distributed algorithms are developed for a wide range of problems in statistics (Zhang et al., 2013; Battey et al., 2018; Lee et al., 2017; Wang et al., 2017; Jordan et al., 2019; Shi et al., 2018; Huang and Huo, 2019; Kannan et al., 2014; Mackey et al., 2015; Wang and Dunson, 2013; Zhang et al., 2015; Zhao et al., 2016; Rosenblatt and Nadler, 2016) and optimization Boyd et al. (2011); Duchi et al. (2011); Zhang and Xiao (2018); Shamir et al. (2014). However, very few of them consider the presence of contamination in distributed data. In fact, data sets that are collected from different sources are more prone to be exposed to (unknown) contamination. This raises a new challenge and motivates us to develop a robust communication-efficient method with distributed contaminated data.

To be more specific, we consider a regression setting and assume that the overall $i.i.d.$ observations $\mathsf{f}(\boldsymbol{x}_i, y_i)\mathsf{g}_{i=1}^N$ (that follows the linear model with contamination (2.1.1)) are distributed across m machines, and each machine has access to only a subsample with size n. Under this framework, the main goal of this chapter is two-fold:

(i) **Distributed robust estimation**: We aim to propose a communication-efficient robust estimator of $\boldsymbol{\beta}$ that is able to achieve the optimal statistical error rate $\mathsf{O}(\sqrt{d/N} + \epsilon)$ as

if one has access to the entire N sample data.

(ii) **Distributed inference**: We want to build valid confidence intervals for linear functionals of $\boldsymbol{\beta}$ with the optimal width $O(1/\sqrt{N} + \epsilon)$ and the least possible communication requirements.

In the growing body of work on distributed estimation, two main approaches have been discussed. The first one is one-shot averaging (Zhang et al., 2013; Chen and Xie, 2014; Zhang et al., 2015; Lee et al., 2017; Battey et al., 2018). It requires only one round of communication to average the estimators from each local machine and, thus, is communication-efficient. However, this approach mainly fits the setting where the number of machines is small and the local sample size is large. Specifically, it requires $m = o(\sqrt{N/d})$ (or equivalently $n \quad \sqrt{Nd}$) to achieve the optimal convergence rate. To remove this restriction, Shamir et al. (2014); Wang et al. (2017); Jordan et al. (2019) proposed a multi-round communication approach as an alternative. They showed that the resulting estimator can achieve the optimal rate after $O(\log m / \log n)$ rounds of communication. However, all these works lack theoretical guarantee and empirical validation in the presence of heavy-tailed errors and outlier contamination. Our paper closes this gap by extending their approach with robust M-estimation. We extend the M-estimation framework (developed in Chapter 2) with the multi-round communication approach in Jordan et al. (2019) to get a robust communication-efficient estimator with the theoretical guarantee for its statistical optimality.

With the distributed estimator, we now consider the distributed inference problem. Most distributed inference methods are based on the asymptotic normality of the distributed estimator Jordan et al. (2019); Huang and Huo (2019). While it is computationally fast, in order to estimate the estimator's variance, it typically needs a round of communication of $O(md^2)$ bits for an precise estimate of the precision matrix of $\boldsymbol{x}$. This raises a significantly higher cost than that in the distributed estimation ($O(md)$) when the dimensionality d is large. An alternative way is to estimate the variance using only the local sample. Though it is communication-efficient, it is often challenging to find a robust and accurate estimate with a (limited) local sample size, growing dimensionality, and the presence of contamination.

An alternative approach without the need for an explicit estimate of variance is resampling methods like bootstrap. Note that bootstrap, though widely used in the traditional

centralized setting, is much less considered in the distributed environment since a naive application generally requires hundreds of rounds of information communication. While recent studies (Kleiner et al., 2014; Sengupta et al., 2016; Yu et al., 2020) have attempted various techniques to adapt bootstrap to distributed data, these works impose more or less scaling restrictions on the number of machines, and their validity under contamination is dubious. This encourages us to develop a new distributed inference procedure that is communication-efficient, robust to contamination, and works without stringent restrictions on the number of machines.

Inspired by the multiplier bootstrap procedure developed in Chapter 2, we propose a distributed multiplier bootstrap procedure to construct confidence intervals for linear functionals of $\boldsymbol{\beta}$ based on the robust estimator. This bootstrap procedure can be implemented on a local machine (e.g. the master machine) and thus does not require further communication between machines after the distributed estimation. While it only bootstraps the local sample data, it incorporates the global information aggregated in the distributed estimation phase. As a result, our method can generate confidence intervals of optimal oracle width $O(1/\sqrt{N} + \epsilon)$ without any further communication cost. We establish the theoretical validity of our distributed bootstrap method when the local sample size is not too small ($n \gtrsim d^2$) and the contamination proportion ϵ is not too large ($\epsilon = o(1/\sqrt{N})$) (with no restriction on m). For cases with a larger contamination proportion, we propose debiased confidence intervals, which can cover the potential (severe) bias caused by contamination while not losing much statistical efficiency, as validated by our theory and experiments. For cases with smaller local sample sizes, we conjecture that a modified version of our distributed bootstrap algorithm may work; see the discussion section for further details.

The remainder of this chapter is organized as follows. Section 3.2 and Section 3.3 present our distributed M-estimator and multiplier bootstrap procedure, respectively, together with their theoretical guarantees. In Section 3.4, we extend the debiased confidence interval method (developed in Chapter 2) to the distributed setting, for cases with a large contamination proportion. Section 3.5 reports our experimental results for both estimation and confidence interval coverage and width. We conclude with a discussion in Section 3.6. Appendix B contains all the proofs.

Notations. We use the same notations as Chapter 2.

3.2 Distributed M-estimator

We now consider a distributed setting where the overall N observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ (*i.i.d.* from the contaminated linear model (2.1.1)) are randomly distributed across m machines, which can communicate with a master machine. Assume that $N = mn$ and each machine has access to a subsample of n *i.i.d.* observations. For $j = 1, \ldots, m$, we use $\{(\boldsymbol{x}_i, y_i)\}_{i \in \mathcal{I}_j}$ to denote the subsample data stored in the j-th machine, where $\mathcal{I}_j$'s are disjoint index sets that satisfy $\bigcup_{j=1}^m \mathcal{I}_j = \{1, \ldots, N\}$ and $|\mathcal{I}_j| = n$.

Faced with such distributed contaminated data set, we propose a distributed robust regression algorithm, by applying the iterative communication-efficient surrogate likelihood (CSL) approach in Jordan et al. (2019) and Wang et al. (2017) to our robust M-estimation framework, developed in Section 2.2. The idea of the CSL approach is that one can approximate the higher order derivatives of the global loss by using those of the local loss, while only the first derivative of the global loss is precisely estimated by aggregating the gradient of each local loss. Formally, for some given initial estimator $\bar{\boldsymbol{\theta}}$, we define the surrogate loss function

$$\widetilde{L}_\tau(\boldsymbol{\theta}) := L_{1,\tau}(\boldsymbol{\theta}) - \left\langle \nabla L_{1,\tau}(\bar{\boldsymbol{\theta}}) - \nabla L_{N,\tau}(\bar{\boldsymbol{\theta}}), \boldsymbol{\theta} \right\rangle, \tag{3.2.1}$$

where

$$L_{N,\tau}(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^{N} \ell_\tau(y_i - \bar{\boldsymbol{x}}_i^\intercal \boldsymbol{\theta}) \quad \text{and} \quad L_{j,\tau}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i \in \mathcal{I}_j} \ell_\tau(y_i - \bar{\boldsymbol{x}}_i^\intercal \boldsymbol{\theta}), \quad j = 1, \ldots, m,$$

are the global and local loss functions, respectively. The resulting estimator is given by

$$\widetilde{\boldsymbol{\theta}}_\tau = (\widetilde{\beta}_{0,\tau}, \widetilde{\boldsymbol{\beta}}_\tau^\intercal)^\intercal \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \widetilde{L}_\tau(\boldsymbol{\theta}). \tag{3.2.2}$$

To calculate the surrogate loss, we need to communicate each local gradient to the master machine to obtain $\nabla L_{N,\tau}(\bar{\boldsymbol{\theta}})$, which takes $O(md)$ bits, a much lower communication cost than the raw data transmission ($O(Nd)$ bits). In the following theorem, we show that compared to the initial estimator $\bar{\boldsymbol{\theta}}$, the estimation error of $\widetilde{\boldsymbol{\theta}}_\tau$ is improved by a factor of $O(\sqrt{d/n} + \epsilon)$.

Theorem 3.2.1. For $r_0, r > 0$, define the events

$$E_0(r_0) = \{\|\bar{\theta} - \theta_\tau\|_S \le r_0\} \quad \text{and} \quad E(r) = \{\|S^{1/2} r L_{N,\tau}(\theta_\tau)\|_2 \le r\}.$$

Under Condition 2.2.1 and 2.2.2, for any $t > 1/2$, suppose $r \ge r_0 \ge \sqrt{(d+t)/n} + \epsilon$, then on the event $E_0(r_0) \setminus E(r)$, the estimator $\widetilde{\theta}_\tau$ given in (3.2.2) satisfies

$$\|\widetilde{\theta}_\tau - \theta_\tau\|_S \le C\left[\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r_0 + r\right] \quad \text{and}$$

$$\left\| c_F S^{1/2}(\widetilde{\theta}_\tau - \theta_\tau) - \frac{1}{N}\sum_{i=1}^N \ell'_\tau(\varepsilon_i - \alpha_\tau) z_i \right\|_2 \le C\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r_0$$

with probability at least $1 - 3e^{-t}$, as long as $\epsilon < c$ and $n \ge C'(d+t)$, where c, C, C' are constants independent of (N, n, d, ϵ, t).

This theorem reveals that, given an initial estimator $\bar{\theta}$ with statistical error $O(\sqrt{d/n}+\epsilon)$, applying the above procedure repeatedly would eventually lead to an estimator with the error rate r, which corresponds to the centralized optimal rate $\sqrt{d/N} + \epsilon$. Specifically, denote $\widetilde{\theta}_\tau^{(0)} = \bar{\theta}$ to be our initial estimator. Then at iteration $k = 1, 2, \ldots$, we may construct the surrogate loss function using the estimator from the previous step as

$$\widetilde{L}_\tau^{(k)}(\theta) := L_{1,\tau}(\theta) - \left\langle r L_{1,\tau}(\widetilde{\theta}_\tau^{(k-1)}) - r L_{N,\tau}(\widetilde{\theta}_\tau^{(k-1)}), \theta \right\rangle,$$

with the resulting estimator $\widetilde{\theta}_\tau^{(k)}$ defined as

$$\widetilde{\theta}_\tau^{(k)} = (\widetilde{\beta}_{0,\tau}, \widetilde{\beta}_\tau^\intercal)^\intercal \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{d+1}} \widetilde{L}_\tau^{(k)}(\theta).$$

Theorem 3.2.2. Under the same conditions of Theorem 3.2.1, on the event $E_0(r_0) \setminus E(r)$, the T-th iterate estimator $\widetilde{\theta}_\tau^{(T)}$ satisfies

$$\|\widetilde{\theta}_\tau^{(T)} - \theta_\tau\|_S \le r, \quad \text{and}$$

$$\left\| c_F S^{1/2}(\widetilde{\theta}_\tau^{(T)} - \theta_\tau) - \frac{1}{N}\sum_{i=1}^N \ell'_\tau(\varepsilon_i - \alpha_\tau) z_i \right\|_2 \le \left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r$$

with probability at least $1 - 3Te^{-t}$, for $T \ge \log(r/r_0)/\log(C(\sqrt{(d+t)/n} + \epsilon)) + 1$, where C is the same constant as in Theorem 3.2.1.

Corollary 3.2.1. Let the initial estimator be a minimizer of the local loss $L_{1,\tau}(\boldsymbol{\theta})$, i.e. $\widetilde{\boldsymbol{\theta}}^{(0)} = \bar{\boldsymbol{\theta}} \in \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} L_{1,\tau}(\boldsymbol{\theta})$. Then under Condition 2.2.1 and 2.2.2, with the above iterative procedures, the T-th iterate estimator $\widetilde{\boldsymbol{\theta}}_\tau^{(T)}$ satisfies

$$\|\widetilde{\boldsymbol{\theta}}_\tau^{(T)} - \boldsymbol{\theta}_\tau\|_2 \cdot \sqrt{\frac{d+t}{N}} + \epsilon, \quad \text{and} \tag{3.2.3}$$

$$\left\| c_F \mathbf{S}^{1/2} (\widetilde{\boldsymbol{\theta}}_\tau^{(T)} - \boldsymbol{\theta}_\tau) - \frac{1}{N} \sum_{i=1}^{N} \ell'_\tau(\varepsilon_i - \alpha_\tau) \boldsymbol{z}_i \right\|_2 \cdot \left(\sqrt{\frac{d+t}{n}} + \epsilon \right) \left(\sqrt{\frac{d+t}{N}} + \epsilon \right) \tag{3.2.4}$$

with probability at least $1 - 3(T+1)e^{-t}$, for any $t > 1/2$, as long as $T \gtrsim \log m / \log(n/(d+t)) \asymp \log m / \log(1/\epsilon)$, $\epsilon < c$ and $n \geq C'(d+t)$, where c, C, C' are constants independent of (N, n, d, ϵ, t).

The above theoretical results indicate that with the local M-estimator as the initializer, after at most $O(\log m)$ rounds of communication, we can obtain an estimator that achieves the centralized optimal convergence rate $O(\sqrt{d/N} + \epsilon)$ for the estimation of $\boldsymbol{\beta}$. There are other options for the initializer. One can choose the average of all the local M-estimators, for instance, which is supposed to result in a faster convergence at the expense of one extra round of $O(md)$ bit communication.

3.3 Distributed Multiplier Bootstrap

In this section, we consider adapting the multiplier bootstrap procedure (presented in Section 2.3) to the distributed setting. Assume that we have obtained an estimator achieving the optimal rate $O(\sqrt{d/N} + \epsilon)$, say $\widetilde{\boldsymbol{\theta}}_\tau^{(T)}$ (for $T \gtrsim \log m$) as described in the previous section. Recall that

$$\widetilde{\boldsymbol{\theta}}_\tau^{(T)} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \widetilde{L}_\tau(\boldsymbol{\theta}), \quad \text{with} \quad \widetilde{L}_\tau(\boldsymbol{\theta}) := L_{1,\tau}(\boldsymbol{\theta}) - \left\langle \nabla L_{1,\tau}(\bar{\boldsymbol{\theta}}) - \nabla L_{N,\tau}(\bar{\boldsymbol{\theta}}), \boldsymbol{\theta} \right\rangle, \ \bar{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_\tau^{(T-1)}.$$

For simplicity of notations, we use $\widetilde{\boldsymbol{\theta}}_\tau$ to denote $\widetilde{\boldsymbol{\theta}}_\tau^{(T)}$, and $\bar{\boldsymbol{\theta}}$ to denote $\widetilde{\boldsymbol{\theta}}_\tau^{(T-1)}$ in what follows. A straightforward implementation of the multiplier bootstrap technique is to generate N

random weights $\{w_i\}_{i=1}^N$ for the overall data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ and consider the new global and local loss weighted by the multipliers $\{w_i\}_{i=1}^N$ to be

$$L_{N,\tau}^{\flat}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} w_i \ell_\tau(y_i \quad \bar{\boldsymbol{x}}_i^{\intercal}\boldsymbol{\theta}) \quad \text{and} \quad L_{1,\tau}^{\flat}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i \in \mathcal{I}_1} w_i \ell_\tau(y_i \quad \bar{\boldsymbol{x}}_i^{\intercal}\boldsymbol{\theta}).$$

Then similar to (3.2.1), we may construct the bootstrap surrogate loss with $L_{1,\tau}^{\flat}(\boldsymbol{\theta})$ and $L_{N,\tau}^{\flat}(\boldsymbol{\theta})$ as

$$\widetilde{L}_{\tau}^{\flat}(\boldsymbol{\theta}) := L_{1,\tau}^{\flat}(\boldsymbol{\theta}) \quad \left\langle \nabla L_{1,\tau}^{\flat}(\bar{\boldsymbol{\theta}}) \quad \nabla L_{N,\tau}^{\flat}(\bar{\boldsymbol{\theta}}), \boldsymbol{\theta} \right\rangle \tag{3.3.1}$$

to get the bootstrap estimator, which is expected to be able to mimic the distribution of the distributed estimator $\widetilde{\boldsymbol{\theta}}_\tau$, given the same initializer $\bar{\boldsymbol{\theta}}$. However, in this way, we have to calculate the new global gradient $\nabla L_{N,\tau}^{\flat}(\bar{\boldsymbol{\theta}})$, which calls for at least $\Omega(md)$ bits of communication for each bootstrap iteration. To achieve a reasonable approximation accuracy, bootstrap is typically conducted hundreds of times and thus the overall communication cost for this bootstrap procedure would be prohibitive. To avoid this issue, we consider generating random weights just for the local sample data stored in the master machine (i.e. $\{(\boldsymbol{x}_i, y_i)\}_{i \in \mathcal{I}_1}$) and define the bootstrap surrogate loss instead as

$$\widetilde{L}_{\tau}^{\flat}(\boldsymbol{\theta}) := L_{1,\tau}^{\flat}(\boldsymbol{\theta}) \quad \left\langle \nabla L_{1,\tau}(\bar{\boldsymbol{\theta}}) \quad \nabla L_{N,\tau}(\bar{\boldsymbol{\theta}}), \boldsymbol{\theta} \right\rangle, \tag{3.3.2}$$

and the resulting distributed bootstrap estimator is defined as

$$\widetilde{\boldsymbol{\theta}}_{\tau}^{\flat} = (\widetilde{\beta}_{0,\tau}^{\flat}, \widetilde{\boldsymbol{\beta}}_{\tau}^{\flat\intercal})^{\intercal} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \widetilde{L}_{\tau}^{\flat}(\boldsymbol{\theta}). \tag{3.3.3}$$

Noticeably, this multiplier bootstrap procedure does not require further communication across the machines, as the global gradient $\nabla L_{N,\tau}(\bar{\boldsymbol{\theta}})$ has been calculated in the distributed estimation phase.

The following theorem presents the theoretical properties of the distributed bootstrap estimator $\widetilde{\boldsymbol{\theta}}_{\tau}^{\flat}$, including an estimation error bound and a non-asymptotic Bahadur representation.

Theorem 3.3.1. Recall that $\mathsf{P}(\ ) = \mathsf{P}(\ |\mathcal{D}_N)$. Under Conditions 2.2.1-2.3.2, for any $t \quad 1/2$, $\widetilde{\boldsymbol{\theta}}_\tau^\flat$ defined in (3.3.3) satisfies that

$$\mathsf{P}\left(\|\widetilde{\boldsymbol{\theta}}_\tau^\flat \quad \boldsymbol{\theta}_\tau\|_{\mathbf{S}} \quad C^\flat\left(\sqrt{\frac{d+t}{n}}+\epsilon\right)\right) \quad 1 \quad 2e^{\ -t} \quad \text{and} \tag{3.3.4}$$

$$\mathsf{P}\left(\left\|c_F\mathbf{S}^{1/2}(\widetilde{\boldsymbol{\theta}}_\tau^\flat \quad \widetilde{\boldsymbol{\theta}}_\tau) \quad \frac{1}{n}\sum_{i\in\mathcal{I}_1}\ell_\tau'(\varepsilon_i \quad \alpha_\tau)e_i\boldsymbol{z}_i\right\|_2 \quad C^\flat\Big(\frac{d+t}{n}+\epsilon^2\Big)\right) \quad 1 \quad 5e^{\ -t} \tag{3.3.5}$$

with probability (over $\mathcal{D}_N$) at least $1 \quad (3T+14)e^{\ -t}$ as long as $T \quad C\log m/\log(n/(d+t))\_$ $\log m/\log(1/\epsilon)$, $\epsilon < c$ and $n \quad C'(d+t)^2$, where $c, C, C', C^\flat > 0$ are constants independent of (N, n, d, ϵ, t).

The Bahadur representations of the distributed estimator (3.2.4) and the distributed bootstrap estimator (3.3.5) imply the following distribution approximations.

$$\boldsymbol{\mu}^{\mathsf{I}}(\widetilde{\boldsymbol{\beta}}_\tau \quad \boldsymbol{\beta}\ ) = \boldsymbol{\lambda}^{\mathsf{I}}(\widetilde{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \quad S_N := \frac{1}{N}\sum_{i=1}^N U_i \quad \mathcal{N}(\mathsf{E}U_1, \frac{1}{N}VarU_1),$$

$$\boldsymbol{\mu}^{\mathsf{I}}(\widetilde{\boldsymbol{\beta}}_\tau^\flat \quad \widetilde{\boldsymbol{\beta}}_\tau) = \boldsymbol{\lambda}^{\mathsf{I}}(\widetilde{\boldsymbol{\theta}}_\tau^\flat \quad \widetilde{\boldsymbol{\theta}}_\tau) \quad S_{n,1}^\flat := \frac{1}{n}\sum_{i\in\mathcal{I}_1} e_iU_i \ \overset{|\mathcal{D}_n}{\sim}\ \mathcal{N}(0, \frac{1}{n^2}\sum_{i=1}^n U_i^2),$$

where $U_i := c_F^{\ 1}\ell_\tau'(\varepsilon_i \quad \alpha_\tau)\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S}^{\ 1/2}\boldsymbol{z}_i$, for $i = 1, \ldots, N$. It's noteworthy that the bootstrap estimator now only has $\sqrt{n}$-consistency, while the distributed M-estimator enjoys $\sqrt{N}$-consistency (assuming that the bias is negligible). This is the price that has to be paid for only bootstrapping a local sample (of size n). One can imagine that for the bootstrap estimator resulting from (3.3.1), it would generate an approximation to the distribution of $N^{\ 1}\sum_{i=1}^N e_iU_i$, which certainly provides a better approximation to the distribution of the distributed estimator, though with an exorbitant communication cost. This reflects the trade-off between the communication cost and the statistical efficiency.

Though with $\sqrt{n}$-consistency, after a re-scaling of $1/\sqrt{m}$, our proposed bootstrap estimator $\widetilde{\boldsymbol{\beta}}_\tau^\flat$ can be used to implicitly estimate the variance of the distributed estimator $\widetilde{\boldsymbol{\beta}}_\tau$, and thus further be able to approximate the distribution of $\widetilde{\boldsymbol{\beta}}_\tau$ when the bias $\mathsf{E}U_1$ is negligible to its variance i.e. $\epsilon = o(1/\sqrt{N})$, as proved by the following theorem.

Theorem 3.3.2. Under the same conditions of Theorem 3.3.1, for any $\boldsymbol{\mu} \in \mathbb{R}^d$ and any $t \geq 1/2$, it holds with probability (over $\boldsymbol{D}_N$) at least $1 - (3T+20)e^{-t}$ that

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\boldsymbol{\mu}^{\intercal}(\widetilde{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*) \leq x \right) - \mathbb{P}\left(\sqrt{\frac{1}{m}} \boldsymbol{\mu}^{\intercal}(\widetilde{\boldsymbol{\beta}}^\flat_\tau - \widetilde{\boldsymbol{\beta}}_\tau) \leq x \right) \right| \leq C\left(\sqrt{\frac{d+t}{n}} + \sqrt{\frac{p}{N}}\epsilon \right) + (3T+19)e^{-t}$$

for some constant $C > 0$ independent of $(N, n, d, \epsilon, t, \boldsymbol{\mu}, x)$.

The above results inspire us to define a $(1-\alpha)$ bootstrap confidence interval for coefficient β_j, based on the distributed estimator $\widetilde{\theta}_\tau := (\widetilde{\beta}_0, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_d)$ and the distributed bootstrap estimator $\widetilde{\theta}^\flat_\tau := (\widetilde{\beta}^\flat_0, \widetilde{\beta}^\flat_1, \ldots, \widetilde{\beta}^\flat_d)$, as

$$\widetilde{I}^\flat_j = \left[\widetilde{\beta}_j - \sqrt{\frac{1}{m}} \widetilde{c}^\flat_j(1-\alpha/2), \widetilde{\beta}_j - \sqrt{\frac{1}{m}} \widetilde{c}^\flat_j(\alpha/2) \right], \ j = 1, \ldots, d, \tag{3.3.6}$$

where $\widetilde{c}^\flat_j(1-\alpha/2)$ and $\widetilde{c}^\flat_j(\alpha/2)$ are $(1-\alpha/2)$-quantile and $(\alpha/2)$-quantile of $\widetilde{\beta}^\flat_j - \widetilde{\beta}_j$ (given $\boldsymbol{D}_N$), respectively, which are expected to approximate those counterparts of $\widetilde{\beta}_j - \beta_j$ well after re-scaling, as stated by the following theorem.

Algorithm 1 Distributed Bootstrap Inference

Input: Local Sample Data $\boldsymbol{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i \in \mathcal{I}_1}$ stored in the master machine, Global gradient $\nabla L_{N,\tau}(\bar{\boldsymbol{\theta}})$, estimator $\widetilde{\boldsymbol{\theta}}_\tau$, number of bootstrap iterations B

1: **for** $b = 1, \ldots, B$ **do**

2: Generate i.i.d weight random variables $\{w_i\}_{i=1}^n$ satisfying $w_i \geq 0, \mathbb{E}(w_i) = 1$, $\mathrm{var}(w_i) = 1$.

3: Obtain the bootstrap estimator $\widetilde{\boldsymbol{\theta}}^\flat_{\tau,b} := (\widetilde{\beta}^\flat_{0,b}, \widetilde{\beta}^\flat_{1,b}, \ldots, \widetilde{\beta}^\flat_{d,b})$ by (3.3.3).

4: **end for**

5: Compute the $\alpha/2$- and $(1-\alpha/2)$-quantile of $\{\widetilde{\beta}^\flat_{j,b} - \widetilde{\beta}_{j,b}\}_{b=1}^B$: $\widetilde{c}^\flat_j(1-\alpha/2)$ and $\widetilde{c}^\flat_j(\alpha/2)$.

Output: A bootstrap confidence interval of β_j is given as

$$\widetilde{I}^\flat_j = \left[\widetilde{\beta}_j - \sqrt{\frac{1}{m}} \widetilde{c}^\flat_j(1-\alpha/2), \widetilde{\beta}_j - \sqrt{\frac{1}{m}} \widetilde{c}^\flat_j(\alpha/2) \right], \ j = 1, \ldots, d.$$

Theorem 3.3.3 (Validity of distributed multiplier bootstrap)**.** Under the same conditions of Theorem 3.3.2, for $j = 1, \ldots, d$, we have

$$\sup_{q \in (0,1)} \left| \mathsf{P}\left(\widetilde{\beta}_j \quad \beta_j \quad \frac{1}{\sqrt{m}} \widetilde{c}_j^{\flat}(q) \right) \quad q \right| \quad C\left(\frac{d+t}{\sqrt{n}} + \sqrt{N}\epsilon \right) + C'Te^{-t},$$

where $C, C' > 0$ are constants independent of (N, n, d, ϵ, t). In particular, if $\epsilon = o(1/\sqrt{N})$, and $d = o(\sqrt{n})$, then we establish the validity of the bootstrap confidence interval for β_j in (3.3.6) :

$$\sup_{\alpha \in (0,1)} \left| \mathsf{P}\left(\beta_j \in \widetilde{I}_j^{\flat} \right) \quad (1 \quad \alpha) \right| = o(1),$$

as $n, N \to \infty$.

Remark 3.3.1 (Sample size requirement)**.** Our proposed distributed bootstrap method requires the local sample size $n \gtrsim d^2$. One can imagine that the bootstrap procedure based on (3.3.1) would only require the total sample size $N \gtrsim d^2$. This is due to the fact that this bootstrap method, by ignoring communication constraints, is nearly equivalent to the traditional centralized data arrangement where one has access to the entire data set. For cases where the local sample size is limited (e.g. $n \gtrsim d^2$ fails) but the total sample size is sufficient (e.g. $N \gtrsim d^2$ holds), we speculate that our proposed distributed bootstrap method may work with a slight modification. See the discussion section 3.6.

Note that Theorem 3.3.3 can be seen as a corollary of Theorem 3.3.2 and can be proved in an exactly same manner of Theorem 2.3.4 and thus its proof is omitted.

3.4 Distributed Debiased Confidence Intervals

In this section, we present the debiased confidence interval method in the distributed setting. Following the ideas in Section 2.4, we consider estimating c_F and $\mathbf{S}$ to approximate the upper bound of the bias due to contamination. To save communication cost, we only use the local sample stored on the master machine. Let

$$\widehat{c}_F = \frac{1}{n} \sum_{i \in \mathcal{I}_1} \ell_\tau''(\widehat{\varepsilon}_i), \quad \widehat{\mathbf{S}} = \frac{1}{n} \sum_{i \in \mathcal{I}_1} \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^{\intercal}, \quad \widetilde{D}_j := \epsilon c_1 \tau \left\| \frac{1}{\widehat{c}_F} \boldsymbol{\lambda}^{\intercal} \widehat{\mathbf{S}}^{-1/2} \right\|_2$$

where $\widehat{\varepsilon}_i = y_i - \bar{\boldsymbol{x}}_i^{\intercal}\widetilde{\boldsymbol{\theta}}_\tau$ is the residual for our distributed robust estimator $\widetilde{\boldsymbol{\theta}}_\tau$, and $\boldsymbol{\lambda} = (0, \ldots, 0, 1, 0, \ldots, 0)$ is a vector in R^{d+1} with only the $(j+1)$-th component being 1, for $1 \leq j \leq d$. Then $\widetilde{D}_j$ serves as an estimate of an upper bound of the contamination bias for estimating the j-th coefficient β_j. This implies that we may define a distributed debiased bootstrap confidence interval for β_j as following:

$$\widetilde{I}_{D,j}^{\flat} = \left[\widetilde{\beta}_j - C_D\left(\frac{1}{m}c_j^{\flat}(1-\alpha/2) + \widetilde{D}_j\right), \widetilde{\beta}_j - C_D\left(\frac{1}{m}c_j^{\flat}(\alpha/2) - \widetilde{D}_j\right)\right], \tag{3.4.1}$$

where $C_D \geq 1$ is an universal constant to be specified.

Theorem 3.4.1. Assume Conditions 2.2.1-2.3.2 hold. Further assume that the largest and smallest eigenvalue of $\mathbf{S}$ are both bounded: $c_{\mathbf{S}} \leq \lambda_{\min}(\mathbf{S}) \leq \lambda_{\max}(\mathbf{S}) \leq C_{\mathbf{S}}$ for some constants $C_{\mathbf{S}}, c_{\mathbf{S}}$. Then for any $j = 1, \ldots, d$, any $\alpha \in (0,1)$, any $C_D > 1$, we have

$$\mathsf{P}\left(\beta_j \in \widetilde{I}_{D,j}^{\flat}\right) \geq (1-\alpha) - C\left(\sqrt{\frac{\log n}{n}} + \epsilon^2\right)$$

as long as $\epsilon \leq c$ and $n \geq C'd^2$ for some large enough C' and small enough c. Here, c, C, C' are independent of (N, n, d, ϵ).

3.5 Numerical Studies

3.5.1 Distributed Robust Estimation

In this section, we investigate the numerical performance of the proposed procedures and compare the following methods: (i) global OLS estimator assuming that the entire $N = mn$ observations are directly available; (ii) global M-estimator using the entire N observations; (iii) the proposed robust distributed M-estimator.

For the M-estimators, we employ Huber's loss (2.2.3) and select the tuning parameter τ via the following heuristic two-step way: (i) initially set $\tau = 1$ and obtain the residuals using the available data; (ii) calculate the median absolute deviation (MAD) of the residuals and use it as the final value of τ. Ideally, this results in τ having the same scale as the variance of the noise ε under distribution F. For the proposed distributed estimators, we choose the

initializer $\widetilde{\boldsymbol{\theta}}^{(0)}$ to be the average of local estimators from each local machine, as suggested by Jordan et al. (2019). We set the number of iteration rounds $T = \lfloor \log m \rfloor + 1$, according to our theory in Section 3.2.

To generate contaminated data $(\boldsymbol{x}_i, y_i)$ following model (2.1.1), we consider $\boldsymbol{\beta} = \mathbf{1}_d$ (a d-dimensional vector of ones) and simulate $\boldsymbol{x}_i$ from standard multivariate normal distribution $N(\mathbf{0}, \boldsymbol{I}_d)$ ($\boldsymbol{I}_d$ is a $d \times d$ identity matrix). Regarding the distribution of the noise ε, we consider the following three heavy-tailed distributions for F: (i) t-distribution with 1.5 degrees of freedom, denoted as $t(1.5)$; (ii) Lognormal distribution with $\mu = 0$ and $\sigma = 1$, i.e., the logarithm of the standard normal distribution; (iii) Pareto distribution with shape parameter 1.5 and scale parameter 1. We investigate the following four types of contamination distribution G.

(i) G is a point mass at the first coordinate of $\boldsymbol{x}$;

(ii) G is a point mass at the sum of each coordinate of $\boldsymbol{x}$;

(iii) $G = Unif[-|y_0|, |y_0|]$, where $y_0 = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon$, with $\varepsilon \sim F$, is the initial response vector without contamination.

(iv) $G = sgn(y_0) \circ F$, where $sgn(\cdot)$ is the sign function.

The first two contamination distributions represent contamination based on covariates $\boldsymbol{x}$, and the last two are about contamination based on the response. To ensure a fair comparison for the distributed M-estimator without intercept, all of the aforementioned distributions for F and G are centered.

We report the estimation error of the coefficient $\boldsymbol{\beta}$ under the l_2-norm, i.e., $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$. Figure 2 shows the results over 100 trials, with $d = 50$, $n = 1000$, $m = 100$, and the contamination proportion $\epsilon = 1/\sqrt{N}$. We provide a summary of what Figure 2 revealed. First, the global OLS estimator always performs the worst since it is not robust to either heavy-tailed distribution or contamination. Second, Second, across all settings for various distributions F and G, our proposed distributed M-estimator always performs nearly as well as the global M-estimator, which has the least error and serves as the benchmark. This demonstrates that in the distributed context, our proposed estimator inherits the effectiveness and robustness of the centrialized M-estimator.

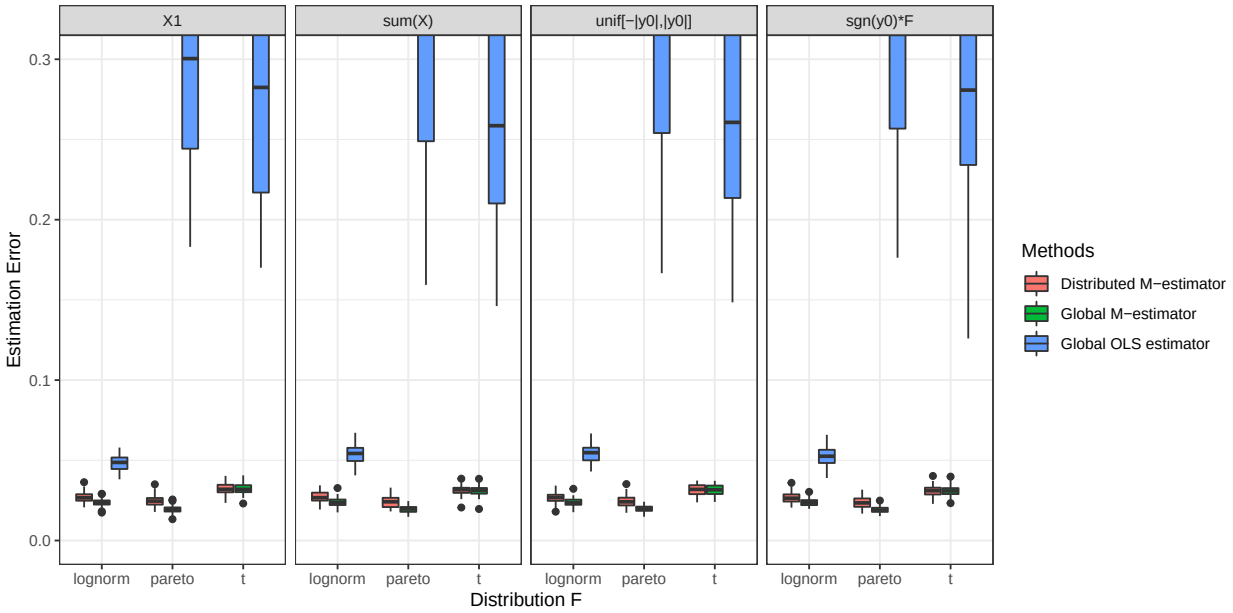


Figure 2: Distributed estimation results under various types of distributions for F and G with $d = 50$, $n = 1000$, $m = 100$, $\epsilon = 1/\sqrt{N}$. Each plot represents one type of contamination distribution G.

We then conduct a sensitivity analysis to study how the contamination proportion and scale affect the estimation accuracy. We consider the contamination proportion $\epsilon \in \{0, 0.02, 0.04, 0.06, 0.08, 0.1\}$, and the contamination distribution G to be a point mass at the first coordinate of $\boldsymbol{x}$ with various scales $\{1, 10, 100\}$, denoted by $G = s \quad x_1$ with $s \in \{1, 10, 100\}$. In other words, under this contamination distribution, the data would act as though the first component of $\boldsymbol{\beta}$ had its value boosted by $s + 1$ times. We choose F to be a t-distribution with 1.5 degrees of freedom, and show the results over 100 random trials in Figure 3, with $d = 50$, $n = 1000$, $m = 100$. From this figure, we see that the performance of the distributed M-estimator is almost as good as the global M-estimator. Both of them are robust to the contamination scale, and consistently have minimal estimation errors across all settings, compared to the OLS estimator. The global OLS estimator is particularly sensitive to the contamination proportion when the contamination scale is high.

For the scenario where the contamination scale is small ($s = 1$), though the estimation accuracy of the OLS estimator is close to M-estimators on average, its performance is very erratic.

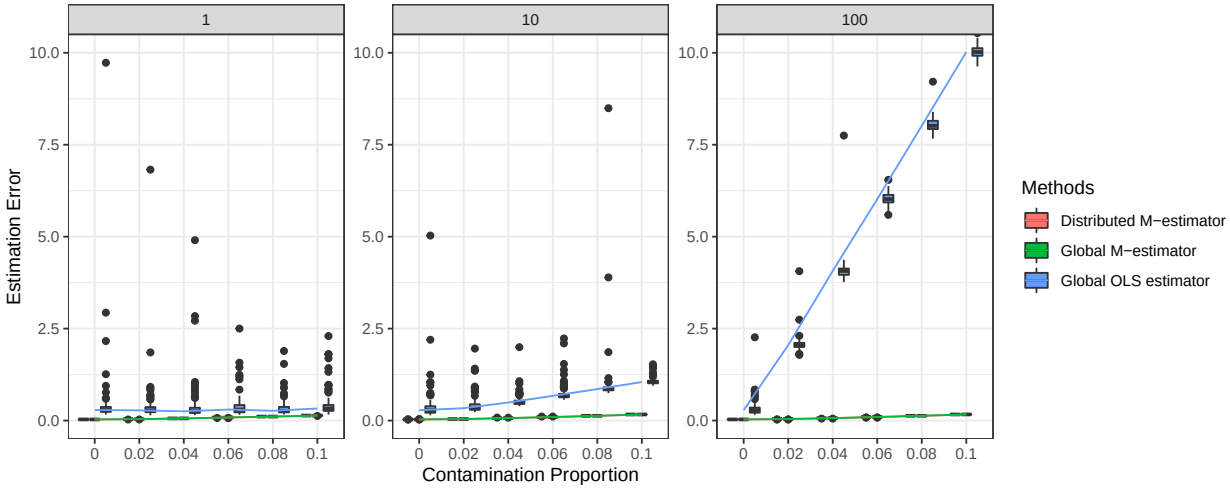


Figure 3: Distributed estimation results under various settings for contamination proportion ϵ and scale s. Results for $\epsilon \in \{0, 0.02, 0.04, 0.06, 0.08, 0.1\}$ and $s \in \{1, 10, 100\}$ are plotted with $F = t(1.5)$, $G = s \cdot x_1$ $d = 50$, $n = 1000$, $m = 100$. Each plot represents one type of contamination scale for G.

At last, we study the effect of the local sample sizes n and the number of machines m on the estimation performances of the aforementioned four methods. With a fixed total sample size of $N = 10^5$ and $m = N/n$, we explore altering the local sample size $n \in \{500, 1000, 2500, 5000, 10000\}$. We choose the distribution F as the logarithm of a standard normal distribution, the contamination distribution G as a point mass at $10x_1$, and the contamination proportion $\epsilon = 1/\sqrt{N}$. Figure 4 presents the results for 100 trials. We notice that when the local sample size grows, the performance of the distributed M-estimator (with intercept) quickly catches up to that of the global M-estimator. In fact, with the exception of a very small local sample size n ($n = 500$), their results are fairly comparable. This aligns with our theoretical analyses in Section 3.2 that the distributed M-estimator has the same convergence rate as the global M-estimator when $n \gtrsim d^2$. Additionally, this figure demon-

strates that the distributed M-estimator without intercept is always inferior to the one with intercept.

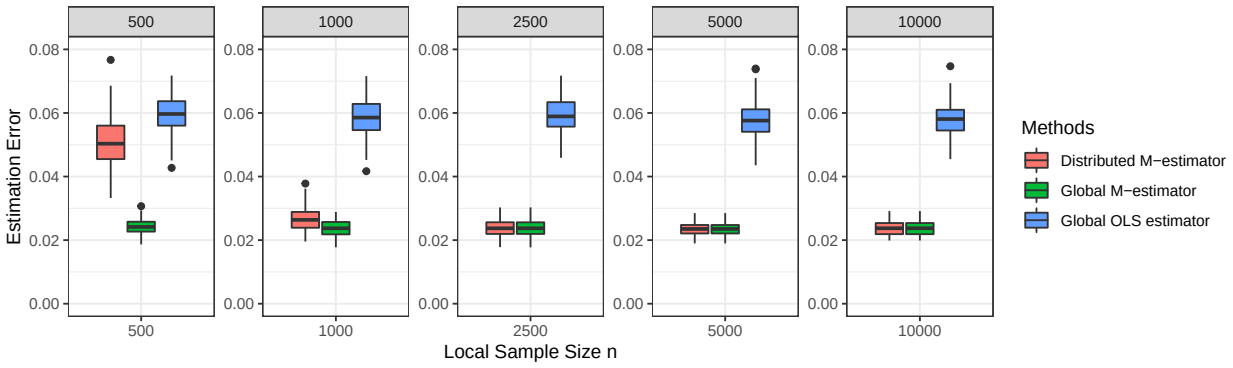


Figure 4: Distributed estimation results for local sample size $n \in \{500, 1000, 2500, 5000, 10000\}$, with $d = 50$, $N = 10^5$, $m = N/n$, $\epsilon = 1/\sqrt{N}$.

3.5.2 Distributed Robust Inference

We now evaluate the performance of the distributed inference procedures with contaminated data. We consider contrasting the following four methods: (i) M-boot: the proposed multiplier bootstrap procedure based on the distributed M-estimator; (ii) Debias-M-boot: the debiased bootstrap confidence intervals (3.4.1) with $C_D = 1$; (iii) M-Normal: the asymptotic normal-based approach for the distributed M-estimator; (iv) OLS-Normal: the standard asymptotic normal-based method for the global OLS-estimator assuming that all N observations are directly accessible.

The method M-Normal is based on the Bahadur representation (3.2.4), from which we have

$$\sqrt{N}\boldsymbol{\lambda}^{\intercal}(\widetilde{\boldsymbol{\theta}}_{\tau}^{(T)} - \boldsymbol{\theta}_{\tau}) \approx \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i \xrightarrow{d} \mathcal{N}(0, \mathsf{E}_{\varepsilon} F U_1^2), \quad \text{with} \quad \mathsf{E}_{\varepsilon} F U_1^2 = c_F^2 C_F \boldsymbol{\lambda}^{\intercal} \mathbf{S}^{-1} \boldsymbol{\lambda},$$

when the contamination bias is negligible i.e., $\epsilon = o(1/\sqrt{N})$, where $U_i := c_F^{-1}\ell_\tau'(\varepsilon_i - \alpha_\tau)\boldsymbol{\lambda}^\top \mathbf{S}^{-1/2}\boldsymbol{z}_i$, $c_F = \mathsf{E}_{\varepsilon \sim F}\ell_\tau''(\varepsilon - \alpha_\tau)$, $C_F = \mathsf{E}_{\varepsilon \sim F}\ell_\tau'(\varepsilon - \alpha_\tau)^2$. To save communication costs, we use the local sample data stored on the master machine to estimate c_F, C_F and $\mathbf{S}$ as follows:

$$\widehat{c}_F = \frac{1}{n}\sum_{i \in \mathcal{I}_1} \ell_\tau''(\widehat{\varepsilon}_i), \quad \widehat{C}_F = \frac{1}{n}\sum_{i \in \mathcal{I}_1} \ell_\tau'(\widehat{\varepsilon}_i)^2, \quad \widehat{\mathbf{S}} = \frac{1}{n}\sum_{i \in \mathcal{I}_1} \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top.$$

where $\widehat{\varepsilon}_i = y_i - \bar{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\theta}}_\tau^{(T)}$ is the residual. Then a resulting $(1-\alpha)$ confidence interval for the j-th coefficient β_j can be given by

$$\widetilde{I}_j^N = \left[\widetilde{\beta}_j - \frac{1}{\sqrt{N}}\Phi^{-1}(1-\alpha/2)\widehat{\sigma}_j, \widetilde{\beta}_j - \frac{1}{\sqrt{N}}\Phi^{-1}(\alpha/2)\widehat{\sigma}_j\right], \text{with } \widehat{\sigma}_j^2 = \widehat{c}_F^{-2}\widehat{C}_F\widehat{\mathbf{S}}^{-1}_{j+1,j+1},$$

where $\Phi^{-1}(\cdot)$ represents the quantile of standard normal distribution, $\widetilde{\beta}_j$ is the $(j+1)$-th component of $\widetilde{\boldsymbol{\theta}}_\tau^{(T)}$, $\widehat{\mathbf{S}}^{-1}_{j+1,j+1}$ is the $(j+1, j+1)$-th entry of the matrix $\widehat{\mathbf{S}}^{-1}$.

To generate data from model (2.1.1), we consider $\boldsymbol{\beta} = \mathbf{1}_d$ and $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. We choose F to be a t-distribution with degrees of freedom 1.5, and the contamination distribution G to be a point mass at $s \cdot x_1$ with $s \in \{1, 10, 100\}$. Since this contamination can be viewed as a perturbation of the first coefficient β_1, we focus on constructing a confidence interval for β_1. Regarding the contamination proportion, recall that the theoretical analyses in Section 3.3 indicate that both method (i) and (iii) would require the condition $\epsilon = o(1/\sqrt{N})$ to make the contamination bias small enough to be covered by the confidence intervals. As a result, we consider the contamination proportion $\epsilon = p/\sqrt{N}$ with $p \in \{0, 1/4, 1/2, 3/4, 1\}$.

Figure 5 presents the coverage proportions and mean widths of 95% confidence intervals for β_1 based on 500 random trials, with $d = 50$, $n = 1000$, $m = 100$. According to the coverage plots, the debiased bootstrap confidence intervals constantly have a coverage probability higher than 95%, whereas the coverage for the other three methods decreases when the contamination proportion ϵ increases, and falls below the nominal level 95% when ϵ approaches $1/\sqrt{N}$. This is consistent with our theory as the condition $\epsilon = o(1/\sqrt{N})$ fails. Compared to M-Normal, M-boot exhibits a more stable performance with a consistently smaller decline in coverage when the contamination proportion grows, which suggests that the bootstrap-based method is more resilient to higher contamination proportions than the

asymptotic normal-based approach. Note that the OLS-Normal confidence intervals, despite having decent coverage (especially when the contamination scale s is small), are generally useless because of their excessive widths.

From the width plots, we also observe that the widths of the debiased bootstrap confidence intervals grow (linearly) with the contamination proportion, while the lengths of the ones without debiasing (M-boot and M-Normal) remain the same. This phenomenon, along with the fact that both M-boot and M-Normal fail to reach the nominal level for higher contamination proportions, supports our argument that when ϵ approaches the scale of $1/\sqrt{N}$, the confidence intervals based on the pure variance approximation are no longer valid, and an appropriate debiasing procedure is required to address the (substantial) bias resulting from contamination.

3.6 Discussion

In this paper, we have presented a robust distributed estimation and inference framework for regression with distributed contaminated data. Specifically, we show a M-estimator with intercept, coupled with the CSL method Jordan et al. (2019) can achieve the centralized minimax rate of convergence under Huber's contamination model. Moreover, we design a communication-efficient multiplier bootstrap procedure for the the distributed robust M-estimator to construct a sharp confidence interval for each coefficient. We theoretically justify the validity of our distributed bootstrap inference method when the local sample size is not too small $d = o(\sqrt{n})$ and the contamination proportion is not too large $\epsilon = o(1/\sqrt{N})$. For the cases with larger contamination proportion, we further propose a debiased procedure, using local sample data for a bias correction, which is proved to be able to generate valid confidence intervals when contamination proportion $\epsilon = o(1)$.

For the cases with smaller local sample size, we may consider modifying our distributed

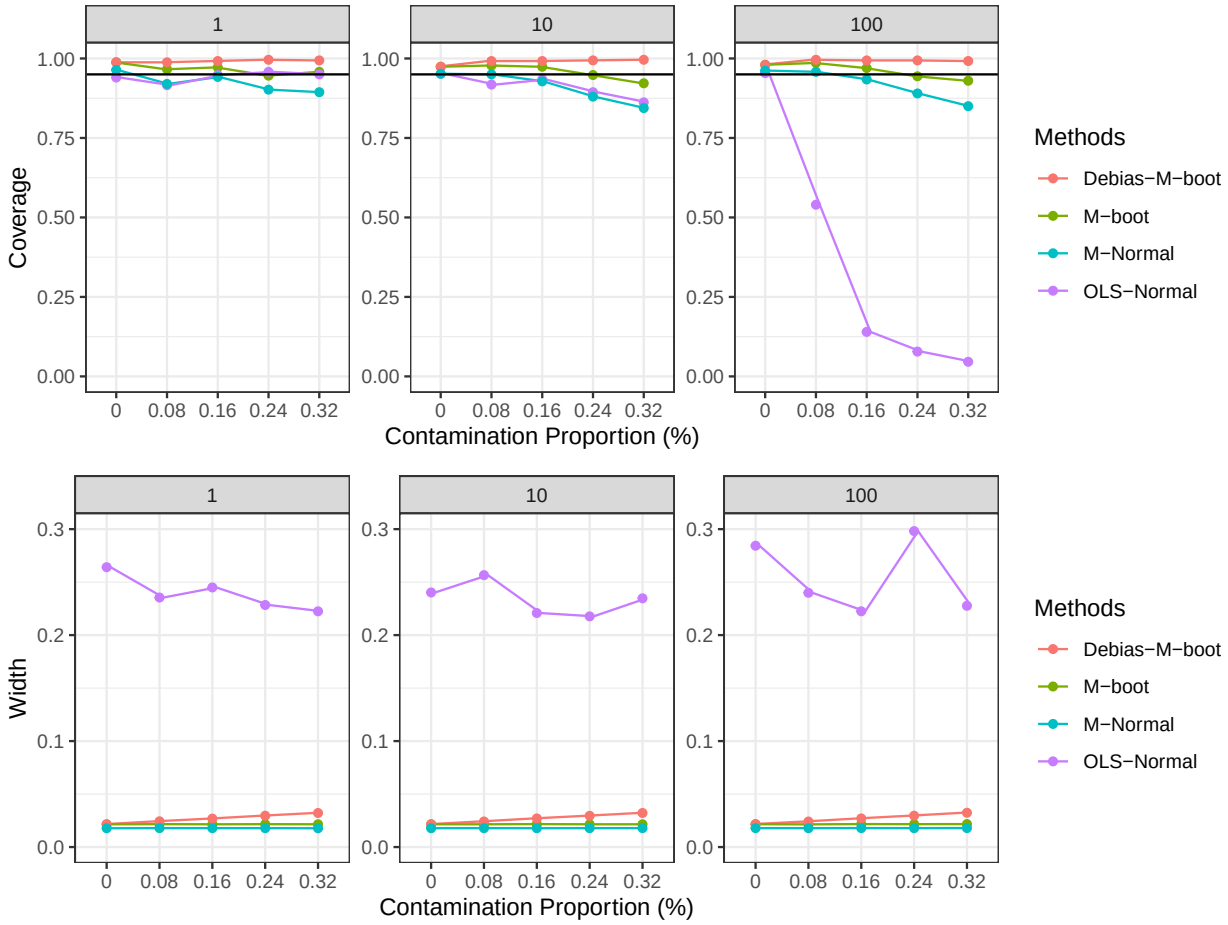


Figure 5: Coverage probabilities (upper) and widths (lower) of 95% confidence intervals for β_1 under various settings for contamination proportion ϵ and scale s. Results for $\epsilon = p/\overline{N}$, $p \in \{0, 1/4, 1/2, 3/4, 1\}$, $G = s \cdot x_1$, $s \in \{1, 10, 100\}$ are presented, with $d = 50$, $n = 1000$, $m = 100$. Each plot represents one type of contamination scale for G. In the coverage plots, the black horizontal line represents the nominal level 95%.

bootstrap procedure and define the following surrogate bootstrap loss:

$$\widetilde{L}^{\flat}_{\tau}(\boldsymbol{\theta}) := L^{\flat}_{1,\tau}(\boldsymbol{\theta}) \quad \left\langle r\, L^{\flat}_{1,\tau}(\bar{\boldsymbol{\theta}}) \quad r\, L^{\flat}_{N,\tau}(\bar{\boldsymbol{\theta}}), \boldsymbol{\theta} \right\rangle, \quad \text{with}$$

$$L^{\flat}_{1,\tau}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i \in I_1} w_i \ell_\tau(y_i \quad \bar{\boldsymbol{x}}_i^{\intercal}\boldsymbol{\theta}) \quad \text{and} \quad L^{\flat}_{N,\tau}(\boldsymbol{\theta}) := \frac{1}{m}\sum_{j=1}^{m} \widetilde{w}_j \left\{ \frac{1}{n}\sum_{i \in I_j} \ell_\tau(y_i \quad \bar{\boldsymbol{x}}_i^{\intercal}\boldsymbol{\theta}) \right\},$$

where $\{w_i\}_{i \in I_1}$ and $\{\widetilde{w}_j\}_{j=1}^m$ are all $i.i.d.$ copies of the weight variable and are independent of the sample data $\mathcal{D}_N$. This approach may be viewed as a communication-efficient approximation to the multiplier bootstrap method (3.3.1) since $\widetilde{L}^{\flat}_{N,\tau}(\boldsymbol{\theta})$ would mimic the global bootstrap loss

$$L^{\flat}_{N,\tau}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} w_i \ell_\tau(y_i \quad \bar{\boldsymbol{x}}_i^{\intercal}\boldsymbol{\theta})$$

well when m is large compared to n. Denote the resulting bootstrap estimator as $\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\mathsf{new}} = (\widetilde{\beta}^{\flat}_{0,\tau,\mathsf{new}}, \widetilde{\boldsymbol{\beta}}^{\flat\intercal}_{\tau,\mathsf{new}})^{\intercal}$, then we expect the corresponding bahadur representation would indicate the following distribution approximation, compared to those for the distributed M-estimator and the proposed bootstrap estimator in Section 3.3:

$$\text{(Distributed M-estimator)} \quad \boldsymbol{\mu}^{\intercal}(\widetilde{\boldsymbol{\beta}}_\tau \quad \boldsymbol{\beta}\ ) = \boldsymbol{\lambda}^{\intercal}(\widetilde{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \quad S_N := \frac{1}{N}\sum_{i=1}^{N} U_i,$$

$$\text{(Proposed bootstrap estimator)} \quad \boldsymbol{\mu}^{\intercal}(\widetilde{\boldsymbol{\beta}}^{\flat}_\tau \quad \widetilde{\boldsymbol{\beta}}_\tau) = \boldsymbol{\lambda}^{\intercal}(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau \quad \widetilde{\boldsymbol{\theta}}_\tau) \quad S^{\flat}_{n,1} := \frac{1}{n}\sum_{i \in I_1} e_i U_i,$$

$$\text{(New bootstrap estimator)} \quad \boldsymbol{\mu}^{\intercal}(\widetilde{\boldsymbol{\beta}}^{\flat}_{\tau,\mathsf{new}} \quad \widetilde{\boldsymbol{\beta}}_\tau) = \boldsymbol{\lambda}^{\intercal}(\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\mathsf{new}} \quad \widetilde{\boldsymbol{\theta}}_\tau) \quad S^{\flat}_m := \frac{1}{m}\sum_{j=1}^{m} \widetilde{e}_j \sum_{i \in I_j} U_i,$$

where $\widetilde{e}_j = \widetilde{w}_j \quad 1$ for $j = 1, \ldots, m$.Therefore, we expect that this method would be suitable to a small n large m setting. We leave the theoretical analysis for future work.

4.0 Robust Adaptive Minimax Density Estimation under Huber's Contamination Model

4.1 Introduction

In this chapter, we study the density estimation problem under Huber's contamination model. Formally, assume that we have $i.i.d.$ observations

$$X_1, ..., X_n \overset{i.i.d.}{\sim} P(\epsilon, f, G) := (1 \quad \epsilon)P_f + \epsilon G, \tag{4.1.1}$$

where P_f is a distribution on R^d with the density of interest f, G is an arbitrary contamination distribution, and ϵ represents the contamination proportion. Under this model, for each observation X_i, there is a probability of ϵ that it is drawn from some arbitrary contamination distribution G. Our goal is to robustly estimate the density f with the contaminated observations.

To characterize the exact influence of contamination on the optimal estimation of the density f, we investigate how the contamination affects the minimax rate of convergence for estimating f. The minimax approach is based on the assumption that f belongs to a given class of densities, and the accuracy of the estimator $\widehat{f}$ is measured by its maximal loss over this density class. To this end, one needs to specify the loss function and the function class that f is assumed to belong to. Here, we choose to use a general L_p ($1 \quad p < \mathsf{1}$) loss and assume f is in an anisotropic Nikol'skii class $\mathsf{N}_{p,d}(\beta_{\mathsf{0}}, L_{\mathsf{0}})$, which is a natural extension of Hölder class under L_p norm on R^d (see Section 4.2 for the precise definition of $\mathsf{N}_{p,d}(\beta_{\mathsf{0}}, L_{\mathsf{0}})$). An analogous estimation procedure and analysis can be made for the pointwise loss function when f is assumed to be in an anisotropic Hölder class; see Section 4.6 for a discussion. Here, $\beta_{\mathsf{0}} = (\beta_{\mathsf{0},1}, \ldots, \beta_{\mathsf{0},d})^{\mathsf{0}}$ is a vector with the jth component representing the smoothness condition of f with respect to the jth variable. We derive the following minimax rate for estimating the density f with contaminated observations

$$\inf_{\widehat{f}} \sup_{\substack{f \mathsf{2} \mathsf{N}_{p,d}(\beta_{\mathsf{0}}, L_{\mathsf{0}}) \\ G}} E_{P(\epsilon,f,G)} \mathsf{k}\, \widehat{f} \quad f\, \mathsf{k}_p \quad \begin{cases} n^{\frac{\beta_{\mathsf{0}}}{q\beta_{\mathsf{0}} + \mathsf{1}}} \_ \epsilon^{\frac{q\beta_{\mathsf{0}}}{q\beta_{\mathsf{0}} + \mathsf{1}}}, & 1 \quad p \quad 2, \\ n^{\frac{\beta_{\mathsf{0}}}{2\beta_{\mathsf{0}} + \mathsf{1}}} \_ \epsilon^{\frac{q\beta_{\mathsf{0}}}{q\beta_{\mathsf{0}} + \mathsf{1}}}, & p > 2. \end{cases}$$

Here, $1/q = 1 \quad 1/p$, and $\bar{\beta}_0 = 1/(\sum_{j=1}^{d} 1/\beta_{0,j})$ represents the average smoothness level of f. In the above formula, the first term $n^{\frac{\beta_0}{q\beta_0+1}}$ (or $n^{\frac{\beta_0}{2\beta_0+1}}$) is the classical minimax rate without contamination (Hasminskii and Ibragimov, 1990; Ibragimov and Khasminski, 1981; Ibragimov and Khas' minskii, 1980). The second term $\epsilon^{\frac{q\beta_0}{q\beta_0+1}}$ captures the effect of contamination on the optimal estimation of f, which depends not only on the contamination proportion ϵ, but also on the smoothness condition of f (i.e. β_0) and the specific loss function (i.e. p) as well.

In Section 4.2, kernel estimators (Devroye, 1985; Silverman, 2018) are shown to be able to achieve the above optimal rate. However, the optimal choice of the bandwidth requires prior knowledge of both the smoothness parameter β_0 and the contamination proportion ϵ, which naturally leads us to consider desired adaptive estimation procedures without using the information of β_0 and ϵ. To demonstrate the difficulty of such a task, we explain below that even if ϵ is known in advance, it is still quite challenging to find adaptive bandwidth selection procedures (adaptive to β_0 only).

(1) Adaptation in $\mathbb{R}^d$ is intrinsically not easy. Notice that $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,d})^{\prime}$ is a vector and adaptation to β_0 means adaptation to d parameters $\{\beta_{0,j}\}_{j=1}^{d}$ simultaneously. As a comparison, Liu and Gao (2019) considered the one-dimensional case (i.e. $d = 1$) and used Lepski's method (Lepskii, 1991, 1992, 1993) for adaptive estimation, which would not work in our case as Lepski's method essentially relies on the order topology of (totally ordered) real numbers, and thus can only adapt to one parameter. When the data are not contaminated, a more recent adaptive method called the Goldenshluger-Lepski method (Goldenshluger and Lepski, 2008; Lepski and Goldenshluger, 2009; Goldenshluger and Lepski, 2011a, 2013) was shown to be able to be adaptive to multiple parameters. But it is not robust to contamination. This leads to our second reason:

(2) Contamination brings extra complication to the analysis of adaptation theory.

- **Methodologically:** Classical adaptive methods like the Goldenshluger-Lepski method only need to consider the bias and variance trade-off of (kernel) estimators. For example, the Goldenshluger-Lepski method (Goldenshluger and Lepski, 2011a) considers selecting the bandwidth as the minimizer of a surrogate loss composed of an approximated bias term and an inflated variance term (called "variance majorant"). However, in our setting,

the contamination of the data distribution would break this well-aligned bias-variance structure and make the original surrogate loss unsuitable.

- **Theoretically:** Most adaptive methods require the *i.i.d.* (drawn from the true modeling distribution, e.g. P_f in our setting) assumption to make it work. This assumption fails due to contamination. If $P(\epsilon, P_f, G)$ had a density (say $p(\epsilon, f, G)$), then we may pretend it to be our target density and apply the adaptive method to the contaminated data, and then analyze the bias between $p(\epsilon, f, G)$ and the density of interest f. Unfortunately, this alternative approach fails either in our case because that under Huber's contamination model, the contamination distribution G can be arbitrary, and thus the overall data distribution $P(\epsilon, P_f, G)$ may not even have a density. This makes it seem entirely impossible to adopt the analysis directly from existing adaptive methods.

In Section 4.3, we tackle those potential challenges by introducing a majorant for the contamination term (called "contamination majorant") and carefully balancing the trade-off among the bias, variance, and contamination terms of kernel estimators. Our adaptive bandwidth selection procedure can be seen as a robust generalization of the Goldenshluger-Lepski method and we show that it can lead to a minimax estimator adaptive to the smoothness parameter β_0 (assuming ϵ is known). In the same spirit, we are able to develop a data-driven procedure adaptive to ϵ (assuming β_0 is known). Notice that in this case, adaption to ϵ (just one parameter) seems to fall in the scope of Lepski's method (Lepskii, 1991, 1992, 1993). However, it is not straightforward to apply it as the key part of Lepski's method is some concentration inequality of the variance term under L_p norm (see e.g. Lemma C.3.1 in the supplementary material, rephrased from Goldenshluger and Lepski (2011a,b)), which does not hold any more due to the contamination effect. We use an innovative induction method to get around this problem, which may be of independent interest.

For the case where both β_0 and ϵ are unknown (or at least one component of β_0 and ϵ are unknown), we prove a surprising result that for any given target convergence rate of the form $n^{-R_1(\beta_0)} \epsilon^{R_2(\beta_0)}$ (with any two given positive functions $R_1(\,)$ and $R_2(\,)$), there is no data-driven procedure that can achieve this rate while being adaptive to a scale of β_0 and ϵ. This phenomenon, first pointed out by Liu and Gao (2019) under the pointwise loss, illustrates the contamination effect on the adaptation theory of density estimation.

Finally, we briefly discuss the structured contamination case, where the contamination distribution G is assumed to have a density g. We show that as long as g has a finite L_p norm, the minimax rate for estimating f can be improved to the order $n^{-\frac{\beta_0}{q\beta_0+1}} \wedge \epsilon$ for $p \geq [1,2]$ and $n^{-\frac{\beta_0}{2\beta_0+1}} \wedge \epsilon$ for $p \geq (2, \infty)$. We further show that even with an additional smoothness assumption (e.g. g also belongs to certain anisotropic Nikol'skii class $\mathcal{N}_{p,d}(\beta_1, L_1)$), the minimax rate would not improve anymore. This implies that the smoothness of the contamination density g has no effect on the optimal estimation of f. The adaptive method is also provided for this case and shown to be minimax adaptive to β_0 (we don't consider adaptation to ϵ here as the oracle choice of bandwidth depends on β_0 only).

4.1.1 Related Works and Contribution on Robust Adaptive Bandwidth Selection

Classical minimax and adaptive minimax density estimation (without contamination) under L_p loss functions on $\mathbb{R}^d$ were considered extensively in literature, see for example, (Devroye, 1985; Devroye and Lugosi, 2012; Goldenshluger and Lepski, 2014; Hasminskii and Ibragimov, 1990; Ibragimov and Khasminski, 1981; Ibragimov and Khas' minskii, 1980; Tsybakov, 2008), especially (Goldenshluger and Lepski, 2014) for a thorough review. However, the adaptive minimax density estimation with contaminated data has rarely been studied, except some very recent works (Liu and Gao, 2019; Chen et al., 2016). In this work, we develop adaptive bandwidth selection procedures for kernel density estimators under Huber's contamination model via a generalization of the Goldenshluger-Lepski method. The Goldenshluger-Lepski method (Goldenshluger and Lepski, 2008; Lepski and Goldenshluger, 2009; Goldenshluger and Lepski, 2011a, 2013), as a multi-dimensional extension of Lepski's method (Lepskii, 1991, 1992, 1993), can be used to construct an adaptive bandwidth selection procedure for kernel estimators on $\mathbb{R}^d$. Specifically, in Goldenshluger and Lepski (2011a), the authors showed that the resulting kernel estimator is minimax adaptive over a scale of the anisotropic Nikol'skii classes. However, as discussed above, this method would face difficulties with contaminated data, both methodologically and theoretically.

To overcome such difficulties, we treat Huber's contamination model (4.1.1) as a two-class

mixture model and analyze the "clean" data and "contaminated" data separately. Notice that Huber's contamination model (4.1.1), as a two-class mixture model, can be re-written in the following hierarchical form

$$X_i \mid \pi_i = 1 \quad P_f, \quad X_i \mid \pi_i = 0 \quad G, \quad i = 1, \ldots, n$$

where $\pi_1, \ldots, \pi_n \overset{i.i.d.}{\sim} Bernoulli(1 \quad \epsilon)$ are latent variables, with each π_i indicating whether the sample observation X_i is drawn from P_f or G. Denote $I_1 = \{1 \quad i \quad n : \pi_i = 1\}$ and $I_0 = \{1 \quad i \quad n : \pi_i = 0\}$ as the index sets for observations generated from P_f and G, respectively. Note that the total number of observations drawn from P_f is $n_1 = \sum_{i=1}^{n} \pi_i$ $Binomial(n, 1 \quad \epsilon)$. As illustrated later, our approaches (see, e.g. (4.2.2)) do not rely on the order of $X_1, \ldots, X_n$. Thus without loss of generality, we assume that $I_1 = \{1, \ldots, n_1\}$ and $I_0 = \{n_1 + 1, \ldots, n\}$. Then given the value of n_1 (or more precisely the values of $\{\pi_i\}_{i=1}^{n}$), we have

$$X_1, ..., X_{n_1} \overset{i.i.d.}{\sim} P_f, \quad X_{n_1+1}, ..., X_n \overset{i.i.d.}{\sim} G. \tag{4.1.2}$$

By conditioning on these latent variables $\{\pi_i\}_{i=1}^{n}$, we can separately consider the clean observations (generated from P_f) from the contaminated ones (generated from G). Our strategy is to apply the technique of the Goldenshluger-Lepski method to the clean observations $\{X_1, \ldots, X_{n_1}\}$, while the contaminated part is separately treated and eventually controlled by a term called contamination majorant. This is analogous to the variance majorant in the Goldenshluger-Lepski method (Goldenshluger and Lepski, 2011a,b), which bounds the variance of kernel estimators uniformly, as summarized in Lemma C.3.1. However, the variance majorant itself is a statistic relying on the whole data set for $p \in (2, \infty)$; thus, it is also bothered by the contamination effect. We show that this contamination effect can be covered by our predefined contamination majorant. With the help of the contamination majorant, we generalize Lemma C.3.1 under Huber's contamination model (see Lemma C.4.1 for further details).

To summarize, we construct a novel contamination majorant term and add it to the original surrogate loss to characterize the effect of contamination on the kernel estimators. From the methodological perspective, we make the new surrogate loss a good proxy of the

actual loss with contaminated data. Theoretically, by analyzing Huber's contamination model hierarchically, we can utilize the results of non-robust Goldenshluger-Lepski method and prove that our new bandwidth selection method can lead to minimax estimators with contaminated data. In what follows, we always consider model (4.1.2) conditioning on n_1 or equivalently $\{\pi_i\}_{i=1}^n$.

4.1.2 Organization and Notation

The rest of the chapter is organized as follows. In Section 4.2, we establish the minimax rate of the density estimation under Huber's contamination model by deriving the upper bound and lower bound, respectively. In Section 4.3, we develop new data-driven procedures to select the bandwidth of kernel estimators when either β_0 or ϵ is known, and prove the impossibility of adaptive methods when neither of them is known. We further provide the results of the structured contamination case, including the minimax rates and adaptive methods in Sections 4.4 and 4.5, respectively. We draw the conclusion in Section 4.6 by discussing how to apply the ideas and techniques we develop in this paper to the case under the pointwise loss function. Proofs of all the results are given in the supplementary material.

Notations. We introduce a few notations that will be used in this chapter. For $a, b \in \mathbb{R}$, let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We use a_+ to denote $\max\{a, 0\}$. For a positive real number a, $\lfloor a \rfloor$ is the largest integer strictly less than a. For two vectors $\beta_0, \beta_1 \in \mathbb{R}^d$, we write $\beta_0 \preceq \beta_1$ to represent that every component of β_0 is smaller than or equal to that of β_1. For two probability measures P_1, P_2 on a σ-algebra $\mathcal{F}$, their total variation distance is defined as $\mathrm{TV}(P_1, P_2) = \sup_{A \in \mathcal{F}} |P_1(A) - P_2(A)|$. We use $\mathbb{1}(\,)$ to denote the indicator function. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if $a_n \le Cb_n$ for all n and some positive constant C independent of n, and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We want to remind the readers that in this chapter, we use P and E to represent probability and expectation, instead of $\mathbb{P}$ and $\mathbb{E}$ as used in Chapter 2 and 3.

4.2 Minimax Rate under Huber's Contamination Model

Assume that we observe $X_1, ..., X_n$ from contamination model (4.1.1). Our goal is to estimate f with the contaminated data $\{X_i\}$, and the accuracy of estimators is measured by the L_p ($1 \leq p < \infty$) loss. In this work, f is assumed to be in an anisotropic Nikol'skii class, which is defined as the following.

Definition 4.2.1. Let $p \in [1, \infty)$, $\beta = (\beta_1, \ldots, \beta_d)^\intercal$, $\beta_j > 0$, and $L > 0$. We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ belongs to the anisotropic Nikol'skii class $\mathcal{N}_{p,d}(\beta, L)$ if

(i) $\|D_j^{(k)} f\|_p \leq L$, for $k = 0, \ldots, \lfloor \beta_j \rfloor$, $j = 1, \ldots, d$;

(ii) for all $j = 1, \ldots, d$ and all $z \in \mathbb{R}$,

$$\left\{ \int \left| D_j^{(\lfloor \beta_j \rfloor)} f(t_1, \ldots, t_j + z, \ldots, t_d) - D_j^{(\lfloor \beta_j \rfloor)} f(t_1, \ldots, t_j, \ldots, t_d) \right|^p \mathrm{d}t \right\}^{1/p} \leq L|z|^{\beta_j - \lfloor \beta_j \rfloor}.$$

Here $D_j^{(k)} f$ denotes the kth-order partial derivative of f with respect to the variable t_j, and $\|f\|_p$ represents the L_p norm of f. If β_j is an integer, then assumption (ii) implies that f's partial derivative $D_j^{(\lfloor \beta_j \rfloor)} f$ is Lipschitz with respect to L_p norm. The function classes $\mathcal{N}_{p,d}(\beta, L)$ were first considered in approximation theory by Nikol'skii (see for example, (Nikol'skii, 2012)). The minimax density estimation problem for such function classes was solved by Ibragimov and Khasminski (Hasminskii and Ibragimov, 1990; Ibragimov and Khasminski, 1981; Ibragimov and Khas' minskii, 1980). The adaptive estimation over function classes $\mathcal{N}_{p,d}(\beta, L)$ was considered later in (Goldenshluger and Lepski, 2011a).

We further define $\mathcal{P}_{p,d}(\beta, L) := \{f : f : \mathbb{R}^d \to \mathbb{R} | f \geq 0, \int f = 1, f \in \mathcal{N}_{p,d}(\beta, L)\}$ to denote the set of density functions that are in some anisotropic Nikol'skii class $\mathcal{N}_{p,d}(\beta, L)$.

Theorem 4.2.1. *Assume that $f \in \mathcal{P}_{p,d}(\beta_0, L_0)$, where $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,d})^\intercal$. Let $1/\bar{\beta}_0 = \sum_{j=1}^d 1/\beta_{0,j}$, $1/q = 1 - 1/p$. If $\bar{\beta}_0 > 1/p$, then the minimax rate is*

$$\inf_{\widehat{f}} \sup_{\substack{f \in \mathcal{P}_{p,d}(\beta_0, L_0) \\ G}} E_{P(\epsilon, f, G)} \|\widehat{f} - f\|_p \asymp \begin{cases} n^{-\frac{\beta_0}{q\beta_0 + 1}} + \epsilon^{\frac{q\beta_0}{q\beta_0 + 1}}, & 1 \leq p \leq 2 \\ n^{-\frac{\beta_0}{2\beta_0 + 1}} + \epsilon^{\frac{q\beta_0}{q\beta_0 + 1}}, & p > 2. \end{cases} \tag{4.2.1}$$

The first term $n^{-\frac{\beta_0}{q\beta_0+1}}$ (or $n^{-\frac{\beta_0}{2\beta_0+1}}$) is the classical minimax rate without contamination. The second term $\epsilon^{\frac{q\beta_0}{q\beta_0+1}}$ captures the effect of contamination on the optimal estimation of f. The way it depends on β_0 implies that the smoother f is, the smaller the contamination effect is. Thus, the smoothness parameter β_0 of f can also be an indicator of the robustness of the optimal estimation of f. Some special cases: (1) $p = 1$: $n^{-\frac{\beta_0}{q\beta_0+1}}$ is of order $O(1)$, which shows the phenomenon that smoothness alone is not sufficient to guarantee consistency of density estimators in $L_1(\mathsf{R}^d)$, as discussed in (Ibragimov and Khasminski, 1981). The term $\epsilon^{\frac{q\beta_0}{q\beta_0+1}}$ becomes ϵ, which matches the general lower bound for L_1 loss under Huber's contamination model, provided by (Chen et al., 2016). (2) $p = 2$: the minimax rate (4.2.1) is of order $n^{-\frac{\beta_0}{2\beta_0+1}} \vee \epsilon^{\frac{2\beta_0}{2\beta_0+1}}$, which indicates that the phase transition boundary occurs at $\epsilon \asymp n^{-1/2}$. In other words, if the contamination proportion ϵ is lower than the level $n^{-1/2}$ asymptotically, then the contamination would not affect the minimax rate. If ϵ is above this level, then no density estimator can achieve the classical minimax rate as the contamination term $\epsilon^{\frac{2\beta_0}{2\beta_0+1}}$ is dominant in the minimax rate (4.2.1). Therefore, to achieve the classical minimax rate, there can be (approximately) at most $n\epsilon \asymp n^{1/2}$ contaminated observations.

We then briefly discuss how we establish the upper and lower bounds of (4.2.1), respectively in the following subsections. The detailed proof is given in Section C.2 of the supplementary material.

4.2.1 Upper Bound

The minimax rate (4.2.1) can be achieved by a kernel density estimator that takes the form

$$\widehat{f}_h(x) = \frac{1}{nV_h}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \tag{4.2.2}$$

where $h = (h_1, \ldots, h_d)$ is the bandwidth vector, $K : \mathsf{R}^d \to \mathsf{R}$ is some kernel function, $V_h = \prod_{j=1}^{d} h_j$, u/v for $u, v \in \mathsf{R}^d$ represents the coordinate-wise division, and $K_h(\cdot) = V_h^{-1} K(\cdot/h)$. In this setting, we consider a family of kernels satisfying the following assumptions:

- (K1) $K(t) := \prod_{j=1}^{d} K_j(t_j)$, $K_j : \mathsf{R} \to \mathsf{R}$ is some kernel function on R.

- (K2) For all $j = 1, \ldots, d$, K_j is a kernel of order $l_j = \lfloor \beta_j \rfloor$ and $\int |u|^{\beta_j} |K_j(u)| du < \infty$. That is, for any $j = 1, \ldots, d$,

$$\int_{\mathbb{R}} K_j = 1, \quad \int_{\mathbb{R}} u^s K_j(u) du = 0, s = 1, \ldots, \lfloor \beta_j \rfloor, \quad \int_{\mathbb{R}} |u|^{\beta_j} |K_j(u)| du \leq L_K < \infty$$

- (K3) For any $j = 1, \ldots, d$, $\|K_j\|_1 \vee \|K_j\|_2 \vee \|K_j\|_p \vee \|K_j\|_\infty \leq L_K < \infty$.

We use the notation $\mathcal{K}_\beta(L_K)$ to denote the collection of all the kernels satisfying (K1)-(K3).

Remark. Given some β, it is not hard to find a kernel in $\mathcal{K}_\beta(L_K)$. For example, let $K_0(u)$ be a bounded and compactly supported function on $\mathbb{R}$ and $\int K_0(u) du = 1$. Following (Kerkyacharian et al., 2001), for some integer $l \geq \max_{1 \leq j \leq d} \beta_j$ we define

$$K_l(u) = \sum_{k=1}^{l} \binom{l}{k} (-1)^{k+1} \frac{1}{k} K_0\left(\frac{u}{k}\right), \quad K(t) = \prod_{j=1}^{d} K_l(t_j).$$

It is easy to check such kernel K_l satisfies (K1) – (K3) with some constant L_K.

Theorem 4.2.2. *Under the assumptions of Theorem 4.2.1, for a kernel estimator $\widehat{f}_h$ (4.2.2) with some kernel $K \in \mathcal{K}_\beta(L_K)$ for some $\beta \geq \beta_0$, we have*

$$\sup_{f \in \mathcal{P}_{p,d}(\beta_0, L_0)} E_{P(\epsilon,f,G)} \|\widehat{f}_h - f\|_p^p \leq C \left\{ \sum_{j=1}^{d} h_j^{\beta_{0,j}} + (nV_h)^{-1/(q-2)} + \epsilon V_h^{-1/q} \right\}^p, \tag{4.2.3}$$

for any h such that $0 < V_h \leq 1$, $\epsilon \leq C_0 < 1$ and $n \geq C_1$, where C_0, C_1 are two absolute constants, and C is a constant depending only on L_0, L_K, C_0.

Remark. By choosing $h_j \asymp n^{-\frac{\beta_0}{\beta_{0,j}((q-2)\beta_0+1)}} \wedge \epsilon^{\frac{q\beta_0}{\beta_{0,j}(q\beta_0+1)}}$, we immediately obtain that the kernel estimator $\widehat{f}_h(x)$ can achieve the minimax rate (4.2.1). Noticeably, the optimal choice of the bandwidth depends on not only the smoothness parameter β_0, but the contamination proportion ϵ as well. We will discuss how to adaptively choose the bandwidth in Section 4.3.

Remark. The first two terms $\sum_{j=1}^{d} h_j^{\beta_{0,j}}$ and $(nV_h)^{-1/(q-2)}$ in (4.2.3) are the classical upper bounds for the bias and variance of the kernel estimator, respectively, when assuming there is no contamination, i.e. $\{X_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} P_f$. The last term $\epsilon V_h^{-1/q}$ reflects the contamination effect on the kernel density estimation.

To be more specific, we introduce the following notations to denote the bias and variance parts of the kernel estimator $\widehat{f}_h$. For $1 \leq m \leq n$, assume $X_1, ..., X_m \overset{i.i.d.}{\sim} p_X$ (the density of clean observations), and define the bias and variance term of the kernel estimator $\frac{1}{m}\sum_{i=1}^{m} K_h(X_i - x)$ as:

$$B_h(p_X, t) := E_{p_X} K_h(t - X) - p_X(t) = \int_{\mathbb{R}^d} K_h(t - x) p_X(x) dx - p_X(t), \tag{4.2.4}$$

$$\xi_{h,m}(p_X, t) := \frac{1}{m}\sum_{i=1}^{m} [K_h(t - X_i) - E_{p_X} K_h(t - X)]. \tag{4.2.5}$$

Recall that Huber's contamination model (4.1.1) can be rewritten as (4.1.2) for the analysis of kernel estimators, conditioning on n_1 or equivalently $\{\pi_i\}_{i=1}^{n}$. With the notations (4.2.4)-(4.2.5) and setting $m = n_1$, we can have the following decomposition of the L_p loss of the kernel estimator (4.2.2)

$$\begin{aligned} \|\widehat{f}_h - f\|_p &\leq \left\| \frac{1}{n}\sum_{i=1}^{n_1} (K_h(t - X_i) - f) \right\|_p + \left\| \frac{1}{n}\sum_{i=n_1+1}^{n} (K_h(t - X_i) - f) \right\|_p \\ &\leq \frac{n_1}{n}\left(\|B_h(f,t)\|_p + \|\xi_{h,n_1}(f,t)\|_p \right) + \frac{n - n_1}{n}\left(\|K_h\|_p + \|f\|_p \right). \end{aligned} \tag{4.2.6}$$

Under the setting of Theorem 4.2.2, it can be shown that

- **Bias term:**

$$\|B_h(f,t)\|_p \asymp \sum_{j=1}^{d} h_j^{\beta_{0,j}}.$$

- **Variance term:**

$$E_{P(\epsilon,f,G)} \|\xi_{h,n_1}(f,t)\|_p \asymp (nV_h)^{-1/(q-2)}.$$

- **Contamination term:**

$$E_{P(\epsilon,f,G)} \frac{n - n_1}{n}\left(\|K_h\|_p + \|f\|_p \right) \asymp \epsilon V_h^{-1/q}.$$

4.2.2 Lower Bound

In (Chen et al., 2018), a general lower bound is provided for Huber's ϵ-contamination model. Given a general statistical experiment $\{P_\theta, \theta \in \Theta\}$ under Huber's contamination model $P(\epsilon, \theta, Q) = (1 - \epsilon)P_\theta + \epsilon G$, a key quantity, called modulus of continuity, is defined as

$$\omega(\epsilon, \Theta) = \sup \{L(\theta_1, \theta_2) : \mathrm{TV}(P_{\theta_1}, P_{\theta_2}) \le \epsilon/(1 - \epsilon); \theta_1, \theta_2 \in \Theta\},$$

where $L(\ ,\ )$ is the loss function. This definition of modulus of continuity dates back to (Donoho, 1994; Donoho and Liu, 1991), and (Chen et al., 2018) shows that the optimal estimation error rate can be generally lower bounded by this quantity with the following theorem.

Lemma 4.2.3 (Theorem 5.1 in (Chen et al., 2018))**.** *Suppose there is some $M(0)$ such that*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \sup_{Q} \mathsf{P}_{P(\epsilon,\theta,Q)} \{L(\widehat{\theta}, \theta) \ge M(\epsilon)\} \ge c$$

holds for $\epsilon = 0$. Then for any $\epsilon \in [0, 1]$, it holds that $M(\epsilon) \ge M(0) - \omega(\epsilon, \Theta)$.

The intuition behind the above result is that Huber's ϵ-contamination model $P(\epsilon, \theta, G)$ can be treated as a perturbation of the true distribution P_θ under the total variation distance at the order of ϵ. In fact, any two distributions $P_{\theta_1}, P_{\theta_2}$ with total variation bounded by $\epsilon/(1 - \epsilon)$ cannot be distinguished under Huber's contamination model. That is, we can always find two distributions G_1, G_2 such that $P_{\theta_1} - P_{\theta_2} = \frac{\epsilon}{1-\epsilon}(G_2 - G_1)$, i.e. $P(\epsilon, \theta_1, G_1) = P(\epsilon, \theta_2, G_2)$. Therefore, a price of $L(\theta_1, \theta_2)$ has to be paid for the estimation of θ.

In our setting, we can show that $\omega(\epsilon, \Theta) \gtrsim \epsilon^{\frac{q\beta_0}{q\beta_0+1}}$ and establish the following lower bound.

Theorem 4.2.4. *Under the assumptions of Theorem 4.2.1, we have*

$$\inf_{\widehat{f}} \sup_{f \in P_{p,d}(\beta_0, L_0)} E_{P(\epsilon,f,G)} \|\widehat{f} - f\|_p^2 \gtrsim n^{-\frac{\beta_0}{(q-2)\beta_0+1}} + \epsilon^{\frac{q\beta_0}{q\beta_0+1}}.$$

Remark. Specifically, we consider the following construction of two densities f_0, f_1. Choose a density $f_0 \in P_{p,d}(\beta_0, L_0/2)$ bounded away from zero, and let

$$f_1(x) = f_0(x) + \gamma V_h^{\beta_0} \prod_{j=1}^{d} \phi_0\Big(\frac{x_j}{h_j}\Big),$$

where $h_j = \epsilon^{\frac{q\beta_0}{\beta_{0,j}(q\beta_0+1)}}$, $\phi_0 : \mathbb{R} \to \mathbb{R}$ is an infinitely differentiable function with the support $[-1, 1]$ and satisfies $\int \phi_0 = 0$. By setting γ sufficiently small, one can obtain that $f_1 \in P_{p,d}(\beta_0, L_0)$ and $\mathrm{TV}(P_{f_0}, P_{f_1}) \leq \frac{\epsilon}{1-\epsilon}$, and thus, $\omega(\epsilon, \Theta) \geq \|f_1 - f_0\|_p \geq \epsilon^{\frac{q\beta_0}{q\beta_0+1}}$.

4.3 Adaptive Density Estimation

In this section, we discuss the adaptive density estimation under Huber's contamination model. To this end, we focus on how to adaptively select the bandwidth of the kernel estimator to make it achieve the optimal minimax rate (4.2.1). As we mentioned in Remark 4.2.1, the optimal choice of the bandwidth requires a prior knowledge of both the smoothness parameter β_0 and the contamination proportion ϵ. Therefore, there are three cases of adaptation to be investigated.

(1) Only the smoothness parameter β_0 is unknown;

(2) Only the contamination proportion ϵ is unknown;

(3) Neither β_0 nor ϵ is known.

For the first two cases where only one parameter is unknown, we develop adaptive bandwidth selection procedures and prove the resulting kernel estimators can achieve the minimax rate in Sections 4.3.1 and 4.3.2. Regarding the last case, in Section 4.3.3 we prove that it is impossible to obtain an adaptive method that achieves the optimal minimax rate. In fact, this negative results still holds even if only one component of the β_0 and ϵ are unknown. This phenomenon, first pointed out by (Liu and Gao, 2019), illustrates the intrinsic challenges in the analysis of adaptation theory under Huber's contamination model.

4.3.1 Adaptation to the Smoothness Parameter Only

4.3.1.1 Bandwidth Selection in One-dimensional Case

Before we introduce our adaptive bandwidth selection procedure, let's start with a simple case when $d = 1$. Theorem 4.2.2 tells us that a kernel estimator $\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i \quad t)$ with bandwidth h can have the following convergence rate:

$$\sup_{f 2 P_{p,d}(\beta_0, L_0) \atop G} E_{P(\epsilon,f,G)} \, k \, \widehat{f}_h \quad f \, k_p \; . \; h^{\beta_0} + (nh)^{\ 1/(q\_2)} + \epsilon h^{\ 1/q}.$$

In order to minimize the above upper bound to achieve the optimal rate (4.2.1), the optimal bandwidth h should satisfy $h^{\ \beta_0} \quad (nh\ )^{\ 1/(q\_2)} + \epsilon h^{\ 1/q}$. Although β_0 is unknown to us, we can utilize the information of ϵ and choose h to be the largest h such that $h^{\beta_0} \; . \; (nh)^{\ 1/(q\_2)} + \epsilon h^{\ 1/q}$, intuitively. This intuition can be formalized by the following selection rule, known as Lepski's adaptive method (Lepskii, 1991):

$$\widehat{h} := \max\left\{ h \; 2 \; H : k\widehat{f}_h \quad \widehat{f}_l k_p \quad C_0 \left(\frac{1}{(nl)^{1/(q\_2)}} + \frac{\epsilon}{l^{1/q}} \right), 8h \quad l \; 2 \; H \right\},$$

where H is some (discrete) bandwidth collection set containing h and C_0 is a sufficiently large constant. The proof strategy of Lepski's adaptive method can be simply described in two stages: (1) we first show that the selected bandwidth $\widehat{h}$ is larger than the oracle bandwidth h with high probability; (2) then conditional on this event, by the definition of $\widehat{h}$, we have

$$E k \widehat{f}_{\widehat{h}} \quad f k_p \quad E k \widehat{f}_{\widehat{h}} \quad \widehat{f}_h \; k_p + E k \widehat{f}_h \quad f k_p \; . \; h^{\ \beta_0} + (nh\ )^{\ 1/(q\_2)} + \epsilon h^{\ 1/q}.$$

While Lepski's method is a smart way to select the optimal bandwidth without knowing β_0, this method essentially relies on the order topology of R, the set h belongs to. More specifically, the analysis critically utilizes the monotone property of each term in the convergence rate (i.e. the bias term h^{β_0} is increasing with h, while the variance term $(nh)^{\ 1/(q\_2)}$ and contamination term $\epsilon h^{\ 1/q}$ are decreasing). In contrast, this method can be questionable for the multi-dimensional case since in R^d the bias term now becomes $\sum_{j=1}^{d} h_j^{\beta_{0,j}}$ while the variance and contamination terms are monotonic in $V_h = \prod_{j=1}^{d} h_j$. Because R^d is not totally

ordered, there is no way to arrange the relative order of each component of the bandwidth h in advance when we do not know each $\beta_{0,j}$. This indicates the difficulty in the adaptation theory when considering a scale of the anisotropic density classes (i.e. β_0 is essentially a vector), compared to the isotropic density classes (i.e. β_0 is essentially a scalar). In the next section, we show how to overcome this difficulty by adopting the idea of the Goldenshluger-Lepski method (Goldenshluger and Lepski, 2011a).

4.3.1.2 Selection Procedure in Multi-dimensional Case

Intuition. Before we formally state our selection procedure, we would like to discuss some intuition behind it. We start with introducing a family of auxiliary estimators below, which play a key role in the Goldenshluger-Lepski method.

$$\widehat{f}_{h,l}(t) := \frac{1}{n}\sum_{i=1}^{n}[K_h \quad K_l](t \quad X_i), \quad \forall h, l \in \mathcal{H},$$

where " " stands for the convolution on $\mathbb{R}^d$, and $\mathcal{H}$ is a bandwidth collection set containing the optimal bandwidth h . A key observation (Goldenshluger and Lepski, 2011a) is that one can approximate the bias term of $\widehat{f}_h$ (i.e. $\sum_{j=1}^{d} h_j^{\beta_{0,j}}$) using that of $\widehat{f}_{h,l} \quad \widehat{f}_l$, as indicated by the following proposition.

Proposition 4.3.1. *For any $h \in \mathcal{H}$, we have*

$$\sup_{l\in\mathcal{H}} \|E_f[K_h \quad K_l](t \quad X) \quad E_f K_l(t \quad X)\|_p \quad \|K\|_1 \|E_f K_h(t \quad X) \quad f(t)\|_p ,$$

for any kernel K with $\|K\|_p \vee \|K\|_1 < 1$, and density f with $\|f\|_p < 1$.

Proposition 4.3.1 suggests that we may consider using $\widehat{R}_{h,p} := \sup_{l\in\mathcal{H}} \|\widehat{f}_{h,l} \quad \widehat{f}_l\|_p$ as a surrogate loss (as a proxy for the actual L_p loss $\|\widehat{f}_h \quad f\|_p$), and then select the bandwidth $\widehat{h}$ to be the minimizer of $\widehat{R}_{h,p}$ over $\mathcal{H}$. It turns out that in this way, only the bias term of $\widehat{f}_h$ is well approximated, while the variance and contamination term would be out of control. To fix this issue with contaminated observations, following the intuition of the Goldenshluger-Lepski method, we propose the following modification to $\widehat{R}_{h,p}$. Define the modified surrogate loss as

$$\widehat{R}^{(0)}_{h,p} := \sup_{l\in\mathcal{H}} \left\{ \|\widehat{f}_{h,l} \quad \widehat{f}_l\|_p \quad \left[\mathrm{Var}(\widehat{f}_{h,l}) + \mathrm{Var}(\widehat{f}_l)\right] \quad \left[\mathrm{C}(\widehat{f}_{h,l}) + \mathrm{C}(\widehat{f}_l)\right] \right\}_+ + \left[\mathrm{Var}(\widehat{f}_h) + \mathrm{C}(\widehat{f}_h)\right],$$

where $\mathrm{Var}(\widehat{f}_l)$ (resp. $\mathrm{Var}(\widehat{f}_{h,l})$), $\mathrm{C}(\widehat{f}_l)$ (resp. $\mathrm{C}(\widehat{f}_{h,l})$) represent the variance term and contamination term of $\widehat{f}_l$ (resp. $\widehat{f}_{h,l}$), that need to be specified later. The high-level intuition is that

$$\begin{aligned}\widehat{R}^{(0)}_{h,p} &\quad \sup_{l \in H}\Big[\mathrm{Bias}(\widehat{f}_{h,l}) \quad \mathrm{Bias}(\widehat{f}_l)\Big]_+ + \Big[\mathrm{Var}(\widehat{f}_h) + \mathrm{C}(\widehat{f}_h)\Big] \\ &\overset{(i)}{} \mathrm{Bias}(\widehat{f}_h) + \mathrm{Var}(\widehat{f}_h) + \mathrm{C}(\widehat{f}_h) \overset{(ii)}{} k\widehat{f}_h \quad fk_p.\end{aligned}$$

Here $\mathrm{Bias}(\widehat{f}_h)$, $\mathrm{Var}(\widehat{f}_h)$, $\mathrm{C}(\widehat{f}_h)$ represent the bias, variance and contamination terms of $\widehat{f}_h$ respectively. The intuition for the approximation (i) follows from Proposition 4.3.1, and (ii) follows from the decomposition of $k\widehat{f}_h \quad fk_p$ shown in (4.2.6).

Selection Rule. Now we formally state our bandwidth selection rule by specifying the variance term and contamination term of $\widehat{f}_l$ and $\widehat{f}_{h,l}$.

Define

$$\begin{aligned}\widehat{R}^{(1)}_{h,p} := \sup_{l \in H}\Big[\Big\|\widehat{f}_{h,l} \quad \widehat{f}_l\Big\|_p \quad 2m_p(h,l) \quad (2+128D_p)m_{\epsilon,p}(h,l)\Big]_+ \\ + 2m_p(h) + (2+128D_p)m_{\epsilon,p}(h),\end{aligned} \tag{4.3.1}$$

where $D_p = 0$ for $p \in [1,2]$ and $15p/(\log p)$ for $p \in (2, 1)$. Then the selected bandwidth $\widehat{h}$ is defined by

$$\widehat{h} := \arg\inf_{h \in H} \widehat{R}^{(1)}_{h,p}. \tag{4.3.2}$$

In the definition of (4.3.1), $m_p(h,l)$ and $m_p(h)$ are the variance terms. One can understand $m_p(h,l)$ as the surrogate of $\mathrm{Var}(\widehat{f}_{h,l}) + \mathrm{Var}(\widehat{f}_l)$, and $m_p(h)$ as the surrogate of $\mathrm{Var}(\widehat{f}_h)$. Similarly, $m_{\epsilon,p}(h,l)$ and $m_{\epsilon,p}(h)$ are the contamination terms, and one can treat $m_{\epsilon,p}(h,l)$ and $m_{\epsilon,p}(h)$ as the surrogates of $\mathrm{C}(\widehat{f}_{h,l}) + \mathrm{C}(\widehat{f}_l)$ and $\mathrm{C}(\widehat{f}_h)$ respectively. Now we give the formal definition of these terms.

- **Contamination terms:**

$$\begin{aligned}m_{\epsilon,p}(h,l) &:= (kKk_1 + 1)kKk_p f\epsilon V_l \quad {}^{1/q} + \epsilon(V_h \_ V_l) \quad {}^{1/q} g, \quad 8h,l \in H, \\ m_{\epsilon,p}(h) &:= \sup_{l \in H} m_{\epsilon,p}(l,h), \quad 8h \in H.\end{aligned}$$

- **Variance terms:**

$$m_p(h,l) := d_p(K_h \quad K_l) + d_p(K_l), \quad \forall h,l \in H,$$
$$m_p(h) := \sup_{l \in H} m_p(l,h), \quad \forall h \in H.$$

The function $d_s(U)$ (where $s \in [1, \infty)$, and U is a function $\mathbb{R}^d \to \mathbb{R}$) is defined by

$$d_s(U) := \begin{cases} r_s(U) & s \in [1,2], \\ \widehat{r}_s(U) & s \in (2, \infty), \end{cases}$$

where

$$r_s(U) := C_s n^{1/s - 1} \|U\|_s, s \in [1,2], \tag{4.3.3}$$

with $C_s = 128$ for $s \in [1,2)$ and $C_2 = 25/3$; for $s \in (2, \infty)$,

$$\begin{aligned} \widehat{r}_s(U) := 32 D_s \Bigg\{ n^{-1/2} \Bigg(\int \left[\frac{1}{n} \sum_{i=1}^{n} U^2(t \quad X_i) \right]^{s/2} dt \Bigg)^{1/s} \\ + 2n^{1/s - 1} \|U\|_s \Bigg\} - 32 n^{-1/2} \|U\|_2, \end{aligned} \tag{4.3.4}$$

with $D_s = 15s/\log s$.

Remark. The variance terms were proposed in (Goldenshluger and Lepski, 2011a), which considers $\widehat{r}_s(U)$ as an empirical counterpart of $r_s(U, p_X)$ for $s \in (2, \infty)$:

$$\begin{aligned} r_s(U, p_X) := 32 D_s \Bigg\{ n^{-1/2} \Bigg(\int \left[\int U^2(t \quad x) p_X(x) dx \right]^{s/2} dt \Bigg)^{1/s} \\ + 2n^{1/s - 1} \|U\|_s \Bigg\} - 32 n^{-1/2} \|U\|_2, \end{aligned} \tag{4.3.5}$$

where p_X is the density function of X_i under the setting that $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p_X$. Notice that this setting no longer holds with the presence of contamination. However, in our setting, we may consider $\widehat{r}_s(U)$ as a contaminated empirical version of $r_s(U, f)$.

Remark. The term $m_p(h,l)$ is called the majorant function for the variance term of $\widehat{f}_{h,l}$ and $\widehat{f}_l$. To be more specific, we define the bias and variance parts of the convolution kernel estimator $\frac{1}{m}\sum_{i=1}^{m}[K_h \quad K_l](x \quad X_i)$ based on $X_1, ..., X_m \overset{i.i.d.}{} p_X$ as the following (similar to (4.2.4)-(4.2.5))

$$B_{h,l}(p_X,t) := E_{p_X}[K_h \quad K_l](t \quad X) \quad p_X(t) = \int_{\mathbb{R}^d}[K_h \quad K_l](t \quad x)p_X(x)dx \quad p_X(t),$$

$$\xi_{h,l,m}(p_X,t) := \frac{1}{m}\sum_{i=1}^{m}[[K_h \quad K_l](t \quad X_i) \quad E_{p_X}[K_h \quad K_l](t \quad X)].$$

Under the setting where $X_1, \ldots, X_n \overset{i.i.d.}{} p_X$ without contamination, it is shown in (Goldenshluger and Lepski, 2011a) that (see also Lemma C.3.1 in the supplementary material)

$$E_{p_X} \sup_{l \in H}[\|\xi_{l,n}(p_X,t)\|_p \quad d_p(K_l)]_+ \quad \delta_{n,p}, \tag{4.3.6}$$

$$E_{p_X} \sup_{(h,l)\in H \times H}[\|\xi_{h,l,n}(p_X,t)\|_p \quad d_p(K_h \quad K_l)]_+ \quad \widetilde{\delta}_{n,p}, \tag{4.3.7}$$

where $\delta_{n,p}, \widetilde{\delta}_{n,p}$ are two small terms decaying exponentially with n, and thus are negligible compared to the terms in the rate (4.2.3). Therefore, one can see that $m_p(h,l)$ "majorates" $\|\xi_{h,l,n}(p_X,t)\|_p + \|\xi_{l,n}(p_X,t)\|_p$ uniformly. However, in our setting with the presence of contamination, (4.3.6) and (4.3.7) do not hold anymore, due to the following two reasons.

(1) Under Huber's contamination model (4.1.1), the data distribution may not have a density as the contamination distribution G can be arbitrary.

(2) For $p \in (2, \infty)$, $d_p(\ ) = \widehat{r}_p(\ )$, which, by its definition (4.3.4), relies on the whole data set and thus is also affected by contamination.

To get around issue (1), we treat pure observations and contaminated observations separately. In particular, under (4.1.2) conditioning on n_1, $X_1, \ldots, X_{n_1} \overset{i.i.d.}{} P_f$ and we consider applying (4.3.6) and (4.3.7) to $\|\xi_{h,l,n_1}(f,t)\|_p$ and $\|\xi_{l,n_1}(f,t)\|_p$, respectively. Regarding issue (2), we characterize the contamination effect by the term $m_{\epsilon,p}(h,l)$ and prove the following key fact in Lemma C.4.1

$$\bar{E}_{P(\epsilon,f,G)} \sup_{l \in H}\left[\|\xi_{l,n_1}(f,t)\|_p + \|\xi_{h,l,n_1}(f,t)\|_p \quad 2m_p(h,l) \quad 128D_p m_{\epsilon,p}(h,l)\right]_+ \ . \ \delta_{n,p} + \widetilde{\delta}_{n,p},$$

where $\bar{E}(\ ) = E\{(\ )\mathbf{1}(\frac{n-n_1}{n} < 2\epsilon)\}$ (noting that the event $\{\frac{n-n_1}{n} < 2\epsilon\}$ holds with high probability). This can be viewed as a robust version of (4.3.6) and (4.3.7) with the presence of contamination. Therefore, in our setting, $2m_p(h,l) + 128D_p m_{\epsilon,p}(h,l)$ is the majorant function for the variance terms $\|\xi_{h,l,n_1}(f,t)\|_p + \|\xi_{l,n_1}(f,t)\|_p$.

Remark. We can treat $m_{\epsilon,p}(h,l)$ as the majorant function for the contamination terms of $\widehat{f}_{h,l}$ and $\widehat{f}_l$, in the sense that the following inequality holds with high probability (see (C.4.7) for further details).

$$\frac{n-n_1}{n}\big(\ \|K_h - K_l\|_p + \|K_l\|_p\ \big) \quad 2m_{\epsilon,p}(h,l).$$

The left side of the above inequality arises from the presence of the contaminated data $\{X_{n_1+1},\ldots,X_n\}$. Similar to the term $\frac{n-n_1}{n}(\|K_h\|_p + \|f\|_p)$, which represents the contamination term in the decomposition of $\|\widehat{f}_h - f\|_p$ shown in (4.2.6), it characterizes the contamination term in the decomposition of $\|\widehat{f}_{h,l} - \widehat{f}_l\|$ in the surrogate loss $\widehat{R}^{(1)}_{h,p}$ (see (C.4.4)). The above inequality tells us that the contamination effect on the kernel estimators can be controlled by the contamination majorant $m_{\epsilon,p}(h,l)$. Coincidentally, the contamination effect on the variance majorant $m_p(h,l)$ ($d_p(\ )$ for $p \in (2, 1]$) can be also accounted for by this contamination majorant $m_{\epsilon,p}(h,l)$, as we discussed in the above remark.

Remark. The definitions of $r_s(U)$ and $r_s(U, p_X)$ are closely related to the following two inequalities.

$$E\left|\sum_{i=1}^m \xi_i\right|^p \quad 2\sum_{i=1}^m E|\xi_i|^p, \quad \forall p \in [1,2], \tag{4.3.8}$$

$$E\left|\sum_{i=1}^m \xi_i\right|^p \quad \left(\frac{15p}{\log p}\right)^p \max\left\{\left(\sum_{i=1}^m E\xi_i^2\right)^{p/2}, \sum_{i=1}^m E|\xi_i|^p\right\}, \quad \forall p \in (2, 1], \tag{4.3.9}$$

where ξ_i are centered independent random variables. The first inequality (Bahr-Esseen inequality) can be found in (von Bahr et al., 1965). The second one is a version of Rosenthal's inequality that can be found in Section 2 of (Masaon, 2009). The constant $15p/\log p$ is the best-known constant in the Rosenthal's inequality shown in (Johnson, 1985).

The above two inequalities are usually used to establish the upper bound for the variance term of kernel estimators. Under Huber's contamination model, we can use these two

inequalities to bound $\|\xi_{h,n_1}(f,t)\|_p$ in (4.2.5), by considering model (4.1.2). To see their connections with the definition of $r_s(U)$ and $r_s(U, p_X)$, just notice that under a general setting $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p_X$, for $p \in [1,2]$,

$$E_{p_X} \|\xi_{h,n}(f,t)\|_p \overset{\text{(Jensen)}}{\leq} \left(\int E_{p_X} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i(t) \right|^p dt \right)^{1/p} \overset{(4.3.8)}{\leq} Cn^{1/p - 1} \|K_h\|_p = C' r_p(K_h),$$

where $\xi_i(t) = K_h(t - X_i) - E_{p_X} K_h(t - X_i)$. Similarly, one can show that $E_{p_X} \|\xi_{h,n}(f,t)\|_p \leq Cr_p(K_h, p_X)$ for $p \in (2, \infty)$.

4.3.1.3 Theoretical Guarantees

Hereafter, we always consider the bandwidth collection of the form (if not specified otherwise) $H := \bigotimes_{j=1}^{d} [h_j^{\min}, h_j^{\max}]$. Denote $V_{\min} := \prod_{j=1}^{d} h_j^{\min}$, $V_{\max} := \prod_{j=1}^{d} h_j^{\max}$. The kernel estimator with the bandwidth selected by the rule (4.3.2) achieves the minimax rate (4.2.1), as shown in the following theorem.

Theorem 4.3.2. *Let $\widehat{f}_h$ be a kernel estimator with some kernel $K \in \mathcal{K}_\beta(L_K)$ being Lipschitz and compactly supported. Let the bandwidth $\widehat{h}$ be de ned in (4.3.2).*

(i) For $p \in [1,2)$, let $h_j^{\min} = 1/n, h_j^{\max} = 1$ for $j = 1, \ldots, d$. For any class $\mathcal{P}_{p,d}(\beta_0, L_0)$ such that $\beta_0 \geq \beta$, we have

$$\sup_{\substack{f \in \mathcal{P}_{p,d}(\beta_0, L_0) \\ G}} E_{P(\epsilon, f, G)} \|\widehat{f}_{\widehat{h}} - f\|_p \lesssim n^{-\frac{\beta_0}{q\beta_0 + 1}} \vee \epsilon^{\frac{q\beta_0}{q\beta_0 + 1}}.$$

(ii) For $p \in [2, \infty)$, we assume $1/\sqrt{n} \geq V_{\max} \geq c_1/(\log n)^{p/2}$, $V_{\min} \geq c_2/n$. Here c_1, c_2 are some constants depending on L_K, d, p only. We further assume that any $f \in \mathcal{P}_{p,d}(\beta_0, L_0)$ is uniformly bounded. Then for any $\epsilon \lesssim (\log n)^{-\frac{p(\beta_0 + 1) - 1}{2}}$, and any class $\mathcal{P}_{p,d}(\beta_0, L_0)$ such that $\beta_0 \geq \beta$, we have

$$\sup_{\substack{f \in \mathcal{P}_{p,d}(\beta_0, L_0) \\ G}} E_{P(\epsilon, f, G)} \|\widehat{f}_{\widehat{h}} - f\|_p \lesssim n^{-\frac{\beta_0}{2\beta_0 + 1}} \vee \epsilon^{\frac{q\beta_0}{q\beta_0 + 1}}.$$

Remark. (i) The optimal bandwidth h needs to be contained in $\mathcal{H}$ to make this adaptive method achieve the optimal minimax rate. One can check that $h \in \mathcal{H}$ is satisfied when we choose $\mathcal{H} = \left[\frac{1}{n}, 1\right]^d$ for $p \in [1,2)$. But for $p \in [2, \infty)$, noting that $V_h \asymp n^{-\frac{1}{(q-2)\beta_0+1}} \vee \epsilon^{\frac{q}{q\beta_0+1}}$, we need to further assume

$$\epsilon^{\frac{q}{q\beta_0+1}} \gtrsim (\log n)^{-p/2}, \quad i.e. \quad \epsilon \gtrsim (\log n)^{-\frac{p(\beta_0+1)}{2}}$$

to guarantee $h \in \mathcal{H}$. This is why we add this extra assumption in (ii) of Theorem 4.3.2.

(ii) The bandwidth collection can be replaced by any subset of $\mathcal{H}$ as long as it contains the optimal bandwidth h. For example, for $p \in [1,2)$, we may choose $\mathcal{H} = \{h = (h_1, \ldots, h_d) \in (0,1]^d : h_j = 2^{-k_j}, k_j = 0, \ldots, \log_2 n, j = 1, \ldots, d\}$ and the resulting kernel estimator is still adaptive minimax optimal.

(iii) For $p \in [2, \infty)$, we need the assumption that f is uniformly bounded, which is required for the proof of the Goldenshluger-Lepski method (Goldenshluger and Lepski, 2011a) even without contamination. Moreover, (Goldenshluger and Lepski, 2014) points out that the condition $\bar{\beta}_0 > 1/p$ implies that any $f \in \mathcal{N}_{p,d}(\beta_0, L_0)$ is uniformly bounded.

4.3.2 Adaptation to the Contamination Proportion Only

4.3.2.1 Bandwidth Selection

In this case, we assume the contamination proportion ϵ is unknown while the smoothness parameter $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,d})^\intercal$ is known. (If any one component of β_0 is also unknown, then there is no adaptive estimator as shown in Section 4.3.3.) Recall in the previous section, we introduce $\|\widehat{f}_{h,l} - \widehat{f}_l\|_p$ to approximate the bias of $\widehat{f}_h$. Since β_0 is given now, we can explicitly write down the exact bias bound. Therefore, using $\sup_{l \in \mathcal{H}} \|\widehat{f}_{h,l} - \widehat{f}_l\|_p$ in the surrogate loss would be meaningless. Instead, we consider $\sup_{l \in \mathcal{H}} \|\widehat{f}_{h,l} - \widehat{f}_h\|_p$ and define the surrogate loss as

$$\widehat{R}^{(2)}_{h,p} := \sup_{l \in \mathcal{H}} \left[\left\| \widehat{f}_{h,l} - \widehat{f}_h \right\|_p - m_b(l) \right]_+ + m_b(h), \tag{4.3.10}$$

and select the bandwidth $\widehat{h}$ by

$$\widehat{h} := \arg\inf_{h \in H} \widehat{R}^{(2)}_{h,p}. \tag{4.3.11}$$

In the definition of (4.3.10), $m_b(l)$ (resp. $m_b(h)$) serves as the majorant function for the bias term of $\widehat{f}_l$ (resp. $\widehat{f}_h$) and is defined as:

$$m_b(l) := \|K\|_1 C_L (2L_K)^d \sum_{j=1}^{d} l_j^{\beta_{0,j}}, \quad \forall l \in H,$$

where C_L is an arbitrary constant larger than L_0.

Remark. The intuition behind the selection procedure is that

$$\begin{aligned}
\widehat{R}^{(2)}_{h,p} &\quad \sup_{l \in H} \left[\left(\mathrm{Bias}(\widehat{f}_{h,l}) \quad \mathrm{Bias}(\widehat{f}_h) \right) \quad \mathrm{Bias}(\widehat{f}_l) + \mathrm{Var}(\widehat{f}_{h,l}) + \mathrm{Var}(\widehat{f}_h) + \mathrm{C}(\widehat{f}_{h,l}) + \mathrm{C}(\widehat{f}_h) \right]_+ \\
&\quad + \mathrm{Bias}(\widehat{f}_h) \\
&\overset{(i)}{\leq} \sup_{l \in H} \left[\mathrm{Var}(\widehat{f}_{h,l}) + \mathrm{Var}(\widehat{f}_h) + \mathrm{C}(\widehat{f}_{h,l}) + \mathrm{C}(\widehat{f}_h) \right] + \mathrm{Bias}(\widehat{f}_h) \\
&\overset{(ii)}{\leq} \mathrm{Var}(\widehat{f}_h) + \mathrm{C}(\widehat{f}_h) + \mathrm{Bias}(\widehat{f}_h) \quad \|\widehat{f}_h\| \quad \|f\|_p,
\end{aligned}$$

where our intuition for the approximation is based on Proposition 4.3.1, and the fact that we can view $K_h \quad K_l$ as a kernel with bandwidth $h \_ l$ in the sense that

$$\mathrm{Var}(\widehat{f}_{h,l}) \ . \ (n(V_h \_ V_l))^{\ 1/(q\_2)}, \quad \mathrm{C}(\widehat{f}_{h,l}) \ . \ \epsilon (V_h \_ V_l)^{\ 1/q}.$$

Theorem 4.3.3. *Given β_0, let $\widehat{f}_h$ be a kernel estimator with some kernel $K \in K_\beta(L_K)$ for some $\beta \quad \beta_0$. Assume K is Lipschitz and compactly supported. Let the bandwidth $\widehat{h}$ be defined in (4.3.11).*

(i) For $p \in [1,2)$, let $h_j^{\min} = 1/n, h_j^{\max} = 1$ for $j = 1, \ldots, d$. For any $\epsilon \quad 1/4$, we have

$$\sup_{f \in P_{p,d}(\beta_0, L_0)} E_{P(\epsilon, f, G)} \|f - \widehat{f}_{\widehat{h}}\|_p \ . \ n^{\ \frac{\beta_0}{q\beta_0+1}} \_ \epsilon^{\frac{q\beta_0}{q\beta_0+1}}. \tag{4.3.12}$$

(ii) For $p \in [2, 1)$, we assume $1/\sqrt{n} \quad V_{\min} \quad c_1/(\log n)^{p/2}$, $V_{\min} \quad c_2/n$. Here c_1, c_2 are some constants depending on L_K, d, p only. We further assume that $f \in P_{p,d}(\beta_0, L_0)$ is uniformly bounded. Then for any $\epsilon \ . \ (\log n)^{\ \frac{p(\beta_0+1)\ 1}{2}}$, we have

$$\sup_{f \in P_{p,d}(\beta_0, L_0)} E_{P(\epsilon, f, G)} \|f - \widehat{f}_{\widehat{h}}\|_p \ . \ n^{\ \frac{\beta_0}{2\beta_0+1}} \_ \epsilon^{\frac{q\beta_0}{q\beta_0+1}}. \tag{4.3.13}$$

4.3.2.2 An Alternative Approach Based on Lepski's Method

Given the smoothness parameter β_0, once we know the oracle scalar V_h , we immediately have the oracle $h = (V_h^{\beta_0/\beta_{0,1}}, \ldots, V_h^{\beta_0/\beta_{0,d}})'$. Therefore, the bandwidth selection in this case is similar to that in the one-dimensional setting as discussed in Section 4.3.1.1 since we only need to select V_h , a one-dimensional parameter. By Theorem 4.2.2 we know for any h with V_h 1,

$$E_{P(\epsilon,f,G)}\|\widehat{f}_h - f\|_p \asymp \sum_{j=1}^{d} h_j^{\beta_{0,j}} + \epsilon V_h^{-1/q} + (nV_h)^{-1/(q-2)}.$$

Intuitively, V_h should be the smallest V_h such that $\epsilon V_h^{-1/q} + (nV_h)^{-1/(q-2)} \asymp \sum_{j=1}^{d} h_j^{\beta_{0,j}}$ holds. Therefore, we can adopt Lepski's method for one-dimensional setting to get the oracle V_h and then h . More specifically, define

$$H = \left\{ \left(V_h^{\frac{\beta_0}{\beta_{0,1}}}, \ldots, V_h^{\frac{\beta_0}{\beta_{0,d}}} \right)' \middle| V_h = \frac{1}{2^k}, k = [\log_2 V_{\min}, \log_2 V_{\max}] \cap \mathbb{N} \right\}, \tag{4.3.14}$$

where $V_{\max} = 1, V_{\min} = 1/n$ for $p \in [1,2)$; $V_{\max} = c_1/(\log n)^{p/2}, V_{\min} = c_2/n$ for $p \in [2, \infty)$. Here c_1, c_2 are some constants depending on L_K, d, p only. Finally, we obtain the bandwidth as

$$\widehat{h} := \min \left\{ h \in H : \|\widehat{f}_h - \widehat{f}_l\|_p \le \frac{c_0}{d} \sum_{j=1}^{d} l_j^{\beta_{0,j}} = c_0 V_l^{\beta_0}, \forall l \le h, l \in H \right\}, \tag{4.3.15}$$

where $l \le h$ is equivalent to $V_l \ge V_h$, and c_0 is a constant depending on L_0, L_K, p, d.

Theorem 4.3.4. *Let $\widehat{f}_h$ be a kernel estimator with some kernel $K \in \mathcal{K}_\beta(L_K)$ being Lipschitz and compactly supported, for some $\beta \ge \beta_0$. Let the bandwidth $\widehat{h}$ be defined in (4.3.15) and H be defined in (4.3.14). Then (i) (4.3.12) holds for any $\epsilon \ge 1/4$; and (ii) (4.3.13) holds for any $\epsilon \gtrsim (\log n)^{-\frac{p(\beta_0+1)-1}{2}}$, assuming that $f \in \mathcal{P}_{p,d}(\beta_0, L_0)$ is uniformly bounded.*

Remark. To analyze the Lepski's method, it is central to establish certain concentration inequality for the variance term $\|\xi_{h,n}(f,t)\|_p$ defined in (4.2.5). We summarize a non-robust version of this result in our Lemma C.3.1, following Lemmas 1 and 2 of (Goldenshluger and Lepski, 2011a). However, we cannot directly apply Lemma C.3.1 in our setting as for $p \geq 2$ $(2, \infty)$ it involves $\widehat{r}_p(K_h)$, which uses the whole data set including the contaminated part, by its definition (4.3.4). Unlike the case for the adaptation to β_0, we cannot use the contamination majorant $m_{\epsilon,p}(h,l)$ to control the contamination effect now. Instead, we develop an innovative induction method to get around this issue. The detailed proof is given in Section C.5 of the supplementary material.

4.3.3 Adaptation to Both the Smoothness Parameter and the Contamination Proportion

In the most general case where neither smoothness parameter β_0 nor contamination proportion ϵ is accessible, adaptive estimation is a much harder task. Inspired by the negative result in (Liu and Gao, 2019), in this section we demonstrate that it is impossible to construct a general rate adaptive estimator for our model. To this end, we start with the following definition which formulates the meaning of being rate adaptive in our specific setting.

Definition 4.3.1. Given two positive functions $R_1(\,), R_2(\,) : \mathbb{R}^d \to \mathbb{R}^+$, an estimator $\widehat{f}$ is called $(R_1(\,), R_2(\,))$-rate adaptive with respect to the parameter set $\Theta = \{\theta_1, \ldots, \theta_m\}$ $\Theta_0 = \{\beta_{0,1}, \ldots, \beta_{0,d}, \epsilon\}$, if there exist some constants $C_0, C_1, \ldots, C_m$ such that for any $n \geq 1$, any $\theta_j \leq C_j, j = 1, \ldots, m$, we have

$$\sup_{\substack{f \in \mathcal{P}_{p,d}(\beta_0, L_0) \\ G}} E_{P(\epsilon,f,G)} \|\widehat{f} - f\|_p \geq C_0\big(n^{-R_1(\beta_0)} - \epsilon^{R_2(\beta_0)}\big).$$

For the minimax rate with $R_1(\beta_0) = \bar{\beta}_0/[(q-2)\bar{\beta}_0+1]$, $R_2(\beta_0) = \bar{\beta}_0/(q\bar{\beta}_0+1)$, it is shown in Section 4.3.1 that there exists some estimator that is $(R_1(\,), R_2(\,))$-rate adaptive with respect to $\{\beta_{0,1}, \beta_{0,2}, \ldots, \beta_{0,d}\}$. In addition, Section 4.3.2 reveals the existence of $(R_1(\,), R_2(\,))$-rate adaptive estimator with respect to $\{\epsilon\}$. However, in searching for rate adaptive estimators with respect to a parameter set containing both ϵ and some component

of β_0, the following theorem shows that such a goal is impossible for any given rate functions $R_1(\,)$ and $R_2(\,)$.

Theorem 4.3.5. *For any two positive functions $R_1(\,)$ and $R_2(\,)$, and any $j = 1, \ldots, d$, there is no estimator that is $(R_1(\,), R_2(\,))$-rate adaptive with respect to $\{\beta_{0,j}, \epsilon\}$. Consequently, there is no estimator that is $(R_1(\,), R_2(\,))$-rate adaptive with respect to $\Theta_0 = \{\beta_{0,1}, \beta_{0,2}, \ldots, \beta_{0,d}, \epsilon\}$.*

Remark. The above theorem says that when both the contamination proportion and the smoothness parameter (even just one component) are unknown, there is no rate adaptive estimator for any target convergence rate of the form $n^{-R_1(\beta_0)}\epsilon^{R_2(\beta_0)}$. This is stronger than the well-known negative results for the classical adaptive estimation under pointwise loss. The classical adaptive estimation theory (Brown and Low, 1996) states that no estimator can adaptively achieve the minimax rate under pointwise loss over a scale of Hölder classes. Nevertheless, some estimators (Lepski and Spokoiny, 1997) can still adaptively achieve the minimax rate up to a logarithm factor.

Remark. The result of Theorem 4.3.5 is based on the fact that Huber's contamination model has no constraint on the contamination distribution, and thus it is possible to represent the same distribution in two classes indexed by two quite different sets of parameters (β_0, ϵ) and $(\widetilde{\beta}_0, \widetilde{\epsilon})$. To be more specific, we may construct a distribution $P(\epsilon, P_f, G) = P(\widetilde{\epsilon}, \widetilde{P}_f, \widetilde{G})$ with $P_f \in \mathcal{P}_{p,d}(\beta_0, L_0)$, $\widetilde{P}_f \in \mathcal{P}_{p,d}(\widetilde{\beta}_0, L_0)$. This specific identifiability issue of these two sets of parameters under Huber's contamination model makes the adaptive estimation to both parameters β_0 and ϵ impossible. See Section C.6 in the supplementary material for complete proof.

4.4 Minimax Rate with Structured Contamination

In this section, we assume the contamination distribution G has a density g with a finite L_p norm. That is, $g \in \mathcal{L}_{p,d}(L_1)$ for some L_1, where $\mathcal{L}_{p,d}(L_1) := \{f : \mathbb{R}^d \to \mathbb{R} | f \geq 0, \int f = 1, \|f\|_p \leq L_1\}$. With this mild additional restriction on G, the minimax rate becomes quite

different from (4.2.1).

Theorem 4.4.1. *Under the above setting, we have the following minimax rate:*

$$\inf_{\widehat{f}} \sup_{\substack{f \in P_{p,d}(\beta_0, L_0) \\ g \in L_{p,d}(L_1)}} E_{p(\epsilon,f,g)} \| \widehat{f} - f \|_p^p \asymp \begin{cases} n^{-\frac{\beta_0}{q\beta_0+1}} - \epsilon, & 1 \le p \le 2 \\ n^{-\frac{\beta_0}{2\beta_0+1}} - \epsilon, & p > 2, \end{cases} \tag{4.4.1}$$

where $p(\epsilon, f, g)$ is used to denote the density $(1-\epsilon)f + \epsilon g$.

4.4.1 Upper Bound

The minimax rate (4.4.1) can be achieved by a kernel density estimator $\widehat{f}_h$ with some $K \in \mathcal{K}_{\beta_0}(L_K)$. We show that

$$\mathrm{Bias}(\widehat{f}_h) := \| E_{p_\epsilon} \widehat{f}_h - f \|_p \lesssim \sum_{j=1}^{d} h_j^{\beta_{0,j}} + \epsilon,$$

$$\mathrm{Var}(\widehat{f}_h) := E_{p_\epsilon} \| \widehat{f}_h - E_{p_\epsilon} \widehat{f}_h \|_p^p \lesssim (nV_h)^{-1/(q-2)},$$

where p_ϵ represents $p(\epsilon, f, g)$. Choosing $h_j \asymp n^{-\frac{\beta_0}{\beta_{0,j}((q-2)\beta_0+1)}}$, we obtain the minimax rate (4.4.1).

Remark. Notice that in this case with structured contamination, we can utilize the smoothness of the contamination distribution G, and treat the contaminated distribution as a whole with a single density p_ϵ. In contrast, under Huber's contamination model, we separate the contaminated distribution into the clean part and the contaminated part, as shown in (4.2.6).

4.4.2 Lower Bound

Regarding the lower bound of (4.4.1), we focus on the second term ϵ and consider

$$\widetilde{f}(x) = f(x) + \epsilon\gamma\phi(x),$$

$$\widetilde{g}(x) = g(x) - (1-\epsilon)\gamma\phi(x).$$

Here, we choose some functions $f \in P_{p,d}(\beta_0, L_0/2)$, $g \in L_{p,d}(L_1/2)$, and the function ϕ is infinitely differentiable satisfying $\int \phi = 0$. If γ is sufficiently small, we also have $\widetilde{f} \in P_{p,d}(\beta_0, L_0)$, $\widetilde{g} \in L_{p,d}(L_1)$. Since it is impossible to distinguish between these two densities $(1-\epsilon)f + \epsilon g$ and $(1-\epsilon)\widetilde{f} + \epsilon\widetilde{g}$, an error of order $\| \widetilde{f} - f \|_p \asymp \epsilon$ cannot be avoided.

4.4.3 Extension to Smooth Contamination Density

For the structured contamination case, a common assumption on the contamination distribution G is that its density g is in a smooth function class similarly defined as the true density f, but with a possibly different smoothness parameter. That is, we assume $g \in \mathcal{P}_{p,d}(\beta_1, L_1)$ for some β_1 and L_1. The following theorem reveals a surprising result that the minimax rate does not depend on β_1 and remains the same as the rate (4.4.1).

Theorem 4.4.2. *Assume $g \in \mathcal{P}_{p,d}(\beta_1, L_1)$, we have the following minimax rate:*

$$\inf_{\widehat{f}} \sup_{\substack{f \in \mathcal{P}_{p,d}(\beta_0, L_0) \\ g \in \mathcal{P}_{p,d}(\beta_1, L_1)}} E_{p(\epsilon,f,g)} \|\widehat{f} - f\|_p \asymp \begin{cases} n^{-\frac{\beta_0}{q\beta_0+1}} \vee \epsilon, & 1 \le p < 2 \\ n^{-\frac{\beta_0}{2\beta_0+1}} \vee \epsilon, & p > 2. \end{cases} \tag{4.4.2}$$

Remark. By comparing three minimax rates (4.2.1), (4.4.1) and (4.4.2), the difference between structured and unstructured (i.e. arbitrary) contamination cases is essentially determined by the existence of the density of contamination distribution. In particular, the smoothness of the contamination density makes no difference with respect to the minimax rate.

4.5 Adaptation with Structured Contamination

As there is little difference in the estimation results between the two assumptions on the contamination density g in the last section, we just take $g \in \mathcal{L}_{p,d}(L_1)$ as an example to discuss the adaptive method. One can check that the procedure can be applied to the case where $g \in \mathcal{P}_{p,d}(\beta_1, L_1)$ without any modification.

To achieve the minimax rate (4.4.1), we need to select the bandwidth $h_j \asymp n^{-\frac{\beta_0}{\beta_{0,j}((q-2)\beta_0+1)}}$, which only requires the prior knowledge of the smoothness parameter β_0. Notice that the choice of bandwidth does not depend on the contamination proportion ϵ; hence it is not surprising to see the classical adaptive procedure of the non-robust Goldenshluger-Lepski method (Goldenshluger and Lepski, 2011a) works in our case too. With a little modification of the proof in (Goldenshluger and Lepski, 2011a), we can show that this procedure leads to

a minimax estimator. For completeness, we briefly describe the adaptive selection procedure below.

We select the bandwidth by

$$\widehat{h} = \arg\inf_{h \in H} \widehat{R}_{h,p}, \quad \text{where} \quad \widehat{R}_{h,p} := \sup_{l \in H} [\| \widehat{f}_{h,l} - \widehat{f}_l \|_p - m_p(h,l)]_+ + m_p(h), \tag{4.5.1}$$

where $m_p(h,l)$ and $m_p(h)$ are defined in Section 4.3.1. We summarize the main results in the following theorem.

Theorem 4.5.1. *Let $\widehat{f}_h$ be a kernel estimator with some kernel $K \in K_\beta(L_K)$ being Lipschitz and compactly supported. Let the bandwidth $\widehat{h}$ be defined in (4.5.1).*

(i) For $p \in [1,2)$, let $h_j^{\min} = 1/n, h_j^{\max} = 1$ for $j = 1, \ldots, d$. For any class $P_{p,d}(\beta_0, L_0)$ with $\beta_0 \geq \beta$, we have

$$\sup_{\substack{f \in P_{p,d}(\beta_0, L_0) \\ g \in L_{p,d}(L_1)}} E_{p(\epsilon,f,g)} \| \widehat{f}_{\widehat{h}} - f \|_p \cdot n^{\frac{\beta_0}{q\beta_0+1}} \lesssim \epsilon.$$

(ii) For $p \in [2, 1)$, we assume $1/\sqrt{n} \geq V_{\max} \geq c_1/(\log n)^{p/2}$, $V_{\min} \leq c_2/n$. Here c_1, c_2 are some constants depending on L_K, d, p only. We further assume that f, g are both uniformly bounded. Then for any class $P_{p,d}(\beta_0, L_0)$ with $\beta_0 \geq \beta$, we have

$$\sup_{\substack{f \in P_{p,d}(\beta_0, L_0) \\ g \in L_{p,d}(L_1)}} E_{p(\epsilon,f,g)} \| \widehat{f}_{\widehat{h}} - f \|_p \cdot n^{\frac{\beta_0}{2\beta_0+1}} \lesssim \epsilon.$$

Remark. (i) In the above theorem, we assume the contamination density g is uniformly bounded for $p \in [2, 1)$. It is due to the requirement of the Goldenshluger-Lepski method. Notice that in Section 4.3 under Huber's contamination model, we only adopt the Goldenshluger-Lepski method to the "clean" data, while for the structured contamination case, we apply it to the whole data set. This is why we only need to assume f is uniformly bounded under Huber's contamination model, but need to assume both f and g are uniformly bounded for the structured contamination case.

(ii) Noticeably, the lower bound of the minimax rate (4.4.1) still holds with this additional boundedness assumption. In fact, one can choose both g_0 and ϕ uniformly bounded in the proof of Theorem 4.4.1.

4.6 Discussion: An extension to the Pointwise Loss

In this work, we constrain our focus on L_p loss functions, establish the minimax rates, and develop new adaptive bandwidth selection procedures under Huber's contamination model. It is worthwhile to point out that this framework can be naturally extended to the pointwise loss setting:

$$L(\widehat{f}, f) = \mathsf{j}\widehat{f}(x_0) \quad f(x_0)\mathsf{j}, \quad \text{for some } x_0 \mathsf{2} \mathsf{R}^d,$$

with f being assumed in a general anisotropic smooth function class with parameter $\beta_0 \mathsf{2} \mathsf{R}^d$. We briefly discuss here how we can apply the ideas and techniques developed in this work and (Goldenshluger and Lepski, 2014) to construct bandwidth selection procedures adaptive to β_0 (one can develop adaptive methods with respect to ϵ similarly), and get the pointwise oracle inequality under Huber's contamination model.

We still need to introduce the convolution auxiliary estimator to approximate the bias. For the variance terms, instead of using the majorant functions $d_p(\ )$ and $m_p(\ ,\ )$, we define new majorant functions $D(\ )$ and $M(\ )$, which were considered in (Goldenshluger and Lepski, 2014):

$$D_h(x) = \sqrt{\frac{xA(K_h, x)\log n}{nV_h}} + \frac{c\log n}{nV_h}, \quad h \mathsf{2} \mathsf{H},$$

where $c > 0$ is a constant to be specified, and

$$A(K_h, x) = \int \mathsf{j}K_h(x \quad t)\mathsf{j}\, f(t)\mathrm{d}t.$$

Similarly, define

$$D_{h,l}(x) = \sqrt{\frac{xA(K_h \quad K_l, x)\log n}{n(V_h \_ V_l)}} + \frac{c\log n}{n(V_h \_ V_l)}, \quad h, l \mathsf{2} \mathsf{H}.$$

As $D_h(x)$ and $D_{h,l}(x)$ rely on the unknown density f, we also introduce the empirical counterparts $\widehat{D}_h(x)$ and $\widehat{D}_{h,l}(x)$ via replacing $A(U,x)$ by $\widehat{A}(U,x) := \frac{1}{n}\sum_{i=1}^n \mathsf{j}U(x \quad X_i)\mathsf{j}$ (for $U = K_h$ or $K_h \quad K_l$). At last, we define

$$M_{h,l}(x) = D_{h,l}(x) + D_l(x), \quad M_h(x) = \sup_{l \mathsf{2} H} M_{l,h}(x);$$

$$\widehat{M}_{h,l}(x) = \widehat{D}_{h,l}(x) + \widehat{D}_l(x), \quad \widehat{M}_h(x) = \sup_{l \mathsf{2} H} \widehat{M}_{l,h}(x).$$

In (Goldenshluger and Lepski, 2014), the authors have shown that without contamination, if we pick the bandwidth h by

$$\widehat{h} = \arg\inf_{h \in H} \widehat{R}_h(x), \quad \text{where} \quad \widehat{R}_h(x) := \sup_{l \in H} [\, | \widehat{f}_{h,l}(x) - \widehat{f}_l(x) | - \widehat{M}_{h,l}(x)]_+ + \widehat{M}_h(x),$$

then one can get the pointwise oracle inequality below

$$| \widehat{f}_{\widehat{h}}(x) - f(x) | \le C \inf_{h \in H} \| f \| B_h(f, \cdot) \|_1 + M_h(x) \} + \zeta(x), \quad \forall x \in \mathbb{R}^d,$$

where the remainder $\zeta(x) > 0$ is small in the sense that

$$E_f \zeta(x) \le C n^{-1/2}, \ \forall x \in \mathbb{R}^d, \quad \text{and} \quad \int_{\mathbb{R}^d} E_f \zeta(x) dx \le C n^{-1/2}.$$

Notice that in the above selection rule, $\widehat{M}_{h,l}(x)$ and $\widehat{M}_h(x)$ represent the majorant functions of the variance terms $\mathrm{Var}(\widehat{f}_{h,l}) + \mathrm{Var}(\widehat{f}_l)$ and $\mathrm{Var}(\widehat{f}_h)$, respectively. Under Huber's contamination model, we introduce the following majorant functions for the contamination terms:

$$\begin{aligned} M_{\epsilon,h,l}(x) &:= C(\|K\|_1 + 1) \| \epsilon V_l \|_1 + \epsilon (V_h - V_l) \|_1, \\ M_{\epsilon,h}(x) &:= \sup_{l \in H} M_{\epsilon,l,h}(x), \end{aligned}$$

when ϵ is known and the smooth parameter β_0 is unknown. We believe that similar to (4.3.1), a surrogate loss function given in the following form:

$$\widehat{R}_h^{(1)}(x) := \sup_{l \in H} \Big[\Big| \widehat{f}_{h,l}(x) - \widehat{f}_l(x) \Big| - 2\widehat{M}_{h,l}(x) - M_{\epsilon,h,l}(x) \Big]_+ + 2\widehat{M}_h(x) + M_{\epsilon,h}(x),$$

with a selected bandwidth $\widehat{h} = \arg\inf_{h \in H} \widehat{R}_h^{(1)}(x)$ would lead to some optimal estimator adaptive to smoothness parameter β_0 given ϵ. (e.g. minimax in some anisotropic Hölder class.) Similar modifications can be made for the adaptive procedure with respect to ϵ, given β_0.

Appendix A Supplement to Chapter 2

A.1 Examples of Robust Loss Functions

We list some smoothed Huber loss and pseudo Huber loss functions that satisfy Condition 2.2.1.

1. (Pseudo-Huber loss I): $\ell(x) = \sqrt{1+x^2} \quad 1$, $x \in \mathbb{R}$. By direct calculations,

$$\ell'(x) = \frac{x}{\sqrt{1+x^2}} \quad \text{and} \quad \ell''(x) = \frac{1}{(1+x^2)^{3/2}}.$$

   Then Condition 2.2.1 holds with $c_1 = 1$, $c_2 = (1+c^2)^{3/2}$ and $c_3 = c$ for any $c > 0$.

2. (Pseudo-Huber loss II): $\ell(x) = \log[(e^x + e^{-x})/2]$, $x \in \mathbb{R}$. The first and second derivatives are, respectively,

$$\ell'(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad \text{and} \quad \ell''(x) = \frac{4}{(e^x + e^{-x})^2}.$$

   Hence, Condition 2.2.1 holds with $c_1 = 1$, $c_2 = 4(e^c + e^{-c})^{-2}$ and $c_3 = c$ for any $c > 0$.

3. (Smoothed Huber loss I):

$$\ell(x) = \begin{cases} x^2/2 - |x|^3/6 & \text{if } |x| \quad 1, \\ |x|/2 - 1/6 & \text{if } |x| > 1. \end{cases}$$

   The first and second derivatives are

$$\ell'(x) = \begin{cases} x - \text{sign}(x) \quad x^2/2 & \text{if } |x| \quad 1, \\ \text{sign}(x)/2 & \text{if } |x| > 1, \end{cases} \qquad \ell''(x) = \begin{cases} 1 - |x| & \text{if } |x| \quad 1, \\ 0 & \text{if } |x| > 1. \end{cases}$$

   Condition 2.2.1 is then satisfied with $c_1 = 1/2$, $c_2 = 1 - c$ and $c_3 = c$ for any $0 < c < 1$.

4. (Smoothed Huber loss II):

$$\ell(x) = \begin{cases} x^2/2 & x^4/24 \qquad \text{if } |x| \quad \sqrt{2}, \\ (2\sqrt{2}/3)|x| & 1/2 \quad \text{if } |x| > \sqrt{2}. \end{cases}$$

The first and second derivatives are

$$\ell'(x) = \begin{cases} x & x^3/6 \qquad \text{if } |x| \quad \sqrt{2}, \\ (2\sqrt{2}/3)\,\mathrm{sign}(x) & \text{if } |x| > \sqrt{2}, \end{cases} \quad \ell''(x) = \begin{cases} 1 & x^2/2 \quad \text{if } |x| \quad \sqrt{2}, \\ 0 & \text{if } |x| > \sqrt{2}. \end{cases}$$

Condition 2.2.1 holds with $c_1 = 2\sqrt{2}/3$ and $c_2 = 1 \quad c^2/2$ and $c_3 = c$ for any $0 < c < \sqrt{2}$.

A.2 Proofs

A.2.1 Proof of Proposition 2.2.1

Let $\widetilde{\boldsymbol{\theta}} := (\beta_0 + \alpha_\tau, \boldsymbol{\beta}^\top)^\top$. It suffices to show that $\widetilde{\boldsymbol{\theta}}$ is the unique minimizer of the function $L(\boldsymbol{\theta}) := \mathsf{E}_{\boldsymbol{x},\varepsilon \ F}[\ell_\tau(y \quad \bar{\boldsymbol{x}}^\top\boldsymbol{\theta})]$. For any $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\top)^\top$, we know

$$L(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{x}}\mathsf{E}_{\varepsilon \ F}\ell_\tau(\varepsilon + (\beta_0 \quad \beta_0) + \boldsymbol{x}^\top(\boldsymbol{\beta} \quad \boldsymbol{\beta}))$$
$$\mathsf{E}_{\boldsymbol{x}}\mathsf{E}_{\varepsilon \ F}\ell_\tau(\varepsilon \quad \alpha_\tau) = L(\widetilde{\boldsymbol{\theta}})$$

The inequality is due to the fact that α_τ is the unique minimizer of the function $\alpha \mapsto \mathsf{E}_{\varepsilon \ F}[\ell_\tau(\varepsilon \quad \alpha)]$. This indicates that $\widetilde{\boldsymbol{\theta}}$ is a minimizer of $L(\boldsymbol{\theta})$. Notice that $\nabla^2 L(\widetilde{\boldsymbol{\theta}}) = \mathsf{E}_{\varepsilon \ F}\ell''_\tau(\varepsilon \quad \alpha_\tau)\mathbf{S}$ is positive definite as $\mathsf{E}_{\varepsilon \ F}\ell''_\tau(\varepsilon \quad \alpha_\tau) \quad c_2\mathsf{P}_{\varepsilon \ F}(|\varepsilon \quad \alpha_\tau| \quad c_3\tau) > 0$. This fact and the convexity of $L(\boldsymbol{\theta})$ imply that $\widetilde{\boldsymbol{\theta}}$ is the unique minimizer of $L(\boldsymbol{\theta})$. Then by definition, $\boldsymbol{\theta}_\tau = \widetilde{\boldsymbol{\theta}} = (\beta_0 + \alpha_\tau, \boldsymbol{\beta}^\top)^\top$. □

A.2.2 Proof of Theorem 2.2.1

Given $r > 0$, define the local parameter set $\Theta_r = \{\theta \in \mathbb{R}^{d+1} : \|\theta - \theta_\tau\|_{\mathbf{S}} \le r\}$. For any prespecified $r > 0$, we can always find an intermediate estimator $\widehat{\theta}_{\tau,\eta} = \theta_\tau + \eta(\widehat{\theta}_\tau - \theta_\tau)$ for some $\eta \in [0,1]$ such that $\|\widehat{\theta}_{\tau,\eta} - \theta_\tau\|_{\mathbf{S}} \le r$. In fact, if $\widehat{\theta}_\tau \in \Theta_r$, we simply take $\eta = 1$; otherwise, we let $\eta = r/\|\widehat{\theta}_\tau - \theta_\tau\|_{\mathbf{S}}$ such that $\|\widehat{\theta}_{\tau,\eta} - \theta_\tau\|_{\mathbf{S}} = r$. As $L_\tau(\theta)$ is convex in θ, by Lemma F.2 in Fan et al. (2018), we have

$$\begin{aligned}\left\langle \nabla L_\tau(\widehat{\theta}_{\tau,\eta}) - \nabla L_\tau(\theta_\tau), \widehat{\theta}_{\tau,\eta} - \theta_\tau \right\rangle &\le \eta \left\langle \nabla L_\tau(\widehat{\theta}_\tau) - \nabla L_\tau(\theta_\tau), \widehat{\theta}_\tau - \theta_\tau \right\rangle \\ &= \eta \left\langle \nabla L_\tau(\theta_\tau), \widehat{\theta}_\tau - \theta_\tau \right\rangle \\ &\le \|\mathbf{S}^{-1/2} \nabla L_\tau(\theta_\tau)\|_2 \left\|\widehat{\theta}_{\tau,\eta} - \theta_\tau\right\|_{\mathbf{S}}\end{aligned} \tag{A.2.1}$$

The proof then boils down to establish an upper bound on

$$\|\mathbf{S}^{-1/2} \nabla L_\tau(\theta_\tau)\|_2 = \left\| \frac{1}{n} \sum_{i=1}^n \ell'_\tau(\varepsilon_i - \alpha_\tau) \boldsymbol{z}_i \right\|_2 \tag{A.2.2}$$

and a lower bound for

$$\inf_{\theta \in \Theta_r} \langle \nabla L_\tau(\theta) - \nabla L_\tau(\theta_\tau), \theta - \theta_\tau \rangle / \|\theta - \theta_\tau\|_{\mathbf{S}}^2 \tag{A.2.3}$$

uniformly over $\theta \in \Theta_r$, where $\boldsymbol{z}_i := \mathbf{S}^{-1/2} \bar{\boldsymbol{x}}_i$ are isotropic random vectors.

First we bound (A.2.2) from above. Let $\xi_i = \ell'_\tau(\varepsilon_i - \alpha_\tau)$, $i = 1, \ldots, n$, be $i.i.d.$ random variables from $\xi := \ell'_\tau(\varepsilon - \alpha_\tau)$. Since α_τ is the unique minimizer of $\alpha \mapsto \mathsf{E}_{\varepsilon \sim F} \rho_\tau(\varepsilon - \alpha)$, we have $\mathsf{E}_{\varepsilon \sim F} \psi_\tau(\varepsilon - \alpha_\tau) = 0$. We first decompose $\|\mathbf{S}^{-1/2} \nabla L_\tau(\theta_\tau)\|_2$ into two parts:

$$\|\mathbf{S}^{-1/2} \nabla L_\tau(\theta_\tau)\|_2 \le \left\| \frac{1}{n} \sum_{i=1}^n \xi_i \boldsymbol{z}_i - \mathsf{E}(\xi_i \boldsymbol{z}_i) \right\|_2 + \|\mathsf{E}(\xi \boldsymbol{z})\|_2.$$

Under the error-contamination model that $\varepsilon|\boldsymbol{x} \sim (1-\epsilon)F + \epsilon G_{\boldsymbol{x}}$, noticing that F is independent of $\boldsymbol{x}$, we have

$$\begin{aligned}
\|\mathsf{E}(\xi\boldsymbol{z})\|_2 &= \|\mathsf{E}\,\mathsf{E}\,[(\xi\boldsymbol{z})|\boldsymbol{x}]\|_2 \\
&= \left\|\mathsf{E}\left\{\boldsymbol{z}(1-\epsilon)\mathsf{E}_{\varepsilon|\boldsymbol{x}\ F}[\ell_\tau^0(\varepsilon-\alpha_\tau)|\boldsymbol{x}] + \boldsymbol{z}\epsilon\mathsf{E}_{\varepsilon|\boldsymbol{x}\ G_{\boldsymbol{x}}}[\ell_\tau^0(\varepsilon-\alpha_\tau)|\boldsymbol{x}]\right\}\right\|_2 \\
&= \left\|\mathsf{E}\left\{\boldsymbol{z}\epsilon\mathsf{E}_{\varepsilon|\boldsymbol{x}\ G_{\boldsymbol{x}}}[\ell_\tau^0(\varepsilon-\alpha_\tau)|\boldsymbol{x}]\right\}\right\|_2 \\
&= \sup_{\boldsymbol{u}\in\mathsf{S}^d} \mathsf{E}\left\{\boldsymbol{u}^\top\boldsymbol{z}\epsilon\mathsf{E}_{\varepsilon|\boldsymbol{x}\ G_{\boldsymbol{x}}}[\ell_\tau^0(\varepsilon-\alpha_\tau)|\boldsymbol{x}]\right\} \\
&\quad \epsilon c_1\tau \sup_{\boldsymbol{u}\in\mathsf{S}^d} \mathsf{E}|\boldsymbol{u}^\top\boldsymbol{z}| \quad \epsilon c_1\tau \sup_{\boldsymbol{u}\in\mathsf{S}^d} \left(\mathsf{E}|\boldsymbol{u}^\top\boldsymbol{z}|^2\right)^{1/2} = \epsilon c_1\tau.
\end{aligned}$$

Denote $\boldsymbol{\gamma} := \frac{1}{n}\sum_{i=1}^n \{\xi_i\boldsymbol{z}_i - \mathsf{E}(\xi_i\boldsymbol{z}_i)\}$. To bound $\|\boldsymbol{\gamma}\|_2 = \sup_{\boldsymbol{u}\in\mathsf{S}^d}\boldsymbol{u}^\top\boldsymbol{\gamma}$, by a standard covering argument, we can find a $(1/2)$-net $\mathcal{N}_{1/2}$ of S^d with $\left|\mathcal{N}_{1/2}\right| \le 5^{d+1}$ such that $\|\gamma\|_2 \le 2\max_{u\in\mathcal{N}_{1/2}}\boldsymbol{u}^\top\boldsymbol{\gamma}$. For every $\boldsymbol{u} \in \mathsf{S}^d$, note that $\boldsymbol{u}^\top\boldsymbol{\gamma} = \sum_{i=1}^n (\xi_i\boldsymbol{u}^\top\boldsymbol{z}_i - \mathsf{E}\xi_i\boldsymbol{u}^\top\boldsymbol{z}_i)$, and $\xi_i\boldsymbol{u}^\top\boldsymbol{z}_i$ are $i.i.d.$ sub-Gaussian variables with parameter $c_1\tau\nu_{\boldsymbol{x}}$. Then by Hoeffding's inequality and taking a union bound, we get

$$\mathsf{P}\left(\|\boldsymbol{\gamma}\|_2 > \delta\right) \le \exp\left\{(d+1)\log 5 - C\frac{n(\delta/2)^2}{(c_1\tau\nu_{\boldsymbol{x}})^2}\right\},$$

where C is some absolute constant. Combing the bound for $\|\mathsf{E}(\xi\boldsymbol{z})\|_2$, we obtain that with probability at least $1-e^{-t}$,

$$\|\mathbf{S}^{1/2}\,r\,L_\tau(\boldsymbol{\theta}_\tau)\|_2 \le C\nu_{\boldsymbol{x}}c_1\tau\left(\sqrt{\frac{d+t}{n}} + \epsilon\right), \tag{A.2.4}$$

for some (new) absolute constant C. Here, we use the fact that $1 = \sup_{\boldsymbol{u}\in\mathsf{S}^d}(\mathsf{E}|\boldsymbol{u}^\top\boldsymbol{z}|^2)^{1/2} \le C_0\nu_{\boldsymbol{x}}$ for some universal constant C_0.

Next we prove that the restricted strong convexity property (A.2.3) holds with high probability. For any $r > 0$, define the event

$$\mathcal{E}_i = \left\{|\langle\bar{\boldsymbol{x}}_i, \boldsymbol{\theta}-\boldsymbol{\theta}_\tau\rangle| \le \frac{c_3\tau}{2r}\|\boldsymbol{\theta}-\boldsymbol{\theta}_\tau\|_{\mathbf{S}}\right\} \cup \left\{|\varepsilon_i - \alpha_\tau| \le c_3\tau/2\right\},$$

where $c_3 > 0$ is as in Condition 2.2.1. By Proposition 2.2.1 and the convexity of ℓ_τ, we have

$$\begin{aligned} h\,r\,L_\tau(\boldsymbol{\theta}) \quad r\,L_\tau(\boldsymbol{\theta}_\tau), \boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau i &= \frac{1}{n}\sum_{i=1}^{n} f\ell_\tau^{0}(y_i \quad \bar{\boldsymbol{x}}_i^{!}\boldsymbol{\theta}_\tau) \quad \ell_\tau^{0}(y_i \quad \bar{\boldsymbol{x}}_i^{!}\boldsymbol{\theta})g\bar{\boldsymbol{x}}_i^{!}(\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau) \\ &\quad \frac{1}{n}\sum_{i=1}^{n} \big\{\ell_\tau^{0}(\varepsilon_i \quad \alpha_\tau) \quad \ell_\tau^{0}(\varepsilon_i \quad \alpha_\tau \quad \bar{\boldsymbol{x}}_i^{!}(\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau))g\bar{\boldsymbol{x}}_i^{!}(\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau)I_{E_i}. \end{aligned}$$

On E_i, note that $jy_i \quad \bar{\boldsymbol{x}}_i^{!}\boldsymbol{\theta}j \quad j\varepsilon_i \quad \alpha_\tau j + j\bar{\boldsymbol{x}}_i^{!}(\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau)j \quad c_3\tau$ for all $\boldsymbol{\theta}\ 2\ \boldsymbol{\Theta}_r$. Since $\ell_\tau^{00}(u) = \ell^{00}(u/\tau) \quad c_2$ for $juj \quad c_3\tau$, it follows that

$$h\,r\,L_\tau(\boldsymbol{\theta}) \quad r\,L_\tau(\boldsymbol{\theta}_\tau), \boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau i \quad \frac{c_2}{n}\sum_{i=1}^{n} h\bar{\boldsymbol{x}}_i, \boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau i^2 I_{E_i}$$

for all $\boldsymbol{\theta}\ 2\ \boldsymbol{\Theta}_r$. To smooth I_{E_i} as a function of $\boldsymbol{\theta}$, for any $R > 0$, we define

$$\varphi_R(u) = \begin{cases} u^2 & \text{if } juj \quad \frac{R}{2}, \\ (u \quad R)^2 & \text{if } \frac{R}{2} \quad u \quad R, \\ (u+R)^2 & \text{if } \quad R \quad u \quad \frac{R}{2}, \\ 0 & \text{if } juj > R, \end{cases} \quad \text{and} \quad \psi_R(u) = I(juj \quad R).$$

Under this notation, since $\varphi_R(u) \quad u^2 I(juj \quad R)$, we have

$$\begin{aligned} &h\,r\,L_\tau(\boldsymbol{\theta}) \quad r\,L_\tau(\boldsymbol{\theta}_\tau), \boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau i \\ &\quad c_2 g(\boldsymbol{\theta}) := \frac{c_2}{n}\sum_{i=1}^{n} \varphi_{c_3\tau k\boldsymbol{\delta}k_{\mathbf{S}}/(2r)}(h\bar{\boldsymbol{x}}_i, \boldsymbol{\delta} i)\psi_{c_3\tau/2}(\varepsilon_i \quad \alpha_\tau) \end{aligned} \tag{A.2.5}$$

for $\boldsymbol{\delta} = \boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau$ and $\boldsymbol{\theta}\ 2\ \boldsymbol{\Theta}_r$. Note that φ_R satisfies $\varphi_R(u) \quad u^2 I(juj \quad R/2)$. Therefore,

$$\begin{aligned} &\mathsf{E}g(\boldsymbol{\theta}) \quad \mathsf{E}\big\{h\bar{\boldsymbol{x}}, \boldsymbol{\delta} i^2 I f j h\bar{\boldsymbol{x}}, \boldsymbol{\delta} ij \quad c_3\tau k\boldsymbol{\delta}k_{\mathbf{S}}/(4r)g\psi_{c_3\tau/2}(\varepsilon \quad \alpha_\tau)\big\} \\ &\quad (1 \quad \epsilon)\mathsf{P}_{\varepsilon \ F}(j\varepsilon \quad \alpha_\tau j \quad c_3\tau/2) \ \mathsf{E}h\bar{\boldsymbol{x}}, \boldsymbol{\delta} i^2 I f j h\bar{\boldsymbol{x}}, \boldsymbol{\delta} ij \quad c_3\tau k\boldsymbol{\delta}k_{\mathbf{S}}/(4r)g \quad \epsilon\mathsf{E}h\bar{\boldsymbol{x}}, \boldsymbol{\delta} i^2 \\ &\quad (1 \quad \epsilon)\kappa_\tau \ [\mathsf{E}h\bar{\boldsymbol{x}}, \boldsymbol{\delta} i^2 \quad \mathsf{E}h\bar{\boldsymbol{x}}, \boldsymbol{\delta} i^2 I f j h\bar{\boldsymbol{x}}, \boldsymbol{\delta} ij > c_3\tau k\boldsymbol{\delta}k_{\mathbf{S}}/(4r)g] \quad \epsilon\mathsf{E}h\bar{\boldsymbol{x}}, \boldsymbol{\delta} i^2 \\ &\quad (1 \quad \epsilon)\kappa_\tau \ \left\{k\boldsymbol{\delta}k_{\mathbf{S}}^2 \quad \frac{(4r)^2}{(c_3\tau)^2 k\boldsymbol{\delta}k_{\mathbf{S}}^2}\mathsf{E}h\bar{\boldsymbol{x}}, \boldsymbol{\delta} i^4\right\} \quad \epsilon k\boldsymbol{\delta}k_{\mathbf{S}}^2. \end{aligned}$$

As $\boldsymbol{z} = \boldsymbol{S}^{1/2}\bar{\boldsymbol{x}}$ is a sub-Gaussian random vector with parameter $\nu_{\boldsymbol{x}}$, it holds $\mathsf{E}\langle \bar{\boldsymbol{x}}, \boldsymbol{u}\rangle^4 = \mathsf{E}\langle \boldsymbol{z}, \boldsymbol{S}^{1/2}\boldsymbol{u}\rangle^4 \leq C_0^4\nu_{\boldsymbol{x}}^4\|\boldsymbol{u}\|_{\mathbf{S}}^4$ for some absolute constant C_0, $\forall \boldsymbol{u} \in \mathsf{R}^{d+1}$. Substituting these estimates into the above inequality yields

$$\mathsf{E}g(\boldsymbol{\theta}) \leq \left\{(1+\epsilon)\kappa_\tau\left[1+C_0^4\nu_{\boldsymbol{x}}^4\left(\frac{4r}{c_3\tau}\right)^2\right]+\epsilon\right\}\|\boldsymbol{\theta}-\boldsymbol{\theta}_\tau\|_{\mathbf{S}}^2 \leq \frac{1}{2}(1+\epsilon)\kappa_\tau\|\boldsymbol{\theta}-\boldsymbol{\theta}_\tau\|_{\mathbf{S}}^2 \tag{A.2.6}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_r$ as long as $\tau \leq (8C_0^2\nu_{\boldsymbol{x}}^2/c_3)r$ and $\epsilon/(1+\epsilon) \leq \kappa_\tau/4$. For the stochastic term $g(\boldsymbol{\theta}) - \mathsf{E}g(\boldsymbol{\theta})$, we define

$$\Delta_r := \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r}\frac{|g(\boldsymbol{\theta})-\mathsf{E}g(\boldsymbol{\theta})|}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_\tau\|_{\boldsymbol{S}}^2} \tag{A.2.7}$$

Notice that Δ_r can be written in the form of

$$\Delta_r = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r}\left|\frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{\theta}}(\bar{\boldsymbol{x}}_i,\varepsilon_i) - \mathsf{E}f_{\boldsymbol{\theta}}(\bar{\boldsymbol{x}}_i,\varepsilon_i)\right|,$$

with

$$f_{\boldsymbol{\theta}}(\bar{\boldsymbol{x}}_i,\varepsilon_i) = \varphi_{c_3\tau/(2r)}\left(\langle\bar{\boldsymbol{x}}_i, \boldsymbol{\delta}/\|\boldsymbol{\delta}\|_{\boldsymbol{S}}\rangle\right)\psi_{c_3\tau/2}(\varepsilon_i - \alpha_\tau),$$

where $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}_\tau$ and we use the fact that $(1/c^2)\varphi_{R/c}(u) = \varphi_R(u/c)$. As $f_{\boldsymbol{\theta}}(\bar{\boldsymbol{x}}_i,\varepsilon_i)$ is bounded by $(c_3\tau/2r)^2$, by McDiarmid's Inequality, we have

$$\mathsf{P}(\Delta_r - \mathsf{E}\Delta_r > t) \leq \exp\left\{-\frac{2nt^2}{(c_3\tau/2r)^4}\right\}. \tag{A.2.8}$$

For the bound of $\mathsf{E}\Delta_r$, we can introduce independent Rademacher random variables $\widetilde{e}_1, \ldots, \widetilde{e}_n$ and apply the symmetrization inequality to get

$$\mathsf{E}\Delta_r \leq 2\mathsf{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r}\left|\frac{1}{n}\sum_{i=1}^n \widetilde{e}_i f_{\boldsymbol{\theta}}(\bar{\boldsymbol{x}}_i,\varepsilon_i)\right|\right].$$

To further bound $\mathsf{E}\Delta_r$, we rewrite $f_{\boldsymbol{\theta}}(\bar{\boldsymbol{x}}_i, \varepsilon_i) = \varphi_{c_3\tau/(2r)}\,(h\bar{\boldsymbol{x}}_i, \boldsymbol{\delta}/k\boldsymbol{\delta}k_{\boldsymbol{S}}i \quad \chi_i)$ with $\chi_i = \psi_{c_3\tau/2}(\varepsilon_i$ $\alpha_\tau)$ 2 $f0, 1g$. Since φ_R is R-Lipschitz, by Talagrand's contraction principle (see, e.g. Theorem 4.12 in Ledoux and Talagrand (2013)), we have

$$\begin{aligned}
\mathsf{E}\Delta_r \quad & 2\frac{c_3\tau}{r}\mathsf{E}\left[\sup_{\boldsymbol{\theta} 2 \boldsymbol{\Theta}_r}\left|\frac{1}{n}\sum_{i=1}^n \widetilde{e}_i h\bar{\boldsymbol{x}}_i, \boldsymbol{\delta}/k\boldsymbol{\delta}k_{\boldsymbol{S}}i\chi_i\right|\right] \\
& 2\frac{c_3\tau}{r}\mathsf{E}\left\|\frac{1}{n}\sum_{i=1}^n \widetilde{e}_i\chi_i\boldsymbol{z}_i\right\|_2 \quad 2\frac{c_3\tau}{r}\left\{\mathsf{E}\left(\frac{1}{n}\sum_{i=1}^n \widetilde{e}_i\chi_i\boldsymbol{z}_i\right)^{\mathsf{I}}\left(\frac{1}{n}\sum_{i=1}^n \widetilde{e}_i\chi_i\boldsymbol{z}_i\right)\right\}^{1/2} \\
& 2\frac{c_3\tau}{r}\left(\mathsf{E}\frac{1}{n^2}\sum_{i=1}^n \boldsymbol{z}_i^{\mathsf{I}}\boldsymbol{z}_i\right)^{1/2} = 2\frac{c_3\tau}{r}\sqrt{\frac{d+1}{n}}
\end{aligned}$$

Combing the above bound of $\mathsf{E}\Delta_r$ with $(A.2.8)$, we get that with probability at least $1 \quad e^{\ t}$,

$$\Delta_r \quad 2\frac{c_3\tau}{r}\sqrt{\frac{d+1}{n}} + \frac{c_3^2\tau^2}{4}\sqrt{\frac{t}{n}}. \tag{A.2.9}$$

Combining (A.2.5)–(A.2.7) and (A.2.9), we conclude that with probability at least $1 \quad e^{\ t}$, uniformly for all $\boldsymbol{\theta}$ 2 $\boldsymbol{\Theta}_r$, we have

$$h\,r\,L_\tau(\boldsymbol{\theta}) \quad r\,L_\tau(\boldsymbol{\theta}_\tau), \boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau i \quad \frac{c_2}{4}(1 \quad \epsilon)\kappa_\tau k\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau k_{\mathbf{S}}^2 \tag{A.2.10}$$

for $n \quad 64.5(1 \quad \epsilon)\ ^2\kappa_\tau{}^2\,[(c_3\tau/r)^2 \_ (c_3\tau/r)^4]\,(d+t+1)$.

With the probabilistic bounds (A.2.4) and (A.2.10) in hand, we can plug them in the basic inequality (A.2.1) to get that with probability at least $1 \quad 2e^{\ t}$,

$$\left\|\widehat{\boldsymbol{\theta}}_{\tau,\eta} \quad \boldsymbol{\theta}_\tau\right\|_{\mathbf{S}} \quad \frac{4C\nu_{\boldsymbol{x}}c_1\tau}{c_2(1 \quad \epsilon)\kappa_\tau}\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) := r_0,$$

as long as $\tau \quad (8C_0^2\nu_{\boldsymbol{x}}^2/c_3)r$ and $\epsilon/(1 \quad \epsilon) \quad \kappa_\tau/4$ and $n \quad 64.5(1 \quad \epsilon)\ ^2\kappa_\tau{}^2[(c_3\tau/r)^2 \_ (c_3\tau/r)^4](d+t+1)$. Pick $r = c_3\tau/(8C_0^2\nu_{\boldsymbol{x}}^2)$, then for $\epsilon < c\kappa_\tau$ and $n \quad C^{\emptyset}\kappa_\tau{}^2(d+t)$, with $c, C^{\emptyset} > 0$ satisfying that $c < 1/5, C^{\emptyset} > 8 \quad 64.5[(8C_0^2\nu_{\boldsymbol{x}}^2)^2 \_ (8C_0^2\nu_{\boldsymbol{x}}^2)^4]$ (such that $\epsilon/(1 \quad \epsilon)$ $\kappa_\tau/(5 \quad \kappa_\tau) \quad \kappa_\tau/4$ and $n \quad 64.5(1 \quad \epsilon)\ ^2\kappa_\tau{}^2[(c_3\tau/r)^2 \_ (c_3\tau/r)^4](d+t+1)$), we get that with probability at least $1 \quad 2e^{\ t}$,

$$\left\|\widehat{\boldsymbol{\theta}}_{\tau,\eta} \quad \boldsymbol{\theta}_\tau\right\|_{\mathbf{S}} \quad r_0 \quad \frac{4C\nu_{\boldsymbol{x}}c_1\tau}{c_2(1 \quad \epsilon)}\left(\sqrt{\frac{1}{C^{\emptyset}}} + c\right) < r,$$

for some sufficiently large $C^{\emptyset}$ and some sufficiently small c, both depending only on $(\nu_{\boldsymbol{x}}, c_1, c_2, c_3)$. This implies that $\widehat{\boldsymbol{\theta}}_\tau$ 2 $\boldsymbol{\Theta}_r$ and thus $\widehat{\boldsymbol{\theta}}_{\tau,\eta} = \widehat{\boldsymbol{\theta}}_\tau$, which gives us that $k\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau k_{\mathbf{S}} \quad r_0$ with probability at least $1 \quad 2e^{\ t}$. This proves the stated result. □

A.2.3 Proof of Theorem 2.2.2

We only show the second term ϵ in the lower bound. We use Le Cam's two point testing method and try to construct two distributions $P_1(\boldsymbol{x}, y) = P(\boldsymbol{\theta}_1, \epsilon, F_1, G_{\boldsymbol{x},1}), P_2(\boldsymbol{x}, y) = P(\boldsymbol{\theta}_2, \epsilon, F_2, G_{\boldsymbol{x},2})$ with $k\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2 k_S \quad \epsilon$ and $P_1(y|\boldsymbol{x}) = P_2(y|\boldsymbol{x})$ with a positive probability. First, we choose $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ such that $k\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2 k_S = \epsilon$, and $F_1 = F_2 = N(0,1)$. Then $P_i(y|\boldsymbol{x}) = (1 \quad \epsilon)N(\bar{\boldsymbol{x}}^\intercal \boldsymbol{\theta}_i, 1) + \epsilon G_{\boldsymbol{x},i}$ for $i = 1, 2$ and some $G_{\boldsymbol{x},1}, G_{\boldsymbol{x},2}$ to be specified later. We define an event

$$A_{\boldsymbol{x}} := \left\{ k\bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2)k_2 \quad \frac{2\epsilon}{1 \quad \epsilon} \right\}.$$

By Pinsker's inequality, on this event $A_{\boldsymbol{x}}$, we always have

$$\begin{aligned} \mathrm{TV}(N(\bar{\boldsymbol{x}}^\intercal\boldsymbol{\theta}_1, 1), N(\bar{\boldsymbol{x}}^\intercal\boldsymbol{\theta}_2, 1)) \quad & \sqrt{\frac{1}{2}\mathrm{KL}(N(\bar{\boldsymbol{x}}^\intercal\boldsymbol{\theta}_1, 1)||N(\bar{\boldsymbol{x}}^\intercal\boldsymbol{\theta}_2, 1))} \\ = & \sqrt{\frac{1}{4}k\bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2)k_2^2} \quad \frac{\epsilon}{1 \quad \epsilon}. \end{aligned}$$

Then by the same argument in the proof of Theorem 5.1 in Chen et al. (2018), we know there exists two distributions $G_{\boldsymbol{x},1}, G_{\boldsymbol{x},2}$ (conditional on $\boldsymbol{x}$) such that

$$(1 \quad \epsilon)N(\bar{\boldsymbol{x}}^\intercal\boldsymbol{\theta}_1, 1) + \epsilon G_{\boldsymbol{x},1} = (1 \quad \epsilon)N(\bar{\boldsymbol{x}}^\intercal\boldsymbol{\theta}_2, 1) + \epsilon G_{\boldsymbol{x},2}.$$

That is $P_1(y|\boldsymbol{x}) = P_2(y|\boldsymbol{x})$ on the event $A_{\boldsymbol{x}}$. Therefore, we can get

$$\begin{aligned} & \inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \ 2\mathbb{R}^{d+1}} \sup_{P(\boldsymbol{x},y)2P} \mathrm{E}_{P(\boldsymbol{x},y)} k\widehat{\boldsymbol{\theta}} \quad \boldsymbol{\theta} \ k_S \\ & \quad \frac{1}{2}\inf_{\widehat{\boldsymbol{\theta}}} \left(\mathrm{E}_{P_1(\boldsymbol{x},y)} k\widehat{\boldsymbol{\theta}} \quad \boldsymbol{\theta}_1 k_S + \mathrm{E}_{P_2(\boldsymbol{x},y)} k\widehat{\boldsymbol{\theta}} \quad \boldsymbol{\theta}_2 k_S \right) \\ & \quad \frac{1}{2}\inf_{\widehat{\boldsymbol{\theta}}} \mathrm{E}_{\boldsymbol{x}} \left[I(A_{\boldsymbol{x}}) \quad \left(\mathrm{E}_{P_1(y|\boldsymbol{x})} k\widehat{\boldsymbol{\theta}} \quad \boldsymbol{\theta}_1 k_S + \mathrm{E}_{P_2(y|\boldsymbol{x})} k\widehat{\boldsymbol{\theta}} \quad \boldsymbol{\theta}_2 k_S \right) \right] \\ & \quad \frac{1}{2}\mathrm{E}_{\boldsymbol{x}} I(A_{\boldsymbol{x}}) \, k\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2 k_S = \frac{1}{2} k\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2 k_S \, \mathrm{P}(A_{\boldsymbol{x}}) \\ & \quad \frac{1}{2} k\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2 k_S \left(1 \quad \frac{\mathrm{E}\, k\bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2)k_2^2}{4\epsilon^2/(1 \quad \epsilon)^2} \right) \\ = & \frac{1}{2} k\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2 k_S \left(1 \quad \frac{k\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2 k_S^2}{4\epsilon^2/(1 \quad \epsilon)^2} \right) = \frac{\epsilon}{2}\left(1 \quad \frac{(1 \quad \epsilon)^2}{4} \right) \ \& \ \epsilon. \end{aligned}$$

The lower bound for ℓ_2 estimation error of $\boldsymbol{\beta}$ can be proved in a similar manner, by choosing $\boldsymbol{\theta}_1 = (0, \boldsymbol{\beta}_1^\intercal)^\intercal, \boldsymbol{\theta}_2 = (0, \boldsymbol{\beta}_2^\intercal)^\intercal$. □

A.2.4 Proof of Theorem 2.3.1

We introduce the following lemma, which will lead to the result of Theorem 2.3.1 immediately and will be also used in other proofs.

Lemma A.2.1. Under the same condition of Theorem 2.3.1, for any $r > 0$, we have

$$\sup_{\boldsymbol{\theta} 2 \boldsymbol{\Theta}_r} \left\| \mathbf{S}^{1/2} \boldsymbol{f} r \boldsymbol{L}_\tau(\boldsymbol{\theta}) \quad r \boldsymbol{L}_\tau(\boldsymbol{\theta}_\tau) \quad \mathsf{E}\ell''_\tau(\varepsilon \quad \alpha_\tau)\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathsf{I}}(\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau)\boldsymbol{g} \right\|_2 \quad C_3 \left(r^2 + \epsilon r + r\sqrt{\frac{d+t}{n}} \right) \tag{A.2.11}$$

with probability at least $1 \quad e^{\ t}$, as long as $n \quad C_4(d+t)$ and $t \quad 1/2$, where $\boldsymbol{\Theta}_r = \boldsymbol{f}\boldsymbol{\theta}\ \boldsymbol{2}\ \mathsf{R}^{d+1} : \boldsymbol{k}\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau \boldsymbol{k}_{\mathbf{S}} \quad r\boldsymbol{g}$ and C_3, C_4 are some constants depending only on $(c_4, \nu_{\boldsymbol{x}}, C_F, L, \tau)$.

Let $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_\tau$ in (A.2.11) (note that $\boldsymbol{r}\boldsymbol{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) = 0$) and denote r_0 as the bound of $\boldsymbol{k}\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau \boldsymbol{k}_{\mathbf{S}}$ in (2.2.4) of Theorem 2.2.1, then we have

$$\left\| \mathbf{S}^{1/2} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell'_\tau(\varepsilon_i \quad \alpha_\tau)\bar{\boldsymbol{x}}_i \quad [\mathsf{E}\ell''_\tau(\varepsilon \quad \alpha_\tau)\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathsf{I}}] (\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \right\} \right\|_2 \quad C_3 \left(r_0^2 + \epsilon r_0 + r_0\sqrt{\frac{d+t}{n}} \right)$$

with probabilty at least $1 \quad 3e^{\ t}$. It is easy to check that

$$\begin{aligned}
&\left\| c_F \mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \quad \mathbf{S}^{1/2} [\mathsf{E}\ell''_\tau(\varepsilon \quad \alpha_\tau)\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathsf{I}}] (\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \right\|_2 \\
&= \epsilon \left\| \mathbf{S}^{1/2} \left[\mathsf{E} \left\{ \mathsf{E}_{\varepsilon | \boldsymbol{x}} \ {}_F \ell''_\tau(\varepsilon \quad \alpha_\tau)\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathsf{I}} \quad \mathsf{E}_{\varepsilon | \boldsymbol{x}} \ {}_{G_{\boldsymbol{x}}} \ell''_\tau(\varepsilon \quad \alpha_\tau)\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathsf{I}} \right\} \right] (\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \right\|_2 \\
&\quad 2c_4\epsilon \sup_{\boldsymbol{u} 2 \mathsf{S}^d} \left[\mathsf{E} \left\| \boldsymbol{u}^{\mathsf{I}} \mathbf{S}^{1/2} \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathsf{I}} \mathbf{S}^{1/2} \right\|_2 \right] \left\| \mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \right\|_2 \\
&= 2c_4\epsilon \sup_{\boldsymbol{u} 2 \mathsf{S}^d} \sup_{\boldsymbol{v} 2 \mathsf{S}^d} [\mathsf{E} \boldsymbol{j}\boldsymbol{u}^{\mathsf{I}} \boldsymbol{z}\boldsymbol{z}^{\mathsf{I}} \boldsymbol{v}\boldsymbol{j}] \left\| \mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \right\|_2 \\
&\quad 2c_4\epsilon \sup_{\boldsymbol{u} 2 \mathsf{S}^d} \sup_{\boldsymbol{v} 2 \mathsf{S}^d} \left(\mathsf{E} \boldsymbol{j}\boldsymbol{u}^{\mathsf{I}} \boldsymbol{z}\boldsymbol{j}^2 \, \mathsf{E} \boldsymbol{j}\boldsymbol{v}^{\mathsf{I}} \boldsymbol{z}\boldsymbol{j}^2 \right)^{1/2} \boldsymbol{k}\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau \boldsymbol{k}_{\mathbf{S}} \\
&\quad 2c_4\epsilon \boldsymbol{k}\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau \boldsymbol{k}_{\mathbf{S}} = 2c_4\epsilon r_0.
\end{aligned} \tag{A.2.12}$$

Here, we use the fact that $\boldsymbol{j}\ell''(u)\boldsymbol{j} \quad c_4$ for any $u\ \boldsymbol{2}\ \mathsf{R}$ ($c_4 = 1$ for Huber's loss). As $r_0 \quad C(\sqrt{\frac{d+t}{n}} + \epsilon)$ with probability at least $1 \quad 2e^{\ t}$ by Theorem 2.2.1 for some constant C depending only on $(\nu_{\boldsymbol{x}}, \kappa_\tau, c_1, c_2, c_3)$, we obtain that (2.3.1) holds with probability at least $1 \quad 3e^{\ t}$. □

A.2.5 Proof of Lemma A.2.1

Denote $B(\boldsymbol{\theta}) := \mathbf{S}^{1/2}\{\nabla L_\tau(\boldsymbol{\theta}) - \nabla L_\tau(\boldsymbol{\theta}_\tau) - \mathsf{E}\ell''_\tau(\varepsilon - \alpha_\tau)\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau)\}$. Our idea is to bound $\|\mathsf{E}B(\boldsymbol{\theta})\|_2$ and $\|B(\boldsymbol{\theta}) - \mathsf{E}B(\boldsymbol{\theta})\|_2$ uniformly for $\boldsymbol{\theta} \in \Theta_r$.

Step 1: Bound $\sup_{\boldsymbol{\theta}\in\Theta_r}\|\mathsf{E}B(\boldsymbol{\theta})\|_2$. Standard calculation shows that

$$\begin{aligned}\mathsf{E}\{\nabla L_\tau(\boldsymbol{\theta}) - \nabla L_\tau(\boldsymbol{\theta}_\tau)\} &= \mathsf{E}\{[\ell'_\tau(\varepsilon - \alpha_\tau + \bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta}_\tau - \boldsymbol{\theta})) - \ell'_\tau(\varepsilon - \alpha_\tau)](-\bar{\boldsymbol{x}})\} \\ &= \mathsf{E}\int_0^1 \ell''_\tau(\varepsilon - \alpha_\tau + t\bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta}_\tau - \boldsymbol{\theta}))\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau)\,dt.\end{aligned}$$

Therefore,

$$\|\mathsf{E}B(\boldsymbol{\theta})\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^d}\mathsf{E}\int_0^1 [\ell''_\tau(\varepsilon - \alpha_\tau + t\bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta}_\tau - \boldsymbol{\theta})) - \ell''_\tau(\varepsilon - \alpha_\tau)]\,\boldsymbol{u}^\intercal\mathbf{S}^{1/2}\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau)\,dt$$

Under Conditions 2.3.1 (i), ℓ'' is L-Lipschitz, and thus ℓ''_τ is L/τ-Lipschitz, then we have

$$\begin{aligned}\|\mathsf{E}B(\boldsymbol{\theta})\|_2 &\le (L/\tau)\sup_{\boldsymbol{u}\in\mathbb{S}^d}\mathsf{E}|\boldsymbol{u}^\intercal\mathbf{S}^{1/2}\bar{\boldsymbol{x}}|\,|\bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau)|^2 \\ &\le (L/\tau)\sup_{\boldsymbol{u}\in\mathbb{S}^d}\mathsf{E}|\boldsymbol{u}^\intercal\boldsymbol{z}|\,|\boldsymbol{z}^\intercal\mathbf{S}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau)|^2 \\ &\le (L/\tau)\sup_{\boldsymbol{u}\in\mathbb{S}^d}\left(\mathsf{E}|\boldsymbol{u}^\intercal\boldsymbol{z}|^2\,\mathsf{E}|\boldsymbol{z}^\intercal\mathbf{S}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau)|^4\right)^{1/2} \\ &\le (L/\tau)\sup_{\boldsymbol{u}\in\mathbb{S}^d}\left(\|\boldsymbol{u}\|_2^2\,C_0^4\nu_{\boldsymbol{x}}^4\|\mathbf{S}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau)\|_2^4\right)^{1/2} \\ &\le (L/\tau)C_0^2\nu_{\boldsymbol{x}}^2\|\boldsymbol{\theta} - \boldsymbol{\theta}_\tau\|_{\mathbf{S}}^2 \le (L/\tau)C_0^2\nu_{\boldsymbol{x}}^2 r^2, \quad \forall\boldsymbol{\theta}\in\Theta_r,\end{aligned} \tag{A.2.13}$$

where we use the fact that $\boldsymbol{z}^\intercal\boldsymbol{v}$ is sub-Gaussian with parameter $\nu_{\boldsymbol{x}}\|\boldsymbol{v}\|_2$ for any $\boldsymbol{v}\in\mathbb{R}^{d+1}$. Under Conditions 2.3.1 (ii), ℓ is Huber's loss, then $\ell''_\tau(u) = I(|u| \le \tau)$. Since $|I(|\varepsilon - \alpha_\tau + t\Delta| \le \tau) - I(|\varepsilon - \alpha_\tau| \le \tau)| \le I(|\varepsilon - \alpha_\tau| \le |\Delta|)$ for any $t\in[0,1]$ with $\Delta := \bar{\boldsymbol{x}}^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau)$, we get

$$\begin{aligned}\|\mathsf{E}B(\boldsymbol{\theta})\|_2 &\le \sup_{\boldsymbol{u}\in\mathbb{S}^d}\mathsf{E}\left\{I(|\varepsilon - \alpha_\tau| \le \tau \le |\varepsilon - \alpha_\tau| + |\Delta|)|\boldsymbol{u}^\intercal\mathbf{S}^{1/2}\bar{\boldsymbol{x}}\Delta|\right\} \\ &\le \sup_{\boldsymbol{u}\in\mathbb{S}^d}(1-\epsilon)\mathsf{E}\left\{\mathsf{P}_{\varepsilon|\boldsymbol{x}}(|\varepsilon - \alpha_\tau| \le \tau \le |\varepsilon - \alpha_\tau| + |\Delta|)\,|\boldsymbol{u}^\intercal\mathbf{S}^{1/2}\bar{\boldsymbol{x}}\Delta|\right\} + \epsilon\mathsf{E}|\boldsymbol{u}^\intercal\mathbf{S}^{1/2}\bar{\boldsymbol{x}}\Delta| \\ &\le \sup_{\boldsymbol{u}\in\mathbb{S}^d}(1-\epsilon)4C_F\mathsf{E}|\boldsymbol{u}^\intercal\boldsymbol{z}|\Delta^2 + \epsilon\mathsf{E}|\boldsymbol{u}^\intercal\boldsymbol{z}\Delta|\end{aligned}$$

Similar to the derivation in (A.2.13), we have $\sup_{\boldsymbol{u}\in\mathbb{S}^d}\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{z}|\boldsymbol{g}\Delta^2 \quad C_0^2\nu_{\boldsymbol{x}}^2r^2$. Also, we have $\sup_{\boldsymbol{u}\in\mathbb{S}^d}\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{z}|\Delta| \quad \sup_{\boldsymbol{u}\in\mathbb{S}^d}\left(\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{z}|^2\,\mathbb{E}\left|\boldsymbol{z}^\top\mathbb{S}^{1/2}(\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau)\right|^2\right)^{1/2} \quad \|\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau\|_{\mathbb{S}} \quad r$ for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}_r$. Therefore, under Conditions 2.3.1 (ii), we get

$$\|\mathbb{E}B(\boldsymbol{\theta})\|_2 \quad 4C_FC_0^2\nu_{\boldsymbol{x}}^2r^2+\epsilon r, \quad \boldsymbol{\theta}\in\boldsymbol{\Theta}_r. \tag{A.2.14}$$

Step 2: Bound $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r}\|B(\boldsymbol{\theta}) \quad \mathbb{E}B(\boldsymbol{\theta})\|_2$. Denote $\boldsymbol{\delta}:=\mathbb{S}^{1/2}(\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau)$. Then we can write

$$B(\boldsymbol{\theta}) \quad \mathbb{E}B(\boldsymbol{\theta})=\mathbb{S}^{\ 1/2}\{\nabla L_\tau(\boldsymbol{\theta}) \quad \nabla L_\tau(\boldsymbol{\theta}_\tau)\} \quad \mathbb{E}\mathbb{S}^{\ 1/2}\{\nabla L_\tau(\boldsymbol{\theta}) \quad \nabla L_\tau(\boldsymbol{\theta}_\tau)\},$$

with

$$\mathbb{S}^{\ 1/2}\{\nabla L_\tau(\boldsymbol{\theta}) \quad \nabla L_\tau(\boldsymbol{\theta}_\tau)\}=\frac{1}{n}\sum_{i=1}^n[\ell'_\tau(\varepsilon_i \quad \alpha_\tau \quad \boldsymbol{z}_i^\top\boldsymbol{\delta}) \quad \ell'_\tau(\varepsilon_i \quad \alpha_\tau)](\quad \boldsymbol{z}_i).$$

Denote $\bar{B}(\boldsymbol{\delta}):=B(\boldsymbol{\theta}) \quad \mathbb{E}B(\boldsymbol{\theta})$. Then $\bar{B}(\mathbf{0})=\mathbf{0}$, $\mathbb{E}\bar{B}(\boldsymbol{\delta})=\mathbf{0}, \boldsymbol{\delta}\in\mathbb{R}^{d+1}$, and

$$\nabla_{\boldsymbol{\delta}}\bar{B}(\boldsymbol{\delta})=\frac{1}{n}\sum_{i=1}^n[\ell''_\tau(\varepsilon_i \quad \alpha_\tau \quad \boldsymbol{z}_i^\top\boldsymbol{\delta})\boldsymbol{z}_i\boldsymbol{z}_i^\top \quad \mathbb{E}\ell''_\tau(\varepsilon_i \quad \alpha_\tau \quad \boldsymbol{z}_i^\top\boldsymbol{\delta})\boldsymbol{z}_i\boldsymbol{z}_i^\top].$$

In addition, for any $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^d$ and $\lambda\in\mathbb{R}$, using the inequality $|e^z \quad 1 \quad z| \quad z^2e^{|z|}/2$ for all $z\in\mathbb{R}$, we have

$$\begin{aligned}&\mathbb{E}\exp\left\{\lambda\sqrt{n}\boldsymbol{u}^\top\nabla_{\boldsymbol{\delta}}\bar{B}(\boldsymbol{\delta})\boldsymbol{v}\right\}\\ &\prod_{i=1}^n\left[1+\frac{c_4^2\lambda^2}{n}\mathbb{E}\left\{(\boldsymbol{u}^\top\boldsymbol{z}_i\boldsymbol{v}^\top\boldsymbol{z}_i)^2+(\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{z}\boldsymbol{v}^\top\boldsymbol{z}|)^2\right\}e^{c_4\frac{|\lambda|}{\sqrt{n}}(|\boldsymbol{u}^\top\boldsymbol{z}_i\boldsymbol{v}^\top\boldsymbol{z}_i|+\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{z}\boldsymbol{v}^\top\boldsymbol{z}|)}\right],\end{aligned}$$

where c_4 is the bound of ℓ''' in the Conditions 2.3.1 (i). For Huber's loss, $c_4=1$. Notice that

$$\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{z}\boldsymbol{v}^\top\boldsymbol{z}| \quad \left(\mathbb{E}(\boldsymbol{u}^\top\boldsymbol{z})^2\mathbb{E}(\boldsymbol{v}^\top\boldsymbol{z})^2\right)^{1/2}=1,$$

and

$$\begin{aligned}&\mathbb{E}(|\boldsymbol{u}^\top\boldsymbol{z}\boldsymbol{v}^\top\boldsymbol{z}|^2+1)e^{c_4\frac{|\lambda|}{\sqrt{n}}(|\boldsymbol{u}^\top\boldsymbol{z}\boldsymbol{v}^\top\boldsymbol{z}|+1)} \quad e^{c_4\frac{|\lambda|}{\sqrt{n}}}\mathbb{E}(|\boldsymbol{u}^\top\boldsymbol{z}\boldsymbol{v}^\top\boldsymbol{z}|^2+1)e^{c_4\frac{|\lambda|}{2}\frac{|\lambda|}{\sqrt{n}}(|\boldsymbol{u}^\top\boldsymbol{z}|^2+|\boldsymbol{v}^\top\boldsymbol{z}|^2)}\\ &e^{c_4\frac{|\lambda|}{\sqrt{n}}}\left\{\left(\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{z}|^4e^{c_4\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{u}^\top\boldsymbol{z}|^2} \quad \mathbb{E}|\boldsymbol{v}^\top\boldsymbol{z}|^4e^{c_4\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{v}^\top\boldsymbol{z}|^2}\right)^{1/2}+\left(\mathbb{E}e^{c_4\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{u}^\top\boldsymbol{z}|^2} \quad \mathbb{E}e^{c_4\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{v}^\top\boldsymbol{z}|^2}\right)^{1/2}\right\}\\ &e^{c_4\frac{|\lambda|}{\sqrt{n}}}\left\{\sup_{\boldsymbol{u}\in\mathbb{S}^d}\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{z}|^4e^{c_4\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{u}^\top\boldsymbol{z}|^2}+\sup_{\boldsymbol{u}\in\mathbb{S}^d}\mathbb{E}e^{c_4\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{u}^\top\boldsymbol{z}|^2}\right\}.\end{aligned}$$

Since $\boldsymbol{z}$ is a sub-Gaussian vector with parameter $\nu_{\boldsymbol{x}}$, we know that for all λ such that $|\lambda| \leq \sqrt{n}/(2c_4\nu_{\boldsymbol{x}}^2)$, we always have $\mathbb{E}e^{c_4\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{u}^\top\boldsymbol{z}|^2} \leq 2$ and $\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{z}|^4 e^{c_4\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{u}^\top\boldsymbol{z}|^2} \leq (\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{z}|^8 \mathbb{E}e^{2c_4\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{u}^\top\boldsymbol{z}|^2})^{1/2} \leq (C_0^8\nu_{\boldsymbol{x}}^8 \cdot 2)^{1/2} = \sqrt{2}C_0^4\nu_{\boldsymbol{x}}^4$ for all $\boldsymbol{u} \in \mathbb{S}^d$ with some universal constant C_0. Therefore, for all $|\lambda| \leq \sqrt{n}/(2c_4\nu_{\boldsymbol{x}}^2)$ and all $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$, we have

$$\begin{aligned}
&\mathbb{E}\exp\left\{\lambda\sqrt{n}\boldsymbol{u}^\top \nabla_{\boldsymbol{\delta}}\bar{B}(\boldsymbol{\delta})\boldsymbol{v}\right\} \\
&\prod_{i=1}^{n}\left[1+\frac{c_4^2\lambda^2}{n}e^{c_4\frac{|\lambda|}{\sqrt{n}}}\left(\sqrt{2}C_0^4\nu_{\boldsymbol{x}}^4+2\right)\right] \leq \exp\left\{c_4^2\lambda^2 e^{1/(2\nu_{\boldsymbol{x}}^2)}\left(\sqrt{2}C_0^4\nu_{\boldsymbol{x}}^4+2\right)\right\} =: \exp\{C^2\lambda^2/2\},
\end{aligned}$$

with $C^2 = 2c_4^2 e^{1/(2\nu_{\boldsymbol{x}}^2)}\left(\sqrt{2}C_0^4\nu_{\boldsymbol{x}}^4+2\right)$. Then by Theorem A.3 in Spokoiny (2013), we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\delta}\in\mathbb{B}^{d+1}(r)} \|\sqrt{n}\bar{B}(\boldsymbol{\delta})\|_2 \geq 6Cr\sqrt{4(d+1)+2t}\right) \leq e^{-t}, \tag{A.2.15}$$

for any $t \geq 1/2$ and any $r > 0$, if $n \geq 8c_4^2\nu_{\boldsymbol{x}}^4(4(d+1)+2t)$.

Combing (A.2.13)-(A.2.15), we obtain that with probability at least $1 - e^{-t}$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \|B(\boldsymbol{\theta})\|_2 \leq C'\left(r^2 + \epsilon r + r\sqrt{\frac{d+t}{n}}\right)$$

with some constant $C' = C'(c_4, \nu_{\boldsymbol{x}}, C_F, L, \tau)$. □

A.2.6 Proof of Theorem 2.3.2

The key of the proof is the following lemma, which gives an bound of the difference between $\nabla L_\tau(\boldsymbol{\theta})$ and $\nabla L_\tau^\flat(\boldsymbol{\theta})$ so that we can apply the results about $\nabla L_\tau(\boldsymbol{\theta})$, like local strong convexity to the bootstrap case.

Lemma A.2.2. Let

$$\xi^\flat(\boldsymbol{\theta}) := \mathbf{S}^{-1/2}\left(\nabla L_\tau^\flat(\boldsymbol{\theta}) - \nabla L_\tau(\boldsymbol{\theta})\right) = \frac{1}{n}\sum_{i=1}^{n} e_i\ell_\tau'(\varepsilon_i - \alpha_\tau - \bar{\boldsymbol{x}}_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau))(-\mathbf{S}^{-1/2}\bar{\boldsymbol{x}}_i), \tag{A.2.16}$$

then under Condition 2.2.1-2.3.2, with probability (over $\mathcal{D}_n$) at least $1 - 2e^{-t}$, the following inequalities hold simultaneously.

(i)

$$\mathsf{P}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \left\|\xi^{\flat}(\boldsymbol{\theta}) - \xi^{\flat}(\boldsymbol{\theta}_\tau)\right\|_2 \leq Cr\sqrt{\frac{d+t}{n}}\right) \geq 1 - e^{-t} \tag{A.2.17}$$

(ii)

$$\mathsf{P}\left(\|\xi^{\flat}(\boldsymbol{\theta}_\tau)\|_2 \leq C\tau\sqrt{\frac{d+t}{n}}\right) \geq 1 - e^{-t} \tag{A.2.18}$$

(iii)

$$\mathsf{P}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \|\xi^{\flat}(\boldsymbol{\theta})\|_2 \leq C(r+\tau)\sqrt{\frac{d+t}{n}}\right) \geq 1 - 2e^{-t} \tag{A.2.19}$$

as long as $n \geq (d+t)^2$, where C is a constant only depending on $(c_1, c_4, \nu_{\boldsymbol{x}}, \nu_e)$.

We first apply Lemma A.2.2 to prove Theorem 2.3.2, leaving the proof of Lemma A.2.2 at the end.

Proof of Theorem 2.3.2 (i): Recall that $\boldsymbol{\Theta}_r = \{\boldsymbol{\theta} \in \mathsf{R}^{d+1} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_\tau\|_{\mathbf{S}} \leq r\}$, where $\boldsymbol{\theta}_\tau = \operatorname{argmin}_{\boldsymbol{\theta}} \mathsf{E}_{\boldsymbol{x},\varepsilon}\{\ell_\tau(y - \bar{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{\theta})\}$. For any prespecified $r > 0$, we use a similar way as we do in the proof of Theorem 2.2.1 to find an intermediate estimator $\widehat{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} = \boldsymbol{\theta}_\tau + \eta(\widehat{\boldsymbol{\theta}}^{\flat}_\tau - \boldsymbol{\theta}_\tau)$ for some $\eta \in [0,1]$, such that $\|\widehat{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau\|_{\mathbf{S}} \leq r$. That is, if $\widehat{\boldsymbol{\theta}}^{\flat}_\tau \in \boldsymbol{\Theta}_r$, we simply take $\eta = 1$; otherwise, we let $\eta = r/\|\widehat{\boldsymbol{\theta}}^{\flat}_\tau - \boldsymbol{\theta}_\tau\|_{\mathbf{S}}$ such that $\|\widehat{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau\|_{\mathbf{S}} = r$. As $L^{\flat}_\tau(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$ (by non-negativity of w_i), applying Lemma F.2 in Fan et al. (2018), we have

$$\begin{aligned}\left\langle \nabla L^{\flat}_\tau(\widehat{\boldsymbol{\theta}}^{\flat}_{\tau,\eta}) - \nabla L^{\flat}_\tau(\boldsymbol{\theta}_\tau), \widehat{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\rangle &\leq \eta \left\langle \nabla L^{\flat}_\tau(\widehat{\boldsymbol{\theta}}^{\flat}_\tau) - \nabla L^{\flat}_\tau(\boldsymbol{\theta}_\tau), \widehat{\boldsymbol{\theta}}^{\flat}_\tau - \boldsymbol{\theta}_\tau \right\rangle \\ &= \eta \left\langle \nabla L^{\flat}_\tau(\boldsymbol{\theta}_\tau), \widehat{\boldsymbol{\theta}}^{\flat}_\tau - \boldsymbol{\theta}_\tau \right\rangle \\ &\quad \left(\|\mathbf{S}^{-1/2}\nabla L_\tau(\boldsymbol{\theta}_\tau)\|_2 + \|\xi^{\flat}(\boldsymbol{\theta}_\tau)\|_2\right) \|\widehat{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau\|_{\mathbf{S}}.\end{aligned}$$

By (A.2.4) and Lemma A.2.2 (ii), we obtain that with probability at least $1 - 3e^{-t}$,

$$\mathsf{P}\left(\left\langle \nabla L_\tau^\flat(\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat) - \nabla L_\tau^\flat(\boldsymbol{\theta}_\tau), \widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat - \boldsymbol{\theta}_\tau \right\rangle \le C\tau\left(\sqrt{\frac{d+t}{n}} + \epsilon\right)\left\|\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat - \boldsymbol{\theta}_\tau\right\|_{\mathbf{S}}\right) \ge 1 - e^{-t},$$

where C is a constant only depending on $(c_1, c_4, \nu_{\boldsymbol{x}}, \nu_e)$. On the other hand, notice that

$$\left\langle \nabla L_\tau^\flat(\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat) - \nabla L_\tau^\flat(\boldsymbol{\theta}_\tau), \widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat - \boldsymbol{\theta}_\tau \right\rangle \le \left\langle \nabla L_\tau(\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat) - \nabla L_\tau(\boldsymbol{\theta}_\tau), \widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat - \boldsymbol{\theta}_\tau \right\rangle + \left\|\xi^\flat(\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat) - \xi^\flat(\boldsymbol{\theta}_\tau)\right\|_2 \left\|\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat - \boldsymbol{\theta}_\tau\right\|_{\mathbf{S}}.$$

By the local strong convexity of $L_\tau(\boldsymbol{\theta})$ (A.2.10) and Lemma A.2.2 (i), we obtain that with probability at least $1 - 3e^{-t}$,

$$\mathsf{P}\left(\left\langle \nabla L_\tau^\flat(\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat) - \nabla L_\tau^\flat(\boldsymbol{\theta}_\tau), \widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat - \boldsymbol{\theta}_\tau \right\rangle \ge \alpha\left\|\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat - \boldsymbol{\theta}_\tau\right\|_{\mathbf{S}}^2 - C'r\sqrt{\frac{d+t}{n}}\left\|\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat - \boldsymbol{\theta}_\tau\right\|_{\mathbf{S}}\right) \ge 1 - e^{-t}, \tag{A.2.20}$$

where $\alpha := \frac{c_2}{4}(1-\epsilon)\kappa_\tau$ and C' is a constant only depending on $(c_1, c_4, \nu_{\boldsymbol{x}}, \nu_e)$. Combing the above results together, we get that with probability at least $1 - 4e^{-t}$,

$$\mathsf{P}\left(\left\|\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat - \boldsymbol{\theta}_\tau\right\|_{\mathbf{S}} \le C''\frac{\tau + r}{\alpha}\left(\sqrt{\frac{d+t}{n}} + \epsilon\right)\right) \ge 1 - 2e^{-t},$$

where C'' is a constant only depending on $(c_1, c_4, \nu_{\boldsymbol{x}}, \nu_e)$. Pick $r = c_3\tau/(8C_0^2\nu_{\boldsymbol{x}}^2)$ as what we do in the proof of Theorem 2.2.1, then for any sufficiently large n and small ϵ, we must have $r_0^\flat := C''(\tau + r)\alpha^{-1}\left(\sqrt{(d+t)/n} + \epsilon\right) < r$, which implies that $\eta = 1$ and $\widehat{\boldsymbol{\theta}}_{\tau,\eta}^\flat = \widehat{\boldsymbol{\theta}}_\tau^\flat$, and thus, the above bound $r_0^\flat$ holds for $\widehat{\boldsymbol{\theta}}_\tau^\flat$. This proves the stated result.

Proof of Theorem 2.3.2 (ii): Recall that $\xi^\flat(\boldsymbol{\theta}) = \mathbf{S}^{-1/2}\left(\nabla L_\tau^\flat(\boldsymbol{\theta}) - \nabla L_\tau(\boldsymbol{\theta})\right)$ as defined in (A.2.16). Then $\xi^\flat(\boldsymbol{\theta}_\tau) = \frac{1}{n}\sum_{i=1}^n \ell_\tau'(\varepsilon_i - \alpha_\tau)e_i\boldsymbol{z}_i$. Denote that $\widetilde{\mathbf{S}} = \mathsf{E}\ell''(\varepsilon - \alpha_\tau)\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\mathsf{T}$. We first show that with probability at least $1 - 8e^{-t}$,

$$\mathsf{P}\left(\left\|\mathbf{S}^{-1/2}\widetilde{\mathbf{S}}(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau) + \xi^\flat(\boldsymbol{\theta}_\tau)\right\|_2 \le C\left(\frac{d+t}{n} + \epsilon^2\right)\right) \ge 1 - 6e^{-t}. \tag{A.2.21}$$

for some constant C independent of (d, t, n, ϵ). In fact, noting that $\nabla L_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) = 0$ and $\nabla L_\tau(\widehat{\boldsymbol{\theta}}_\tau) = 0$, we have

$$\xi^\flat(\widehat{\boldsymbol{\theta}}_\tau) = \mathbf{S}^{1/2}\left(\nabla L_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) \quad \nabla L_\tau(\widehat{\boldsymbol{\theta}}_\tau)\right) = \mathbf{S}^{1/2}\nabla L_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) = \mathbf{S}^{1/2}\left(\nabla L_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) \quad \nabla L_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)\right).$$

Therefore,

$$\begin{aligned}
&\left\|\mathbf{S}^{1/2}\widetilde{\mathbf{S}}(\widehat{\boldsymbol{\theta}}_\tau^\flat \quad \widehat{\boldsymbol{\theta}}_\tau) + \xi^\flat(\boldsymbol{\theta}_\tau)\right\|_2 \\
&\left\|\mathbf{S}^{1/2}\left\{\widetilde{\mathbf{S}}(\widehat{\boldsymbol{\theta}}_\tau^\flat \quad \widehat{\boldsymbol{\theta}}_\tau) + \nabla L_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) \quad \nabla L_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)\right\}\right\|_2 + \left\|\xi^\flat(\widehat{\boldsymbol{\theta}}_\tau) \quad \xi^\flat(\boldsymbol{\theta}_\tau)\right\|_2 \\
&\left\|\mathbf{S}^{1/2}\left\{\widetilde{\mathbf{S}}(\widehat{\boldsymbol{\theta}}_\tau^\flat \quad \widehat{\boldsymbol{\theta}}_\tau) + \nabla L_\tau(\widehat{\boldsymbol{\theta}}_\tau) \quad \nabla L_\tau(\widehat{\boldsymbol{\theta}}_\tau^\flat)\right\}\right\|_2 \\
&+ \left\{\left\|\xi^\flat(\widehat{\boldsymbol{\theta}}_\tau) \quad \xi^\flat(\boldsymbol{\theta}_\tau)\right\|_2 + \left\|\xi^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) \quad \xi^\flat(\widehat{\boldsymbol{\theta}}_\tau)\right\|_2\right\} \\
:=&\Gamma_1 + \Gamma_2.
\end{aligned}$$

Denote that $\Delta(\boldsymbol{\theta}) := \|\mathbf{S}^{1/2}\widetilde{\mathbf{S}}(\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau) + \nabla L_\tau(\boldsymbol{\theta}) \quad \nabla L_\tau(\boldsymbol{\theta}_\tau)\|_2$. Then clearly, we have $\Gamma_1 \quad \Delta(\widehat{\boldsymbol{\theta}}_\tau^\flat) + \Delta(\widehat{\boldsymbol{\theta}}_\tau)$. By Lemma A.2.1 (A.2.11), we know that with probability at least $1 \quad e^{\ t}$,

$$\sup_{\boldsymbol{\theta} 2 \boldsymbol{\Theta}_r} \Delta(\boldsymbol{\theta}) \quad C_3\left(r^2 + \epsilon r + r\sqrt{\frac{d+t}{n}}\right)$$

Then using the bound of $\|\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau\|_{\mathbf{S}}$ and $\|\widehat{\boldsymbol{\theta}}_\tau^\flat \quad \boldsymbol{\theta}_\tau\|_{\mathbf{S}}$, given by Theorem 2.2.1 and 2.3.2 (i), we have

$$\mathsf{P}\left(\Gamma_1 \quad C^{\prime}\left(\frac{d+t}{n} + \epsilon^2\right)\right) \quad 1 \quad 3e^{\ t}$$

with probability at least $1 \quad 6e^{\ t}$, for some constant $C^{\prime}$ independent of (d, t, n, ϵ).

As for Γ_2, by Triangle inequality, we have $\Gamma_2 \quad 2\left\|\xi^\flat(\widehat{\boldsymbol{\theta}}_\tau) \quad \xi^\flat(\boldsymbol{\theta}_\tau)\right\|_2 + \left\|\xi^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) \quad \xi^\flat(\boldsymbol{\theta}_\tau)\right\|_2$. Then by applying Lemma A.2.2 (i), with $r = \max\{r_0, r_0^\flat\}$ in (A.2.17), where $r_0, r_0^\flat$ are the bounds of $\|\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau\|_{\mathbf{S}}$ and $\|\widehat{\boldsymbol{\theta}}_\tau^\flat \quad \boldsymbol{\theta}_\tau\|_{\mathbf{S}}$, given respectively by Theorem 2.2.1 and Theorem 2.3.2 (i), we have

$$\mathsf{P}\left(\Gamma_2 \quad C^{\prime\prime}\left(\frac{d+t}{n} + \epsilon^2\right)\right) \quad 1 \quad 3e^{\ t}$$

with probability at least $1 - 8e^{-t}$, for some constant C'' independent of (d, t, n, ϵ). Combing the above bounds for Γ_1 and Γ_2, we have shown (A.2.21). Comparing (A.2.21) and the desired result (2.3.3), it suffices to show that with probability at least $1 - 6e^{-t}$,

$$\mathsf{P}\left(\left\|\mathbf{S}^{1/2}\left(\widetilde{\mathbf{S}} - c_F\mathbf{S}\right)(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau)\right\|_2 \le C\left(\frac{d+t}{n} + \epsilon^2\right)\right) \ge 1 - 2e^{-t}$$

In fact, by a exact same argument as (A.2.12) in the proof of Theorem 2.3.1 (just replacing $\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau$ by $\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau$), we can get

$$\left\|\mathbf{S}^{1/2}\left(\widetilde{\mathbf{S}} - c_F\mathbf{S}\right)(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau)\right\|_2 \le 2c_4\epsilon\|\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau\|_{\mathbf{S}}.$$

Notice that $\|\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau\|_{\mathbf{S}} \le \|\widehat{\boldsymbol{\theta}}_\tau^\flat - \boldsymbol{\theta}_\tau\|_{\mathbf{S}} + \|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau\|_{\mathbf{S}} \asymp \sqrt{(d+t)/n} + \epsilon$ with high probability. We are done. □

A.2.7 Proof of Lemma A.2.2

(i) Denote $\bar{B}^\flat(\boldsymbol{\delta}) := \xi^\flat(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau) - \xi^\flat(\boldsymbol{\theta}_\tau)$ with $\boldsymbol{\delta} := \mathbf{S}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau)$. Then $\bar{B}^\flat(\mathbf{0}) = \mathbf{0}$, $\mathsf{E}\,\bar{B}^\flat(\boldsymbol{\delta}) = \mathbf{0}, \forall\boldsymbol{\delta} \in \mathbb{R}^{d+1}$, where $\mathsf{E}(\cdot) = \mathsf{E}(\cdot|\mathcal{D}_n)$. It is easy to check that

$$\nabla_{\boldsymbol{\delta}}\bar{B}^\flat(\boldsymbol{\delta}) = \frac{1}{n}\sum_{i=1}^n e_i\ell_\tau''(\varepsilon_i - \alpha_\tau - \boldsymbol{z}_i^\intercal\boldsymbol{\delta})\boldsymbol{z}_i\boldsymbol{z}_i^\intercal.$$

By the sub-Gaussian property of $\{e_i\}_{i=1}^n$, we know that there exists a universal constant C_0 such that for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{S}^d$ and $\lambda \in \mathbb{R}$,

$$\begin{aligned}&\mathsf{E}\,\exp\left\{\lambda\sqrt{n}\boldsymbol{u}^\intercal\nabla_{\boldsymbol{\delta}}\bar{B}^\flat(\boldsymbol{\delta})\boldsymbol{v}\right\}\\ &\prod_{i=1}^n\exp\left\{\frac{c_4^2\lambda^2}{n}C_0^2\nu_e^2(\boldsymbol{u}^\intercal\boldsymbol{z}_i)^2(\boldsymbol{v}^\intercal\boldsymbol{z}_i)^2\right\} \le \exp\{C^2M_4(\boldsymbol{z})\lambda^2/2\},\end{aligned}$$

where $C = c_4C_0\nu_e$, $M_4(\boldsymbol{z}) := \sup_{\boldsymbol{u}\in\mathbb{S}^d}\frac{1}{n}\sum_{i=1}^n(\boldsymbol{u}^\intercal\boldsymbol{z}_i)^4$. Then by a conditional version of Theorem A.3 in Spokoiny (2013), we have

$$\mathsf{P}\left(\sup_{\boldsymbol{\delta}\in\mathbb{B}^{d+1}(r)}\|\sqrt{n}\bar{B}^\flat(\boldsymbol{\delta})\|_2 \le 6C\sqrt{M_4(\boldsymbol{z})}r\sqrt{4(d+1)+2t}\right) \ge 1 - e^{-t}, \tag{A.2.22}$$

for any $t \ge 1/2$. Then it remains to show that $M_4(\boldsymbol{z}) \le C'\nu_{\boldsymbol{x}}^4$ with probability at least $1 - e^{-t}$ for some absolute constant C'. It is easy to check that $(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{u}^\intercal\boldsymbol{z}_i)^4)^{1/4}$

$(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{v}^\intercal \boldsymbol{z}_i)^4)^{1/4} + \boldsymbol{k}\boldsymbol{u} \quad \boldsymbol{v}\boldsymbol{k}_2(M_4(\boldsymbol{z}))^{1/4}$ for any $\boldsymbol{u}, \boldsymbol{v} \ 2 \ \mathsf{S}^d$. Therefore, if we take a (1/2) -net $\boldsymbol{N}_{1/2}$ of S^d with $\left|\boldsymbol{N}_{1/2}\right| \quad 5^{d+1}$, we will get

$$M_4(\boldsymbol{z}) \quad 2^4 \max_{\boldsymbol{u}2\boldsymbol{N}_{1/2}} \frac{1}{n}\sum_{i=1}^n(\boldsymbol{u}^\intercal \boldsymbol{z}_i)^4.$$

As $\boldsymbol{z}$ is a sub-Gaussian vector with parameter $\nu_{\boldsymbol{x}}$, we know $\sup_{\boldsymbol{u}2\mathsf{S}^d}\mathsf{E}(\boldsymbol{u}^\intercal \boldsymbol{z})^4 \quad C_0^4\nu_{\boldsymbol{x}}^4$ and

$$\mathsf{E}\exp\left\{\frac{1}{4C_0^2\nu_{\boldsymbol{x}}^2}\boldsymbol{j}(\boldsymbol{u}^\intercal \boldsymbol{z})^4 \quad \mathsf{E}(\boldsymbol{u}^\intercal \boldsymbol{z})^4\boldsymbol{j}^{1/2}\right\} \quad \mathsf{E}\exp\left\{\frac{1}{4C_0^2\nu_{\boldsymbol{x}}^2}\left[(\boldsymbol{u}^\intercal \boldsymbol{z})^2 + (\mathsf{E}(\boldsymbol{u}^\intercal \boldsymbol{z})^4)^{1/2}\right]\right\}$$
$$e^{1/4}\left(\mathsf{E}\exp\left\{\frac{(\boldsymbol{u}^\intercal \boldsymbol{z})^2}{\nu_{\boldsymbol{x}}^2}\right\}\right)^{1/4} \quad (2e)^{1/4} \quad 2,$$

assuming that $C_0 \quad 1$. This implies that $\boldsymbol{k}(\boldsymbol{u}^\intercal \boldsymbol{z})^4 \quad \mathsf{E}(\boldsymbol{u}^\intercal \boldsymbol{z})^4\boldsymbol{k}_{\Psi_{1/2}} \quad 16C_0^4\nu_{\boldsymbol{x}}^4$ for any $\boldsymbol{u} \ 2 \ \mathsf{S}^d$, where the $\Psi_{1/2}$-norm is defined as $\boldsymbol{k}Y\boldsymbol{k}_{\Psi_r} = \inf\boldsymbol{f}C > 0 : \mathsf{E}\exp(\boldsymbol{j}Y\boldsymbol{j}/C)^r \quad 2\boldsymbol{g}$. Then by (3.6) of Adamczak et al. (2011) and taking a union bound, we obtain

$$\mathsf{P}\left(\max_{\boldsymbol{u}2\boldsymbol{N}_{1/2}}\frac{1}{n}\sum_{i=1}^n[(\boldsymbol{u}^\intercal \boldsymbol{z}_i)^4 \quad \mathsf{E}(\boldsymbol{u}^\intercal \boldsymbol{z}_i)^4] \quad t\right) \quad \exp\left\{(d+1)\log 5 \quad c\min\left[\frac{nt^2}{b^2}, \left(\frac{nt}{b}\right)^{1/2}\right]\right\},$$

where $b := \boldsymbol{k}(\boldsymbol{u}^\intercal \boldsymbol{z})^4 \quad \mathsf{E}(\boldsymbol{u}^\intercal \boldsymbol{z})^4\boldsymbol{k}_{\Psi_{1/2}} \quad 16C_0^4\nu_{\boldsymbol{x}}^4$ and c is a univeral constant. This further implies that with probability at least $1 \quad 2e^{\ t}$,

$$2^{\ 4}M_4(\boldsymbol{z}) \quad \sup_{\boldsymbol{u}2\mathsf{S}^d}\mathsf{E}(\boldsymbol{u}^\intercal \boldsymbol{z})^4 + c^{\prime}\nu_{\boldsymbol{x}}^4\left(\sqrt{\frac{d+t}{n}} + \frac{(d+t)^2}{n}\right) \quad c^{\prime\prime}\nu_{\boldsymbol{x}}^4, \tag{A.2.23}$$

for some (new) universal constant $c^{\prime}, c^{\prime\prime}$, as long as $n \quad (d+t)^2$.

(ii) Noting that $\xi^\flat(\boldsymbol{\theta}_\tau) = \frac{1}{n}\sum_{i=1}^n e_i\ell_\tau^{\prime}(\varepsilon_i \quad \alpha_\tau)(\quad \boldsymbol{z}_i)$ and e_i is sub-Gaussian with parameter ν_e, by Hoeffding's inequality and a standard ϵ-net argument, we know

$$\mathsf{P}\left(\boldsymbol{k}\xi^\flat(\boldsymbol{\theta}_\tau)\boldsymbol{k}_2 \quad Cc_1\tau\nu_e\sqrt{M_2(\boldsymbol{z})}\sqrt{\frac{d+t}{n}}\right) \quad 1 \quad e^{\ t},$$

where C is some absolute constant and c_1 is the bound of $\boldsymbol{j}\ell^{\prime}(\ )\boldsymbol{j}$ in Condition 2.2.1, and $M_2(\boldsymbol{z}) := \sup_{\boldsymbol{u}2\mathsf{S}^d}\frac{1}{n}\sum_{i=1}^n(\boldsymbol{u}^\intercal \boldsymbol{z}_i)^2$. Then we just need to show that $M_2(\boldsymbol{z}) \quad C^{\prime}\nu_{\boldsymbol{x}}^2$ with probability at least $1 \quad e^{\ t}$ for some absolute constant $C^{\prime}$. As $\boldsymbol{z}$ is a sub-Gaussian vector

with parameter $\nu_{\boldsymbol{x}}$, we know that $\sup_{\boldsymbol{u}2\mathsf{S}^d} \mathsf{E}(\boldsymbol{u}^{\mathsf{I}}\boldsymbol{z})^2 \quad C_0^2\nu_{\boldsymbol{x}}^2$ and $(\boldsymbol{u}^{\mathsf{I}}\boldsymbol{z})^2$ is sub-exponential. Then by Bernstein's inequality and a standard ϵ-net argument, we get

$$\mathsf{P}\left(\max_{\boldsymbol{u}2N_{1/2}} \frac{1}{n}\sum_{i=1}^{n}[(\boldsymbol{u}^{\mathsf{I}}\boldsymbol{z}_i)^2 \quad \mathsf{E}(\boldsymbol{u}^{\mathsf{I}}\boldsymbol{z}_i)^2] \quad t\right) \quad \exp\left\{(d+1)\log 5 \quad cn\min\left[\frac{t^2}{\nu_{\boldsymbol{x}}^4}, \frac{t}{\nu_{\boldsymbol{x}}^2}\right]\right\},$$

with some absolute constant c. Note that $M_2(\boldsymbol{z}) \quad 4\max_{\boldsymbol{u}2N_{1/2}} n^{\ 1}\sum_{i=1}^{n}(\boldsymbol{u}^{\mathsf{I}}\boldsymbol{z}_i)^2$. This implies that with probability at least $1 \quad e^{\ t}$,

$$M_2(\boldsymbol{z}) \quad 4\sup_{\boldsymbol{u}2\mathsf{S}^d} \mathsf{E}(\boldsymbol{u}^{\mathsf{I}}\boldsymbol{z})^2 + c^{\emptyset}\nu_{\boldsymbol{x}}^2\left(\sqrt{\frac{d+t}{n}} + \frac{d+t}{n}\right) \quad c^{\emptyset\emptyset}\nu_{\boldsymbol{x}}^2,$$

for some (new) universal constant $c^{\emptyset}, c^{\emptyset\emptyset}$, as long as $n \quad (d+t)$.

(iii) is a direct consequence of (i) and (ii). □

A.2.8 Proof of Theorem 2.3.3

For any $\boldsymbol{\mu}\ 2\ \mathsf{R}^d$, let $\boldsymbol{\lambda} = (0, \boldsymbol{\mu}^{\mathsf{I}})^{\mathsf{I}}$. Noting that $\boldsymbol{\mu}^{\mathsf{I}}(\widehat{\boldsymbol{\beta}}_\tau \quad \boldsymbol{\beta}\ ) = \boldsymbol{\lambda}^{\mathsf{I}}(\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau)$ and $\boldsymbol{\mu}^{\mathsf{I}}(\widehat{\boldsymbol{\beta}}_\tau^{\flat}$ $\widehat{\boldsymbol{\beta}}_\tau) = \boldsymbol{\lambda}^{\mathsf{I}}(\widehat{\boldsymbol{\theta}}_\tau^{\flat} \quad \widehat{\boldsymbol{\theta}}_\tau)$, it suffices to show that

$$\sup_{x2\mathsf{R}}\left|\mathsf{P}\left(\boldsymbol{\lambda}^{\mathsf{I}}(\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \quad x\right) \quad \mathsf{P}\left(\boldsymbol{\lambda}^{\mathsf{I}}(\widehat{\boldsymbol{\theta}}_\tau^{\flat} \quad \widehat{\boldsymbol{\theta}}_\tau) \quad x\right)\right| \quad C\left(\frac{d+t}{n} + \frac{p_-}{n\epsilon}\right) + 11e^{\ t}.$$

Define

$$S_n := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{c_F}\ell_\tau^{\emptyset}(\varepsilon_i \quad \alpha_\tau)\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S}^{\ 1/2}\boldsymbol{z}_i, \quad S_n^{\flat} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{c_F}\ell_\tau^{\emptyset}(\varepsilon_i \quad \alpha_\tau)\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S}^{\ 1/2}e_i\boldsymbol{z}_i$$

where $c_F = \mathsf{E}_{\varepsilon\ F}\ell_\tau^{\emptyset\emptyset}(\varepsilon \quad \alpha_\tau)$. By Theorem 2.3.1, we know

$$\left|\boldsymbol{\lambda}^{\mathsf{I}}(\widehat{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \quad S_n\right| \quad \left\|\frac{1}{c_F}\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S}^{\ 1/2}\right\|_2 \quad C_2\left(\frac{d+t}{n} + \epsilon^2\right) \tag{A.2.24}$$

with probability at least $1 \quad 3e^{\ t}$, for any sufficiently large n and small ϵ. Similarly, by Theorem 2.3.2 (ii), we know

$$\mathsf{P}\left(\left|\boldsymbol{\lambda}^{\mathsf{I}}(\widehat{\boldsymbol{\theta}}_\tau^{\flat} \quad \widehat{\boldsymbol{\theta}}_\tau) \quad S_n^{\flat}\right| \quad \left\|\frac{1}{c_F}\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S}^{\ 1/2}\right\|_2 \quad C_2^{\flat}\left(\frac{d+t}{n} + \epsilon^2\right)\right) \quad 1 \quad 8e^{\ t} \tag{A.2.25}$$

with probability at least $1 \quad 8e^{\ t}$, for any sufficiently large n and small ϵ.

Denote $U_i = c_F \ {}^1\ell_\tau^0(\varepsilon_i \quad \alpha_\tau)\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S} \ {}^{1/2}\boldsymbol{z}_i$. As S_n is in a re-scaled form of a sum of $i.i.d.$ variables U_i, and

$$\rho := \mathsf{E}\boldsymbol{j}U_i \quad \mathsf{E}U_i\boldsymbol{j}^3 \quad 2^3\mathsf{E}\boldsymbol{j}U_i\boldsymbol{j}^3 \quad 2^3\left(C_0c_1\tau\nu_{\boldsymbol{x}}\left\|\frac{1}{c_F}\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S} \ {}^{1/2}\right\|_2\right)^3 < \boldsymbol{1} \tag{A.2.26}$$

for some universal constant C_0. Then by Berry-Esseen Theorem (see e.g. Tyurin (2011)), we get

$$\sup_{x2\mathsf{R}}\left|\mathsf{P}\left(\frac{\boldsymbol{p}\overline{n}}{\sigma}(S_n \quad \mathsf{E}S_n) \quad x\right) \quad \Phi(x)\right| \quad \frac{\rho}{\sigma^3\boldsymbol{p}\overline{n}}, \tag{A.2.27}$$

where $\Phi(\ )$ is the CDF of standard normal distribution, and $\sigma^2 := \mathrm{Var}(U_i)$. It is easy to check that

$$\mathsf{E}U_i^2 \quad (1 \quad \epsilon)\mathsf{E}\left\{\frac{1}{c_F^2}(\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S} \ {}^{1/2}\boldsymbol{z})^2\mathsf{E}_{\varepsilon\boldsymbol{j}\boldsymbol{x}} \ {}_F[(\ell_\tau^0(\varepsilon \quad \alpha_\tau))^2\boldsymbol{j}\boldsymbol{x}]\right\} = (1 \quad \epsilon)\sigma_F^2\left\|\frac{1}{c_F}\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S} \ {}^{1/2}\right\|_2^2, \tag{A.2.28}$$

with $\sigma_F^2 := \mathsf{E}_\varepsilon \ {}_F\boldsymbol{f}(\ell_\tau^0(\varepsilon \quad \alpha_\tau))^2\boldsymbol{g}$, and

$$\begin{aligned}\boldsymbol{j}\mathsf{E}U_i\boldsymbol{j} &= \left|\mathsf{E}\left\{\frac{1}{c_F}\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S} \ {}^{1/2}\boldsymbol{z}\left((1 \quad \epsilon)\mathsf{E}_{\varepsilon\boldsymbol{j}\boldsymbol{x}} \ {}_F[\ell_\tau^0(\varepsilon \quad \alpha_\tau)\boldsymbol{j}\boldsymbol{x}] + \epsilon\mathsf{E}_{\varepsilon\boldsymbol{j}\boldsymbol{x}} \ {}_{G_{\boldsymbol{x}}}[\ell_\tau^0(\varepsilon \quad \alpha_\tau)\boldsymbol{j}\boldsymbol{x}]\right)\right\}\right| \\ &\overset{(\ )}{=} \left|\mathsf{E}\left\{\frac{1}{c_F}\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S} \ {}^{1/2}\boldsymbol{z}\left(\epsilon\mathsf{E}_{\varepsilon\boldsymbol{j}\boldsymbol{x}} \ {}_{G_{\boldsymbol{x}}}[\ell_\tau^0(\varepsilon \quad \alpha_\tau)\boldsymbol{j}\boldsymbol{x}]\right)\right\}\right| \\ &\quad c_1\tau\epsilon\left(\mathsf{E}\frac{1}{c_F^2}(\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S} \ {}^{1/2}\boldsymbol{z})^2\right)^{1/2} = \epsilon c_1\tau\left\|\frac{1}{c_F}\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S} \ {}^{1/2}\right\|_2,\end{aligned} \tag{A.2.29}$$

where $(\ )$ follows from the fact that F is independent of $\boldsymbol{x}$ and $\mathsf{E}_\varepsilon \ {}_F\ell_\tau^0(\varepsilon \quad \alpha_\tau) = 0$. Therefore, one can check that for small enough ϵ (any $\epsilon \quad c = c(c_1, \tau, c_F, \sigma_F)$), it holds

$$\frac{1}{4}\sigma_F^2\left\|\frac{1}{c_F}\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S} \ {}^{1/2}\right\|_2^2 \quad \sigma^2 \quad \mathsf{E}U_i^2 \quad c_1^2\tau^2\left\|\frac{1}{c_F}\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S} \ {}^{1/2}\right\|_2^2. \tag{A.2.30}$$

Then by (A.2.26)-(A.2.30), we have

$$\sup_{x2\mathsf{R}}\left|\mathsf{P}\left(S_n \quad x\right) \quad \Phi\left(\frac{\boldsymbol{p}\overline{n}}{\sigma}(x \quad \mathsf{E}U_1)\right)\right| \quad \frac{\rho}{\sigma^3\boldsymbol{p}\overline{n}} \quad C_3\frac{1}{\boldsymbol{p}\overline{n}}, \tag{A.2.31}$$

where $C_3 = {\sigma_F}^3(4C_0c_1\tau\nu_{\boldsymbol{x}})^3$. Similarly, for $S_n^\flat = \frac{1}{n}\sum_{i=1}^n e_iU_i$, let $\widehat{\rho} := \frac{1}{n}\sum_{i=1}^n \mathsf{E}\, |e_iU_i|^3 = \mathsf{E}|e|^3\frac{1}{n}\sum_{i=1}^n |U_i|^3 < 1$, and $\widehat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^n \mathsf{E}\, |e_iU_i|^2 = \frac{1}{n}\sum_{i=1}^n U_i^2$. Then by a conditional version of Berry-Esseen Theorem, we get

$$\sup_{x\in\mathsf{R}}\left|\mathsf{P}\left(\frac{\sqrt{n}}{\widehat{\sigma}}S_n^\flat \quad x\right) \quad \Phi(x)\right| \quad \frac{\widehat{\rho}}{\widehat{\sigma}^3\sqrt{n}}, \tag{A.2.32}$$

with probability 1. Then we use standard concentration inequality to show $|\widehat{\rho}$ $\mathsf{E}|e_i|^3\mathsf{E}|U_i|^3| \asymp 1/\sqrt{n}$ and $|\widehat{\sigma}^2 \quad \sigma^2| \asymp 1/\sqrt{n} + \epsilon$, both with high probability. In fact, as $\boldsymbol{z}$ is a sub-Gaussian vector with parameter $\nu_{\boldsymbol{x}}$, we know that U_i is sub-Gaussian with parameter $\nu_u := c_1\tau\nu_{\boldsymbol{x}}\|\widetilde{\boldsymbol{\lambda}}\|_2$, where $\widetilde{\boldsymbol{\lambda}}^\mathsf{T} := c_F\ {}^1\boldsymbol{\lambda}^\mathsf{T}\mathbf{S}^{1/2}$. This implies $\mathsf{E}|U_i|^3 \quad (C_0\nu_u)^3$, and

$$\mathsf{E}\exp\left\{\frac{1}{4C_0^2\nu_u^2}\left|U_i|^3 \quad \mathsf{E}|U_i|^3\right|^{2/3}\right\} \quad \mathsf{E}\exp\left\{\frac{1}{4C_0^2\nu_u^2}\left[|U_i|^2 + (\mathsf{E}|U_i|^3)^{2/3}\right]\right\}$$
$$e^{1/4}\left(\mathsf{E}\exp\left\{\frac{U_i^2}{\nu_u^2}\right\}\right)^{1/4} \quad (2e)^{1/4} \quad 2,$$

assuming that $C_0 \quad 1$. This further implies that $\||U_i|^3 \quad \mathsf{E}|U_i|^3\|_{\Psi_{2/3}} \quad (2C_0\nu_u)^3$ for any $\boldsymbol{u} \in \mathsf{S}^d$, where the $\Psi_{2/3}$-norm is defined as $\|Y\|_{\Psi_r} = \inf\{C > 0 : \mathsf{E}\exp(|Y|/C)^r \quad 2\}(r = 2/3)$. Then by (3.6) of Adamczak et al. (2011), we obtain

$$\mathsf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n[|U_i|^3 \quad \mathsf{E}|U_i|^3]\right| \quad t\right) \quad 2\exp\left\{ \quad c\min\left[\frac{nt^2}{b^2}, \left(\frac{nt}{b}\right)^{2/3}\right]\right\},$$

where $b = \||U_i|^3 \quad \mathsf{E}|U_i|^3\|_{\Psi_{2/3}} \quad (2C_0\nu_u)^3$ and c is some universal constant. Therefore, we obtain that with probability at least $1 \quad 2e^{\ t}$,

$$\widehat{\rho} \quad \mathsf{E}|e_1|^3\mathsf{E}|U_1|^3 + c^\prime(C_0^3\nu_u^3)\left(\sqrt{\frac{t}{n}} + \frac{t^{3/2}}{n}\right) \quad C^\prime\nu_e^3\nu_u^3 = C^\prime\nu_e^3(c_1\tau\nu_{\boldsymbol{x}}\|\widetilde{\boldsymbol{\lambda}}\|_2)^3 \tag{A.2.33}$$

as long as $n \quad t^2$, where $c^\prime, C^\prime$ are two absolute constants. Then we start to bound $|\widehat{\sigma}^2 \quad \sigma^2|$. Obviously, U_i^2 is sub-exponential with $\|U_i^2 \quad \mathsf{E}U_i^2\|_{\Psi_1} \quad C\|U_i^2\|_{\Psi_1} \quad Cc_1^2\tau^2\nu_{\boldsymbol{x}}^2\|\widetilde{\boldsymbol{\lambda}}\|_2^2$ for some universal constant C. Therefore, by Bernstein's inequality,

$$\mathsf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n[U_i^2 \quad \mathsf{E}U_i^2]\right| \quad t\right) \quad 2\exp\left\{ \quad cn\min\left[\frac{t^2}{B^2}, \frac{t}{B}\right]\right\}, \tag{A.2.34}$$

with some absolute constant c, and $B = \|U_i^2 - \mathsf{E}U_i^2\|_{\psi_1}$. Since $|\widehat{\sigma}^2 - \mathsf{E}U_1^2| = |EU_1|^2 \le (\epsilon c_1 \tau \|\widetilde{\lambda}\|_2)^2$ by (A.2.29), we get that with probability at least $1 - 2e^{-t}$,

$$|\widehat{\sigma}^2 - \sigma^2| \le c'(c_1\tau\|\widetilde{\lambda}\|_2)^2(\nu_{\boldsymbol{x}} - 1)^2\left(\epsilon^2 + \sqrt{\frac{t}{n}} + \frac{t}{n}\right) \le C_4\|\widetilde{\lambda}\|_2^2\left(\epsilon^2 + \sqrt{\frac{t}{n}}\right), \tag{A.2.35}$$

as long as $n \ge t$, where c' is a universal constant, and $C_4 := 2c'c_1^2\tau^2(\nu_{\boldsymbol{x}} - 1)^2$. Therefore, by (A.2.30), (A.2.33) and (A.2.35), we get an upper bound of $\widehat{\rho}$ and a lower bound of $\widehat{\sigma}$ with high probability, then we can bound the right side of (A.2.32) and get

$$\sup_{x\in\mathsf{R}}\left|\mathsf{P}\left(S_n^\flat \le x\right) - \Phi\left(\frac{\sqrt{n}}{\widehat{\sigma}}x\right)\right| \le C_5\frac{1}{\sqrt{n}}, \tag{A.2.36}$$

with probability at least $1 - 4e^{-t}$, where C_5 is a constant only depending on $(c_1, \tau, \nu_{\boldsymbol{x}}, \nu_e, \sigma_F)$. At last, we introduce Lemma A.7 in Spokoiny et al. (2015) to bound the Kolmogorov distance between two mean-zero normal distributions with variance σ_1^2 and σ_2^2:

$$\sup_{x\in\mathsf{R}}\left|\Phi\left(\frac{x}{\sigma_1}\right) - \Phi\left(\frac{x}{\sigma_2}\right)\right| \le \frac{1}{2}\left|\frac{\sigma_1^2}{\sigma_2^2} - 1\right|, \tag{A.2.37}$$

if $|\sigma_1^2/\sigma_2^2 - 1| \le 1/2$. By (A.2.30) and (A.2.35), we know $|\widehat{\sigma}^2/\sigma^2 - 1| \le C_6(\epsilon^2 + \sqrt{t/n}) \le 1/2$, for sufficiently large n and small ϵ, where $C_6 := 4C_4/\sigma_F^2 = 8c'c_1^2\tau^2(\nu_{\boldsymbol{x}} - 1)^2/\sigma_F^2$. Therefore, we know that with probability at least $1 - 2e^{-t}$

$$\sup_{x\in\mathsf{R}}\left|\Phi\left(\frac{x}{\widehat{\sigma}}\right) - \Phi\left(\frac{x}{\sigma}\right)\right| \le C_6\left(\epsilon^2 + \sqrt{\frac{t}{n}}\right). \tag{A.2.38}$$

Combing the above results, we finally get that for any $x \in \mathsf{R}$,

$$\begin{aligned}
&\mathsf{P}\left(\lambda^\dagger(\widehat{\theta}_\tau - \theta_\tau) \le x\right) \\
&\le \mathsf{P}\left(S_n \le x + C_2\|\widetilde{\lambda}\|_2\left(\frac{d+t}{n} + \epsilon^2\right)\right) + 3e^{-t} \quad \text{(by (A.2.24))} \\
&\le \Phi\left(\frac{\sqrt{n}}{\sigma}\left(x + C_2\|\widetilde{\lambda}\|_2\left(\frac{d+t}{n} + \epsilon^2\right) - \mathsf{E}U_1\right)\right) + \frac{C_3}{\sqrt{n}} + 3e^{-t} \quad \text{(by (A.2.31))} \\
&\le \Phi\left(\frac{\sqrt{n}}{\sigma}\left(x - C_2^\flat\|\widetilde{\lambda}\|_2\left(\frac{d+t}{n} + \epsilon^2\right)\right)\right) + \frac{C_3}{\sqrt{n}} + 3e^{-t} \\
&\quad + \frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\left((C_2 + C_2^\flat)\|\widetilde{\lambda}\|_2\left(\frac{d+t}{n} + \epsilon^2\right) + |\mathsf{E}U_1|\right)
\end{aligned}$$

where the last inequality is due to the anti-concentration inequality of standard normal: $|\Phi(a) - \Phi(b)| \le |b - a|/\sqrt{2\pi}$. Since $\sigma \le (\sigma_F/2)\|\widetilde{\lambda}\|_2$ by (A.2.30) and $|\mathbb{E}U_1| \le \epsilon c_1 \tau \|\widetilde{\lambda}\|_2$ by (A.2.29), we get

$$
\begin{aligned}
&\mathbb{P}\left(\langle \lambda, (\widehat{\theta}_\tau - \theta_\tau)\rangle \le x\right) \\
&\le \Phi\left(\frac{\sqrt{n}}{\sigma}\left(x + C_2^\flat \|\widetilde{\lambda}\|_2 \left(\frac{d+t}{n} + \epsilon^2\right)\right)\right) + \frac{C_3\sqrt{d}}{n} + C_2^\natural \sqrt{n}\left(\frac{d+t}{n} + \epsilon\right) + 3e^{-t} \\
&\le \Phi\left(\frac{\sqrt{n}}{\widehat{\sigma}}\left(x + C_2^\flat \|\widetilde{\lambda}\|_2 \left(\frac{d+t}{n} + \epsilon^2\right)\right)\right) + \frac{C_3\sqrt{d}}{n} + C_2^\natural \sqrt{n}\left(\frac{d+t}{n} + \epsilon\right) \\
&\qquad + C_6\left(\epsilon^2 + \sqrt{\frac{t}{n}}\right) + 3e^{-t} \quad \text{(w.p. } \ge 1 - 2e^{-t}\text{, by (A.2.38))} \\
&\le \mathbb{P}\left(S_n^\flat \le x + C_2^\flat \|\widetilde{\lambda}\|_2 \left(\frac{d+t}{n} + \epsilon^2\right)\right) + \frac{(C_3 + C_5)\sqrt{d}}{n} + C_2^\natural \sqrt{n}\left(\frac{d+t}{n} + \epsilon\right) \\
&\qquad + C_6\left(\epsilon^2 + \sqrt{\frac{t}{n}}\right) + 3e^{-t} \quad \text{(w.p. } \ge 1 - 6e^{-t}\text{, by (A.2.36))} \\
&\le \mathbb{P}\left(\langle \lambda, (\widehat{\theta}_\tau^\flat - \widehat{\theta}_\tau)\rangle \le x\right) + \frac{(C_3 + C_5)\sqrt{d}}{n} + C_2^\natural \sqrt{n}\left(\frac{d+t}{n} + \epsilon\right) \\
&\qquad + C_6\left(\epsilon^2 + \sqrt{\frac{t}{n}}\right) + 11e^{-t} \quad \text{(w.p. } \ge 1 - 14e^{-t}\text{, by (A.2.25))} \\
&\le \mathbb{P}\left(\langle \lambda, (\widehat{\theta}_\tau^\flat - \widehat{\theta}_\tau)\rangle \le x\right) + C\left(\frac{d+t}{\sqrt{n}} + \sqrt{n}\epsilon\right) + 11e^{-t} \quad \text{(w.p. } \ge 1 - 14e^{-t}\text{)},
\end{aligned}
$$

for some constant C independent of $(d, t, n, \epsilon, x, \lambda)$. By a similar argument, we can show the opposite way.

□

A.2.9 Proof of Theorem 2.3.4

By Theorem 2.3.3, we know there exists an event E_t satisfying $\mathbb{P}(E_t) \ge 1 - 14e^{-t}$, and on this event E_t, we have

$$
\sup_{x \in \mathbb{R}} \left|\mathbb{P}\left(\widehat{\beta}_j - \beta_j \le x\right) - \mathbb{P}\left(\widehat{\beta}_j^\flat - \widehat{\beta}_j \le x\right)\right| \le \Delta_t := C\left(\frac{d+t}{\sqrt{n}} + \sqrt{n}\epsilon\right) + 11e^{-t}.
$$

We first show that on this event E_t, for any $q \in (0,1)$, we always have

$$c_j(q - \Delta_t) \leq \hat{c}_j^{\flat}(q) \leq c_j(q + \Delta_t). \tag{A.2.39}$$

In fact, one can check that on the event E_t,

$$P\left(\hat{\beta}_j^{\flat} \leq \hat{\beta}_j - c_j(q + \Delta_t)\right) \leq P\left(\hat{\beta}_j - \beta_j \leq c_j(q + \Delta_t)\right) \leq \Delta_t \leq (q + \Delta_t) - \Delta_t = q.$$

Recall the definition of the q-quantile function $c_j^{\flat}(q)$, the above implies that $c_j(q+\Delta_t) \geq \hat{c}_j^{\flat}(q)$. Similarly, for any $\delta > 0$,

$$P\left(\hat{\beta}_j^{\flat} \geq \hat{\beta}_j - c_j(q - \Delta_t) - \delta\right) \geq P\left(\hat{\beta}_j - \beta_j \geq c_j(q - \Delta_t) - \delta\right) + \Delta_t < (q - \Delta_t) + \Delta_t = q,$$

which implies that $c_j(q - \Delta_t) - \delta < c_j^{\flat}(q)$, for any $\delta > 0$, and thus, $c_j(q - \Delta_t) \leq c_j^{\flat}(q)$. With (A.2.39) in hand, we start to derive the bound of $|P(\hat{\beta}_j - \beta_j \leq c_j^{\flat}(q)) - q|$. First, we have

$$\begin{aligned} P\left(\hat{\beta}_j - \beta_j \leq c_j^{\flat}(q)\right) - q &\leq P\left(\hat{\beta}_j - \beta_j \leq c_j(q - \Delta_t)\right) - P(E_t^c) - q \\ &\leq (q - \Delta_t) - 14e^{-t} - q = -\Delta_t - 14e^{-t}. \end{aligned}$$

Similarly,

$$\begin{aligned} P\left(\hat{\beta}_j - \beta_j \leq c_j^{\flat}(q)\right) - q &\geq P\left(\hat{\beta}_j - \beta_j \leq c_j(q + \Delta_t)\right) + P(E_t^c) - q \\ &\geq P\left(\hat{\beta}_j - \beta_j \leq c_j(q + \Delta_t) - \delta\right) + L(\delta) + P(E_t^c) - q \\ &> (q + \Delta_t) + L(\delta) + 14e^{-t} - q = \Delta_t + L(\delta) + 14e^{-t}, \quad \forall \delta > 0, \end{aligned}$$

where $L(\delta) := \sup_{x \in \mathbb{R}} P\left(\left|\hat{\beta}_j - \beta_j - x\right| \leq \delta\right)$ is the Lévy concentration function of the random variable $\hat{\beta}_j - \beta_j$. Here, we introduce $L(\delta)$ just to avoid the case that the CDF of $\hat{\beta}_j - \beta_j$ is not continuous at the point $c_j(q+\Delta_t)$. To prove the stated result in Theorem 2.3.4, it just remains to properly bound $L(\delta)$. This is achievable as by Bahadur representation of $\hat{\beta}_j - \beta_j$ (see Theorem 2.3.1), it is close to $S_n := \frac{1}{n}\sum_{i=1}^n U_i$ with $U_i = c_F^{-1}\ell_\tau'(\varepsilon_i - \alpha_\tau)\boldsymbol{\lambda}^\top \mathbf{S}^{-1/2}\boldsymbol{z}_i$, where $\boldsymbol{\lambda} = (0, \ldots, 0, 1, 0, \ldots, 0)$(the $(j+1)$-th component is 1), as defined in the proof of Theorem 2.3.3. And thus it can be approximated by a re-scaled normal random variable and we can

use the anti-concentration property of the normal distribution. Formally, by (A.2.24) and (A.2.31) in the proof of Theorem 2.3.3, for any $x \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{P}\left(\left|\widehat{\beta}_j - \beta_j - x\right| - \delta\right) &\leq \mathbb{P}\left(|S_n - x| - \delta + C_2\|\widetilde{\boldsymbol{\lambda}}\|_2\left(\frac{d+t}{n} + \epsilon^2\right)\right) + 3e^{-t} \\
&\leq \Phi\left(\frac{\sqrt{n}}{\sigma}\left\{x - \mathbb{E}U_1 + \delta + C_2\|\widetilde{\boldsymbol{\lambda}}\|_2\left(\frac{d+t}{n} + \epsilon^2\right)\right\}\right) \\
&\quad - \Phi\left(\frac{\sqrt{n}}{\sigma}\left\{x - \mathbb{E}U_1 - \delta - C_2\|\widetilde{\boldsymbol{\lambda}}\|_2\left(\frac{d+t}{n} + \epsilon^2\right)\right\}\right) + \frac{2C_3}{\sqrt{n}} + 3e^{-t} \\
&\leq \frac{2\sqrt{n}}{\sqrt{2\pi}\sigma}C_2\delta + \frac{2\sqrt{n}}{\sqrt{2\pi}\sigma}C_2\|\widetilde{\boldsymbol{\lambda}}\|_2\left(\frac{d+t}{n} + \epsilon^2\right) + \frac{2C_3}{\sqrt{n}} + 3e^{-t},
\end{aligned}$$

where $\widetilde{\boldsymbol{\lambda}} := c_F\,\|\mathbf{S}^{1/2}\boldsymbol{\lambda}\|$ and $\sigma^2 := \mathrm{Var}(U_1)$. Since $\sigma \geq (\sigma_F/2)\|\widetilde{\boldsymbol{\lambda}}\|_2$ by (A.2.30), where $\sigma_F^2 := \mathbb{E}_{\varepsilon\sim F}[(\ell_\tau'(\varepsilon - \alpha_\tau))^2]$, we then get

$$L(\delta) \leq \frac{2\sqrt{n}}{\sqrt{2\pi}\sigma}C_2\delta + C'\left(\frac{d+t}{\sqrt{n}} + \sqrt{n}\epsilon^2\right) + 3e^{-t}, \tag{A.2.40}$$

where C' is a constant decided by C_2, C_3 and σ_F. Combing all the above results, we finally obtain that for any $q \in (0,1)$,

$$\left|\mathbb{P}\left(\widehat{\beta}_j - \beta_j - c_j^{\flat}(q)\right) - q\right| \leq \Delta_t + \inf_{\delta>0} L(\delta) + 14e^{-t} \leq C\left(\frac{d+t}{\sqrt{n}} + \sqrt{n}\epsilon\right) + 28e^{-t}.$$

□

Appendix B Supplement to Chapter 3

B.1 Proof of Theorem 3.2.1

For any $r > 0$, we can find an intermediate estimator $\widetilde{\boldsymbol{\theta}}_{\tau,\eta} = \boldsymbol{\theta}_\tau + \eta(\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau)$ with some $\eta \in [0,1]$ such that $\widetilde{\boldsymbol{\theta}}_{\tau,\eta} \in \boldsymbol{\Theta}_r := \{\boldsymbol{\theta} \in \mathbb{R}^{d+1} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_\tau\|_{\mathbf{S}} \le r\}$. In fact, if $\widetilde{\boldsymbol{\theta}}_\tau \in \boldsymbol{\Theta}_r$, we may pick $\eta = 1$ and $\widetilde{\boldsymbol{\theta}}_{\tau,\eta} = \widetilde{\boldsymbol{\theta}}_\tau$; otherwise, let $\eta = r/\|\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau\|_{\mathbf{S}}$ and thus $\|\widetilde{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}_\tau\|_{\mathbf{S}} = r$. As $\widetilde{L}_\tau(\boldsymbol{\theta})$ is convex (since $L_{1,\tau}(\boldsymbol{\theta})$ is convex), by Lemma F.2 in Fan et al. (2018), we have

$$\begin{aligned}\left\langle r\widetilde{L}_\tau(\widetilde{\boldsymbol{\theta}}_{\tau,\eta}) - r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau), \widetilde{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\rangle &\le \eta \left\langle r\widetilde{L}_\tau(\widetilde{\boldsymbol{\theta}}_\tau) - r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau), \widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau \right\rangle \\ &= \eta \left\langle r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau), \widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau \right\rangle \\ &\le \left\|\mathbf{S}^{-1/2} r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau)\right\|_2 \left\|\widetilde{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}_\tau\right\|_{\mathbf{S}}.\end{aligned}$$

To bound $\|\mathbf{S}^{-1/2} r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau)\|_2$, we define $B_N(\boldsymbol{\theta})$ and $B_1(\boldsymbol{\theta})$ as follows:

$$B_N(\boldsymbol{\theta}) := \mathbf{S}^{-1/2}\left(rL_{N,\tau}(\boldsymbol{\theta}) - rL_{N,\tau}(\boldsymbol{\theta}_\tau)\right) - c_F\mathbf{S}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau), \tag{B.1.1}$$

$$B_1(\boldsymbol{\theta}) := \mathbf{S}^{-1/2}\left(rL_{1,\tau}(\boldsymbol{\theta}) - rL_{1,\tau}(\boldsymbol{\theta}_\tau)\right) - c_F\mathbf{S}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\tau), \tag{B.1.2}$$

where $c_F = \mathbb{E}_{\varepsilon - F}\ell''_\tau(\varepsilon - \alpha_\tau)$. By (A.2.11) and (A.2.12), we have

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \|B_N(\boldsymbol{\theta})\|_2 \le C_3' r\left(r + \epsilon + \sqrt{\frac{d+t}{N}}\right) \quad \text{and} \quad \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \|B_1(\boldsymbol{\theta})\|_2 \le C_3' r\left(r + \epsilon + \sqrt{\frac{d+t}{n}}\right) \tag{B.1.3}$$

with probability at least $1 - 2e^{-t}$, as long as $n \ge C_4(d+t)$ and $t \ge 1/2$, where C_3', C_4 are some constants depending only on $(c_4, \nu_{\boldsymbol{x}}, C_F, L, \tau)$. Therefore, we can get

$$\begin{aligned}&\left\|\mathbf{S}^{-1/2} r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau)\right\|_2 \\ &= \left\|\mathbf{S}^{-1/2}\left(rL_{1,\tau}(\boldsymbol{\theta}_\tau) - rL_{1,\tau}(\bar{\boldsymbol{\theta}}) + rL_{N,\tau}(\bar{\boldsymbol{\theta}}) - rL_{N,\tau}(\boldsymbol{\theta}_\tau) + rL_{N,\tau}(\boldsymbol{\theta}_\tau)\right)\right\|_2 \\ &\le \|B_1(\bar{\boldsymbol{\theta}})\|_2 + \|B_N(\bar{\boldsymbol{\theta}})\|_2 + \left\|\mathbf{S}^{-1/2} rL_{N,\tau}(\boldsymbol{\theta}_\tau)\right\|_2 \\ &\le 2C_3' r_0\left(r_0 + \epsilon + \sqrt{\frac{d+t}{n}}\right) + r\ ,\end{aligned} \tag{B.1.4}$$

on the events $E_0(r_0) \setminus E(r)$, with probability at least $1 - 2e^{-t}$. On the other hand, notice that $r\widetilde{L}_\tau(\widetilde{\boldsymbol{\theta}}_{\tau,\eta}) - r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau) = rL_{1,\tau}(\widetilde{\boldsymbol{\theta}}_{\tau,\eta}) - rL_{1,\tau}(\boldsymbol{\theta}_\tau)$. Then by (A.2.10), we know that with probability at least $1 - e^{-t}$,

$$\inf_{\boldsymbol{\theta}\in\Theta_r} h\, r\widetilde{L}_\tau(\boldsymbol{\theta}) - r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau), \boldsymbol{\theta} - \boldsymbol{\theta}_\tau i \geq \frac{c_2}{4}(1-\epsilon)\kappa_\tau k\boldsymbol{\theta} - \boldsymbol{\theta}_\tau k_{\mathbf{S}}^2, \tag{B.1.5}$$

as long as $\tau \leq (8C_0^2\nu_{\boldsymbol{x}}^2/c_3)r$ and $\epsilon/(1-\epsilon) \leq \kappa_\tau/4$ and $n \geq 64.5(1-\epsilon)^{-2}\kappa_\tau^{-2}[(c_3\tau/r)^2 \vee (c_3\tau/r)^4](d+t+1)$. Pick $r = c_3\tau/(8C_0^2\nu_{\boldsymbol{x}}^2)$, then for $\epsilon < c\kappa_\tau$ and $n \geq C'\kappa_\tau^{-2}(d+t)$, where $c, C' > 0$ are two constants depending only on $(\nu_{\boldsymbol{x}}, c_3)$, we obtain that on the events $E_0(r_0) \setminus E(r)$,

$$\|\widetilde{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}_\tau\|_{\mathbf{S}} \leq \frac{4}{c_2(1-\epsilon)\kappa_\tau}\left[2C_3' r_0\left(r_0 + \epsilon + \sqrt{\frac{d+t}{n}}\right) + r\right] := \widetilde{r}_0 \tag{B.1.6}$$

with probability at least $1 - 3e^{-t}$. For $r \geq r_0 \geq \sqrt{(d+t)/n} + \epsilon$, we would get

$$\widetilde{r}_0 \leq \left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r_0 + r \leq r_0 < r,$$

and thus, by the definition of $\widetilde{\boldsymbol{\theta}}_{\tau,\eta}$, we must have $\widetilde{\boldsymbol{\theta}}_{\tau,\eta} = \widetilde{\boldsymbol{\theta}}_\tau$. This proves the first result. As for the proof of second result, we observe that

$$\begin{aligned}
&\left\|c_F\mathbf{S}^{1/2}(\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau) - \mathbf{S}^{-1/2}\left(r\widetilde{L}_\tau(\widetilde{\boldsymbol{\theta}}_\tau) - r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau)\right)\right\|_2 \\
=&\left\|c_F\mathbf{S}^{1/2}(\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau) - \mathbf{S}^{-1/2}\left(rL_{1,\tau}(\widetilde{\boldsymbol{\theta}}_\tau) - rL_{1,\tau}(\boldsymbol{\theta}_\tau)\right)\right\|_2 = \left\|B_1(\widetilde{\boldsymbol{\theta}}_\tau)\right\|_2.
\end{aligned}$$

Also, using the fact that $r\widetilde{L}_\tau(\widetilde{\boldsymbol{\theta}}_\tau) = \mathbf{0}$, we can get

$$\begin{aligned}
&\left\|c_F\mathbf{S}^{1/2}(\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau) + \mathbf{S}^{-1/2}rL_{N,\tau}(\boldsymbol{\theta}_\tau)\right\|_2 \\
&\left\|c_F\mathbf{S}^{1/2}(\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau) + \mathbf{S}^{-1/2}r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau)\right\|_2 + \left\|\mathbf{S}^{-1/2}\left(r\widetilde{L}_\tau(\boldsymbol{\theta}_\tau) - rL_{N,\tau}(\boldsymbol{\theta}_\tau)\right)\right\|_2 \\
&\left\|B_1(\widetilde{\boldsymbol{\theta}}_\tau)\right\|_2 + \left\|B_1(\bar{\boldsymbol{\theta}}) - B_N(\bar{\boldsymbol{\theta}})\right\|_2 \\
&\sup_{\boldsymbol{\theta}\in\Theta_{\widetilde{r}_0}} kB_1(\boldsymbol{\theta})k_2 + \sup_{\boldsymbol{\theta}\in\Theta_{r_0}} kB_1(\boldsymbol{\theta})k_2 + \sup_{\boldsymbol{\theta}\in\Theta_{r_0}} kB_N(\boldsymbol{\theta})k_2 \\
&\leq \left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r_0,
\end{aligned}$$

with the same probability, under the condition that $r \geq r_0 \geq \sqrt{(d+t)/n} + \epsilon$. □

B.2 Proof of Theorem 3.2.2

First, by Theorem 3.2.1, we obtain that on the event $E_0(r_0) \setminus E(r)$, the estimator $\widetilde{\theta}_\tau^{(1)}$satisfies

$$\|\widetilde{\theta}_\tau^{(1)} - \theta_\tau\|_{\mathbf{S}} \leq C\left[\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r_0 + r\right] := r_1 \tag{B.2.1}$$

and

$$\left\| c_F \mathbf{S}^{1/2}(\widetilde{\theta}_\tau^{(1)} - \theta_\tau) - \frac{1}{N}\sum_{i=1}^N \ell'_\tau(\varepsilon_i - \alpha_\tau)\boldsymbol{z}_i \right\|_2 \leq C\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r_0$$

with probability at least $1 - 3e^{-t}$, where C is a constant independent of (N, n, d, ϵ, t). Then for sufficiently large n and small ϵ, we have $r \leq r_1 \leq r_0 \leq \sqrt{(d+t)/n} + \epsilon$. Also, (B.2.1) implies that $\widetilde{\theta}_\tau^{(1)} \in \Theta_{r_1}$ with probability at least $1 - 3e^{-t}$. Then again by Theorem 3.2.1, we get that on the event $E_1(r_1) \setminus E(r)$(where $E_1(r_1) := \{\widetilde{\theta}_\tau^{(1)} \in \Theta_{r_1}\}$), the estimator $\widetilde{\theta}_\tau^{(2)}$satisfies

$$\|\widetilde{\theta}_\tau^{(2)} - \theta_\tau\|_{\mathbf{S}} \leq C\left[\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r_1 + r\right] := r_2$$

and

$$\left\| c_F \mathbf{S}^{1/2}(\widetilde{\theta}_\tau^{(2)} - \theta_\tau) - \frac{1}{N}\sum_{i=1}^N \ell'_\tau(\varepsilon_i - \alpha_\tau)\boldsymbol{z}_i \right\|_2 \leq C\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r_1$$

with probability at least $1 - 3e^{-t}$. It is easy to see that $r \leq r_2 \leq r_0 \leq \sqrt{(d+t)/n} + \epsilon$. Therefore, by repeatedly using Theorem 3.2.1 at iteration $k = 3, \ldots, T$, we finally get that on the event $E_{T-1}(r_{T-1}) \setminus E(r)$(where $E_{T-1}(r_{T-1}) := \{\widetilde{\theta}_\tau^{(T-1)} \in \Theta_{r_{T-1}}\}$), the estimator $\widetilde{\theta}_\tau^{(T)}$satisfies

$$\|\widetilde{\theta}_\tau^{(T)} - \theta_\tau\|_{\mathbf{S}} \leq C\left[\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r_{T-1} + r\right] := r_T \tag{B.2.2}$$

and

$$\left\| c_F \mathbf{S}^{1/2}(\widetilde{\theta}_\tau^{(T)} - \theta_\tau) - \frac{1}{N}\sum_{i=1}^N \ell'_\tau(\varepsilon_i - \alpha_\tau)\boldsymbol{z}_i \right\|_2 \leq C\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) r_{T-1} \tag{B.2.3}$$

with probability at least $1 - 3e^{-t}$. By induction, it is easy to check that $r_T = \gamma^T r_0 + C\frac{1-\gamma^T}{1-\gamma} r$ with $\gamma := C(\sqrt{(d+t)/n} + \epsilon)$. For $T \geq \log(r/r_0)/\log(C(\sqrt{(d+t)/n} + \epsilon)) + 1$, we have $r_T \leq r_{T-1} \leq r$. Plugging this in (B.2.2) and (B.2.3), we show that the inequalities stated in Theorem 3.2.2 hold with probability at least $1 - 3Te^{-t}$ by taking a union bound for each iteration. □

B.3 Proof of Corollary 3.2.1

By Theorem 2.2.1, we can get that for the initial estimator $\widetilde{\boldsymbol{\theta}}_{\tau}^{(0)} = \bar{\boldsymbol{\theta}}$ 2 argmin$_{\boldsymbol{\theta} 2 \mathsf{R}^{d+1}} L_{1,\tau}(\boldsymbol{\theta})$, the event $E_0(r_0) = f\bar{\boldsymbol{\theta}}$ 2 $\Theta_{r_0}g$ holds with probability at least 1 2e ^{t}, with r_0 . $\sqrt{(d+t)/n}+$ ϵ. Also, by (A.2.4), we get that the event E $(r\ ) = fk\mathbf{S}$ $^{1/2}$ r $L_{N,\tau}(\boldsymbol{\theta}_{\tau})k_2$ r g holds with probability at least 1 e ^{t}, with r . $\sqrt{(d+t)/N}+\epsilon$. Then using the results from Theorem 3.2.2, we can immediately get (3.2.3) and (3.2.4). □

B.4 Proof of Theorem 3.3.1

Similar to what we did in the proof of Theorem 2.3.2, we start by defining the following key quantities which characterize the difference between the bootstrap losses and the original losses (for $\widetilde{L}_{\tau}^{\flat}(\boldsymbol{\theta})$ and $L_{1,\tau}^{\flat}(\boldsymbol{\theta})$, respectively):

$$\widetilde{\xi}^{\flat}(\boldsymbol{\theta}) := \mathbf{S}^{\ 1/2}\Big(r\,\widetilde{L}_{\tau}^{\flat}(\boldsymbol{\theta}) \quad r\,\widetilde{L}_{\tau}(\boldsymbol{\theta})\Big) \quad \text{and} \quad \xi_1^{\flat}(\boldsymbol{\theta}) := \mathbf{S}^{\ 1/2}\big(r\,L_{1,\tau}^{\flat}(\boldsymbol{\theta}) \quad r\,L_{1,\tau}(\boldsymbol{\theta})\big).$$

By the definition of $\widetilde{L}_{\tau}^{\flat}(\boldsymbol{\theta})$, we know that $\widetilde{\xi}^{\flat}(\boldsymbol{\theta}) = \xi_1^{\flat}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ 2 R^{d+1}. Then applying Lemma A.2.2 on $\xi_1^{\flat}(\boldsymbol{\theta})$, we can get the following properties about $\widetilde{L}_{\tau}^{\flat}(\boldsymbol{\theta})$ hold with probability at least 1 2e ^{t}:

$$\mathsf{P}\left(\sup_{\boldsymbol{\theta} 2 \Theta_r}\left\|\widetilde{\xi}^{\flat}(\boldsymbol{\theta}) \quad \widetilde{\xi}^{\flat}(\boldsymbol{\theta}_{\tau})\right\|_2 \quad Cr\sqrt{\frac{d+t}{n}}\right) \quad 1 \quad e^{\ t}, \tag{B.4.1}$$

and

$$\mathsf{P}\left(k\widetilde{\xi}^{\flat}(\boldsymbol{\theta}_{\tau})k_2 \quad C\tau\sqrt{\frac{d+t}{n}}\right) \quad 1 \quad e^{\ t}, \tag{B.4.2}$$

as long as n $(d+t)^2$, where C is a constant only depending on $(c_1, c_4, \nu_{\boldsymbol{x}}, \nu_e)$. Then we are ready to prove the stated results in Theorem 3.3.1.

Proof of (3.3.4)**:** We still first define an intermediate estimator for some pre-specified $r > 0$: $\widetilde{\boldsymbol{\theta}}_{\tau,\eta}^{\flat} = \boldsymbol{\theta}_{\tau} + \eta(\widetilde{\boldsymbol{\theta}}_{\tau}^{\flat} \quad \boldsymbol{\theta}_{\tau})$ with $\eta = \min f r/k\widetilde{\boldsymbol{\theta}}_{\tau}^{\flat} \quad \boldsymbol{\theta}_{\tau}k_{\mathbf{S}}, 1g$. That is, if $\widetilde{\boldsymbol{\theta}}_{\tau}^{\flat}$ 2 Θ_r, then $\eta = 1$ and $\widetilde{\boldsymbol{\theta}}_{\tau,\eta}^{\flat} = \widetilde{\boldsymbol{\theta}}_{\tau}^{\flat}$; otherwise, $\eta = r/k\widetilde{\boldsymbol{\theta}}_{\tau}^{\flat} \quad \boldsymbol{\theta}_{\tau}k_{\mathbf{S}}$ and $k\widetilde{\boldsymbol{\theta}}_{\tau,\eta}^{\flat} \quad \boldsymbol{\theta}_{\tau}k_{\mathbf{S}} = r$. Under this construction,

we always have $\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} \in \boldsymbol{\Theta}_r$. As $\widetilde{L}^{\flat}_{\tau}(\boldsymbol{\theta})$ is convex (since each $w_i \geq 0$ and $L^{\flat}_{1,\tau}(\boldsymbol{\theta})$ is convex), by Lemma F.2 in Fan et al. (2018), we have

$$\begin{aligned}\left\langle \nabla \widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta}) - \nabla \widetilde{L}^{\flat}_{\tau}(\boldsymbol{\theta}_\tau), \widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\rangle &\leq \eta \left\langle \nabla \widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau}) - \nabla \widetilde{L}^{\flat}_{\tau}(\boldsymbol{\theta}_\tau), \widetilde{\boldsymbol{\theta}}^{\flat}_{\tau} - \boldsymbol{\theta}_\tau \right\rangle \\ &= \eta \left\langle \nabla \widetilde{L}^{\flat}_{\tau}(\boldsymbol{\theta}_\tau), \widetilde{\boldsymbol{\theta}}_{\tau} - \boldsymbol{\theta}_\tau \right\rangle \\ &\quad \left(\left\| \mathbf{S}^{-1/2} \nabla \widetilde{L}_{\tau}(\boldsymbol{\theta}_\tau) \right\|_2 + \left\| \widetilde{\xi}^{\flat}(\boldsymbol{\theta}_\tau) \right\|_2 \right) \| \widehat{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}_\tau \|_{\mathbf{S}}.\end{aligned}$$

By (B.1.4) in the proof of Theorem 3.2.1, we know that

$$\left\| \mathbf{S}^{-1/2} \nabla \widetilde{L}_{\tau}(\boldsymbol{\theta}_\tau) \right\|_2 \leq 2C_3^{\prime} r_0 \left(r_0 + \epsilon + \sqrt{\frac{d+t}{n}} \right) + r \ ,$$

on the events $\mathcal{E}_0(r_0) \cap \mathcal{E}(r)$, with probability at least $1 - 2e^{-t}$, where $C_3^{\prime}$ is some constant depending only on $(c_4, \nu_{\boldsymbol{x}}, C_F, L, \tau)$. Recall that the initial estimator $\bar{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}^{(T-1)}_{\tau}$, then by the proof of Theorem 3.2.2 and Corollary 3.2.1, we know that the events $\mathcal{E}_0(r_0) \cap \mathcal{E}(r)$ hold with $r_0 \asymp r \asymp \sqrt{(d+t)/N} + \epsilon$ with probability at least $1 - 3(T+1)e^{-t}$. Then combing this fact with (B.4.2), we obtain that

$$\mathbb{P}\left(\left\langle \nabla \widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta}) - \nabla \widetilde{L}^{\flat}_{\tau}(\boldsymbol{\theta}_\tau), \widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\rangle \leq \widetilde{C} \left(\sqrt{\frac{d+t}{n}} + \epsilon \right) \left\| \widehat{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\|_{\mathbf{S}} \right) \geq 1 - e^{-t},$$

with probability at least $1 - (3T+7)e^{-t}$, for some constant $\widetilde{C}$ independent of (N, n, d, ϵ, t). On the other hand, notice that

$$\begin{aligned}\left\langle \nabla \widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta}) - \nabla \widetilde{L}^{\flat}_{\tau}(\boldsymbol{\theta}_\tau), \widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\rangle &\leq \left\langle \nabla \widetilde{L}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta}) - \nabla \widetilde{L}_{\tau}(\boldsymbol{\theta}_\tau), \widetilde{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\rangle \\ &\quad \left\| \widetilde{\xi}^{\flat}(\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta}) - \widetilde{\xi}^{\flat}(\boldsymbol{\theta}_\tau) \right\|_2 \left\| \widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\|_{\mathbf{S}}.\end{aligned}$$

By the local strong convexity of $\widetilde{L}_{\tau}(\boldsymbol{\theta})$ (B.1.5) and the inequality (B.4.1), we obtain that if we pick $r = c_3\tau/(8C_0^2\nu_{\boldsymbol{x}}^2)$, then with probability at least $1 - 3e^{-t}$,

$$\mathbb{P}\left(\left\langle \nabla \widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta}) - \nabla \widetilde{L}^{\flat}_{\tau}(\boldsymbol{\theta}_\tau), \widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\rangle \geq \alpha \left\| \widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\|^2_{\mathbf{S}} - Cr\sqrt{\frac{d+t}{n}} \left\| \widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} - \boldsymbol{\theta}_\tau \right\|_{\mathbf{S}} \right) \geq 1 - e^{-t},$$

as long as $\epsilon < c$ and $n \quad C^{\emptyset}(d+t)^2$, where $\alpha := \frac{c_2}{4}(1 \quad \epsilon)\kappa_\tau$ and $c, C, C^{\emptyset}$ are constants independent of (N, n, d, ϵ, t). Combing the above results together, we get that with probability at least $1 \quad (3T+10)e^{\ t}$,

$$\mathsf{P}\left(\left\|\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} \quad \boldsymbol{\theta}_\tau\right\|_{\mathbf{S}} \quad C^{\flat}\left(\sqrt{\frac{d+t}{n}}+\epsilon\right)\right) \quad 1 \quad 2e^{\ t},$$

for some constant $C^{\flat}$ independent of (N, n, d, ϵ, t). For sufficiently large n and small ϵ, we always have $C^{\flat}\left(\sqrt{(d+t)/n}+\epsilon\right) < r := c_3\tau/(8C_0^2\nu_{\boldsymbol{x}}^2)$, which indicates that $\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau,\eta} = \widetilde{\boldsymbol{\theta}}^{\flat}_{\tau}$ and thus the above property is also true for $\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau}$.

Proof of (3.3.5)**:** Using the fact that $\mathsf{r}\,\widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_{\tau}) = 0$ and $\mathsf{r}\,\widetilde{L}_{\tau}(\widetilde{\boldsymbol{\theta}}_{\tau}) = 0$, we can get

$$\widetilde{\xi}^{\flat}(\widetilde{\boldsymbol{\theta}}_\tau) = \mathbf{S}^{\ 1/2}\left(\mathsf{r}\,\widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}_\tau) \quad \mathsf{r}\,\widetilde{L}_{\tau}(\widetilde{\boldsymbol{\theta}}_\tau)\right) = \mathbf{S}^{\ 1/2}\,\mathsf{r}\,\widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}_\tau) = \mathbf{S}^{\ 1/2}\left(\mathsf{r}\,\widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}_\tau) \quad \mathsf{r}\,\widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau)\right).$$

Therefore,

$$\begin{aligned}
&\left\|c_F\mathbf{S}^{1/2}(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau \quad \widetilde{\boldsymbol{\theta}}_\tau) + \widetilde{\xi}^{\flat}(\boldsymbol{\theta}_\tau)\right\|_2 \\
&\quad \left\|c_F\mathbf{S}^{1/2}(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau \quad \widetilde{\boldsymbol{\theta}}_\tau) + \mathbf{S}^{\ 1/2}\left(\mathsf{r}\,\widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}_\tau) \quad \mathsf{r}\,\widetilde{L}^{\flat}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau)\right)\right\|_2 + \left\|\widetilde{\xi}^{\flat}(\widetilde{\boldsymbol{\theta}}_\tau) \quad \widetilde{\xi}^{\flat}(\boldsymbol{\theta}_\tau)\right\|_2 \\
&\quad \left\|c_F\mathbf{S}^{1/2}(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau \quad \widetilde{\boldsymbol{\theta}}_\tau) + \mathbf{S}^{\ 1/2}\left(\mathsf{r}\,\widetilde{L}_{\tau}(\widetilde{\boldsymbol{\theta}}_\tau) \quad \mathsf{r}\,\widetilde{L}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau)\right)\right\|_2 \\
&\quad + \left(\left\|\widetilde{\xi}^{\flat}(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau) \quad \widetilde{\xi}^{\flat}(\widetilde{\boldsymbol{\theta}}_\tau)\right\|_2 + \left\|\widetilde{\xi}^{\flat}(\widetilde{\boldsymbol{\theta}}_\tau) \quad \widetilde{\xi}^{\flat}(\boldsymbol{\theta}_\tau)\right\|_2\right) \\
&:= \Gamma_1 + \Gamma_2.
\end{aligned}$$

Notice that $\mathsf{r}\,\widetilde{L}_{\tau}(\widetilde{\boldsymbol{\theta}}_\tau) \quad \mathsf{r}\,\widetilde{L}_{\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau) = \mathsf{r}\,L_{1,\tau}(\widetilde{\boldsymbol{\theta}}_\tau) \quad \mathsf{r}\,L_{1,\tau}(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau)$. And recall the definition of $B_1(\boldsymbol{\theta})$ (B.1.2) and the bound of $B_1(\boldsymbol{\theta})$ (B.1.3), then we can get

$$\Gamma_1 \quad \left\|B_1(\widetilde{\boldsymbol{\theta}}_\tau)\right\|_2 + \left\|B_1(\widetilde{\boldsymbol{\theta}}^{\flat}_\tau)\right\|_2 \quad C_3^{\emptyset} r\left(r+\epsilon+\sqrt{\frac{d+t}{n}}\right)$$

with probability at least $1 \quad 2e^{\ t}$, as long as $n \quad C_4(d+t)$ and $t \quad 1/2$, where $C_3^{\emptyset}, C_4$ are some constants depending only on $(c_4, \nu_{\boldsymbol{x}}, C_F, L, \tau)$, and r is an upper bound for $\mathsf{k}\widetilde{\boldsymbol{\theta}}_\tau$ $\boldsymbol{\theta}_\tau\mathsf{k}_{\mathbf{S}} \_ \mathsf{k}\widetilde{\boldsymbol{\theta}}^{\flat}_\tau \quad \boldsymbol{\theta}_\tau\mathsf{k}_{\mathbf{S}}$ to be specified. Using the bound (3.3.4) for $\widetilde{\boldsymbol{\theta}}^{\flat}_\tau$ and the bound (3.2.3) for $\widetilde{\boldsymbol{\theta}}_\tau = \widetilde{\boldsymbol{\theta}}^{(T)}_\tau$ in Corollary 3.2.1, we get

$$\mathsf{P}\left(\Gamma_1 \quad \widetilde{C}^{\emptyset}\left(\frac{d+t}{n}+\epsilon^2\right)\right) \quad 1 \quad 2e^{\ t}$$

with probability at least 1 $(3T+12)e^{-t}$, for some constant $\widetilde{C}^{(0)}$ independent of (N, n, d, ϵ, t). As for Γ_2, note that $\Gamma_2 \le 2\left\|\widetilde{\xi}^{(b)}(\widetilde{\boldsymbol{\theta}}_\tau) - \widetilde{\xi}^{(b)}(\boldsymbol{\theta}_\tau)\right\|_2 + \left\|\widetilde{\xi}^{(b)}(\widetilde{\boldsymbol{\theta}}^{(b)}_\tau) - \widetilde{\xi}^{(b)}(\boldsymbol{\theta}_\tau)\right\|_2$. Then by (B.4.1) with r as an upper bound for $\|\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau\|_{\mathbf{S}} \vee \|\widetilde{\boldsymbol{\theta}}^{(b)}_\tau - \boldsymbol{\theta}_\tau\|_{\mathbf{S}}$, we can similarly get

$$\mathsf{P}\left(\Gamma_2 \le \widetilde{C}^{(00)}\left(\frac{d+t}{n} + \epsilon^2\right)\right) \ge 1 - 3e^{-t}$$

with probability at least 1 $(3T+14)e^{-t}$, for some constant $\widetilde{C}^{(00)}$ independent of (N, n, d, ϵ, t). Combing the above results and using the fact that $\widetilde{\xi}^{(b)}(\boldsymbol{\theta}_\tau) = \xi^{(b)}_1(\boldsymbol{\theta}_\tau) = \frac{1}{n}\sum_{i \in \mathcal{I}_1} \ell^{(1)}_\tau(\varepsilon_i - \alpha_\tau) e_i \boldsymbol{z}_i$, we prove (3.3.5). □

B.5 Proof of Theorem 3.3.2

We employ a strategy akin to the one we used to prove Theorem 2.3.3. For any $\boldsymbol{\mu} \in \mathsf{R}^d$, let $\boldsymbol{\lambda} = (0, \boldsymbol{\mu}^\intercal)^\intercal$. Then it suffices to show that

$$\sup_{x \in \mathsf{R}} \left|\mathsf{P}\left(\boldsymbol{\lambda}^\intercal(\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau) \le x\right) - \mathsf{P}\left(\frac{1}{\sqrt{m}}\boldsymbol{\lambda}^\intercal(\widetilde{\boldsymbol{\theta}}^{(b)}_\tau - \widetilde{\boldsymbol{\theta}}_\tau) \le x\right)\right| \le C\left(\frac{d+t}{\sqrt{n}} + \sqrt{N}\epsilon\right) + (3T+19)e^{-t}.$$

Define

$$S_N := \frac{1}{N}\sum_{i=1}^{N} \frac{1}{c_F}\ell^{(1)}_\tau(\varepsilon_i - \alpha_\tau)\boldsymbol{\lambda}^\intercal \boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}_i, \quad S^{(b)}_1 := \frac{1}{n}\sum_{i \in \mathcal{I}_1} \frac{1}{c_F}\ell^{(1)}_\tau(\varepsilon_i - \alpha_\tau)\boldsymbol{\lambda}^\intercal \boldsymbol{\Sigma}^{-1/2} e_i \boldsymbol{z}_i$$

where $c_F = \mathsf{E}_{\varepsilon \sim F}\ell^{(10)}_\tau(\varepsilon - \alpha_\tau)$. By the Bahadur representations of $\widetilde{\boldsymbol{\theta}}_\tau$ (3.2.4) and $\widetilde{\boldsymbol{\theta}}^{(b)}_\tau$ (3.3.5), we know

$$\left|\boldsymbol{\lambda}^\intercal(\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau) - S_N\right| \le \|\widetilde{\boldsymbol{\lambda}}\|_2 \cdot \widetilde{C}\left(\sqrt{\frac{d+t}{N}} + \epsilon\right)\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) \tag{B.5.1}$$

and

$$\mathsf{P}\left(\left|\boldsymbol{\lambda}^\intercal(\widetilde{\boldsymbol{\theta}}^{(b)}_\tau - \widetilde{\boldsymbol{\theta}}_\tau) - S^{(b)}_1\right| \le \|\widetilde{\boldsymbol{\lambda}}\|_2 \cdot C^{(b)}\left(\frac{d+t}{n} + \epsilon^2\right)\right) \ge 1 - 5e^{-t} \tag{B.5.2}$$

with probability at least 1 $(3T+14)e^{-t}$, where $\widetilde{\boldsymbol{\lambda}}^\intercal = c_F^{-1}\boldsymbol{\lambda}^\intercal \boldsymbol{\Sigma}^{-1/2}$, and $\widetilde{C}, C^{(b)}$ are constants independent of $(N, n, d, \epsilon, t, \boldsymbol{\lambda})$.

Denote $U_i := c_F^{-1}\ell_\tau''(\varepsilon_i - \alpha_\tau)\boldsymbol{\lambda}^\top \mathbf{S}^{-1/2}\boldsymbol{z}_i$, for $i = 1, \ldots, N$, and $\sigma^2 := \mathrm{Var}(U_i)$, $\widehat{\sigma}^2 := \frac{1}{n}\sum_{i \in I_1} \mathsf{E}\,|e_i U_i|^2 = \frac{1}{n}\sum_{i \in I_1} U_i^2$. By the normal approximation results about S_N (A.2.31) and $S_1^\flat$ (A.2.36) from the proof of Theorem 2.3.3, we know

$$\sup_{x \in \mathbb{R}} \left| \mathsf{P}\left(S_N - x\right) - \Phi\left(\frac{\sqrt{N}}{\sigma}(x - \mathsf{E}U_1)\right) \right| \le C_3 \frac{1}{\sqrt{N}}, \tag{B.5.3}$$

and

$$\sup_{x \in \mathbb{R}} \left| \mathsf{P}\left(S_1^\flat - x\right) - \Phi\left(\frac{\sqrt{n}}{\widehat{\sigma}}x\right) \right| \le C_5 \frac{1}{\sqrt{n}}, \tag{B.5.4}$$

with probability at least $1 - 4e^{-t}$, where C_3 and C_5 is a constant only depending on $(c_1, \tau, \nu_{\boldsymbol{x}}, \nu_e, \sigma_F)$. Denote

$$r_N := \widetilde{C}\left(\sqrt{\frac{d+t}{N}} + \epsilon\right)\left(\sqrt{\frac{d+t}{n}} + \epsilon\right), \quad \text{and} \quad r_n := C^\flat\left(\frac{d+t}{n} + \epsilon^2\right).$$

For any $\boldsymbol{\lambda} \in \mathbb{R}^{d+1}$ and any $x \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathsf{P}\left(\boldsymbol{\lambda}^\top(\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau) \le x\right) \\
&\quad \mathsf{P}\left(S_N \le x + \|\widetilde{\boldsymbol{\lambda}}\|_2 r_N\right) + (3T+14)e^{-t} \quad \text{(by (B.5.1))} \\
&\quad \Phi\left(\frac{\sqrt{N}}{\sigma}\left(x + \|\widetilde{\boldsymbol{\lambda}}\|_2 r_N - \mathsf{E}U_1\right)\right) + \frac{C_3}{\sqrt{N}} + (3T+14)e^{-t} \quad \text{(by (B.5.3))} \\
&\quad \Phi\left(\frac{\sqrt{n}}{\sigma}\left(\sqrt{m}x - \|\widetilde{\boldsymbol{\lambda}}\|_2 r_n\right)\right) + \frac{C_3}{\sqrt{N}} + (3T+14)e^{-t} \\
&\qquad + \frac{1}{\sqrt{2\pi}\sigma}\left(\sqrt{N}\left(\|\widetilde{\boldsymbol{\lambda}}\|_2 r_N + |\mathsf{E}U_1|\right) + \sqrt{n}\|\widetilde{\boldsymbol{\lambda}}\|_2 r_n\right),
\end{aligned}$$

where the last inequality is due to the anti-concentration inequality of standard normal: $|\Phi(a) - \Phi(b)| \le |b - a|/\sqrt{2\pi}$. Since $\sigma \ge (\sigma_F/2)\|\widetilde{\boldsymbol{\lambda}}\|_2$ by (A.2.30) and $|\mathsf{E}U_1| \le \epsilon c_1 \tau \|\widetilde{\boldsymbol{\lambda}}\|_2$ by (A.2.29), we get

$$\mathsf{P}\left(\boldsymbol{\lambda}^\top(\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau) \le x\right) \le \Phi\left(\frac{\sqrt{n}}{\sigma}\left(\sqrt{m}x - \|\widetilde{\boldsymbol{\lambda}}\|_2 r_n\right)\right) + (3T+14)e^{-t} + \widetilde{C}^\natural\left(\sqrt{N}\epsilon + \frac{d+t}{\sqrt{n}}\right),$$

for some constant $\widetilde{C}^{(0)}$ independent of $(N, n, d, \epsilon, t, \boldsymbol{\lambda}, x)$. By (A.2.38), (B.5.2) and (B.5.4), we get

$$
\begin{aligned}
&\Phi\left(\frac{\sqrt{n}}{\sigma}\left(\sqrt{m}x \quad k\widetilde{\boldsymbol{\lambda}}k_2 \quad r_n\right)\right) \\
&\quad \Phi\left(\frac{\sqrt{n}}{\widehat{\sigma}}\left(\sqrt{m}x \quad k\widetilde{\boldsymbol{\lambda}}k_2 \quad r_n\right)\right) + C_6\left(\epsilon^2 + \sqrt{\frac{t}{n}}\right) \quad \text{(w.p. } \quad 1 \quad 2e^{\quad t}\text{, by (A.2.38))} \\
&\quad \mathsf{P}\left(S_1^{\flat} \quad \sqrt{m}x \quad k\widetilde{\boldsymbol{\lambda}}k_2 \quad r_n\right) + C_5\frac{1}{\sqrt{n}} + C_6\left(\epsilon^2 + \sqrt{\frac{t}{n}}\right) \quad \text{(w.p. } \quad 1 \quad 6e^{\quad t}\text{, by (B.5.4))} \\
&\quad \mathsf{P}\left(\boldsymbol{\lambda}^{\mathsf{T}}(\widetilde{\boldsymbol{\theta}}_\tau^{\flat} \quad \widetilde{\boldsymbol{\theta}}_\tau) \quad \sqrt{m}x\right) + C_5\frac{1}{\sqrt{n}} + C_6\left(\epsilon^2 + \sqrt{\frac{t}{n}}\right) \\
&\quad + 5e^{\quad t} \text{ (w.p. } \quad 1 \quad (3T+20)e^{\quad t}\text{, by (B.5.2))}.
\end{aligned}
$$

Therefore, we finally obtain that

$$
\mathsf{P}\left(\boldsymbol{\lambda}^{\mathsf{T}}(\widetilde{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \quad x\right) \quad \mathsf{P}\left(\boldsymbol{\lambda}^{\mathsf{T}}(\widetilde{\boldsymbol{\theta}}_\tau^{\flat} \quad \widetilde{\boldsymbol{\theta}}_\tau) \quad \sqrt{m}x\right) + C\left(\sqrt{N}\epsilon + \frac{d+t}{\sqrt{n}}\right) + (3T+19)e^{\quad t}
$$

with probability at least $1 \quad (3T+20)e^{\quad t}$. By a similar argument, we can show the opposite direction of the above inequality, which completes the proof. □

B.6 Proof of Theorem 3.4.1

We begin with introducing the following lemma, which gives the approximation error of $\widetilde{D}_j$.

Lemma B.6.1. Under the same conditions of Theorem 3.4.1, for any $\boldsymbol{\lambda} \in \mathsf{R}^{d+1}$, we have

$$
\left|\left\|\frac{1}{c_F}\boldsymbol{\lambda}^{\mathsf{T}}\mathbf{S}^{\quad 1/2}\right\|_2 \quad \left\|\frac{1}{\widehat{c}_F}\boldsymbol{\lambda}^{\mathsf{T}}\widehat{\mathbf{S}}^{\quad 1/2}\right\|_2\right| \quad Ck\boldsymbol{\lambda}k_2\left(\sqrt{\frac{d+t}{n}} + \epsilon\right), \tag{B.6.1}
$$

with probability at least $1 \quad (3T+8)e^{\quad t}$, for some constant C independent of $(d, t, N, n, \epsilon, \boldsymbol{\lambda})$.

We leave the proof of this lemma at the end of this proof. Now we start to prove Theorem 3.4.1. We decompose it into three main steps.

Step 1: Following the proof of Theorem 3.3.2, denote

$$r_N := \widetilde{C}\left(\sqrt{\frac{d+t}{N}}+\epsilon\right)\left(\sqrt{\frac{d+t}{n}}+\epsilon\right) \quad \text{and} \quad r_n := C^\flat\left(\frac{d+t}{n}+\epsilon^2\right)$$

as the bound in (B.5.1) and (B.5.2) respectively. For any $\boldsymbol{\lambda}\ 2\ \mathsf{R}^{d+1}$, denote $\widetilde{\boldsymbol{\lambda}}^{\mathsf{I}} = c_F{}^{1}\boldsymbol{\lambda}^{\mathsf{I}}\mathbf{S}^{\ 1/2}$. By comparing the CDF of $\boldsymbol{\lambda}(\widetilde{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau)$ and $\boldsymbol{\lambda}(\widetilde{\boldsymbol{\theta}}^\flat_\tau \quad \widetilde{\boldsymbol{\theta}}_\tau)$, we can get

$$\begin{aligned}
&\mathsf{P}\left(\boldsymbol{\lambda}^{\mathsf{I}}(\widetilde{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \quad x\right)\\
&\quad \mathsf{P}\left(S_N \quad x+k\widetilde{\boldsymbol{\lambda}}k_2 \quad r_N\right)+(3T+14)e^{\ t} \quad \text{(by (B.5.1))}\\
&\quad \Phi\left(\frac{p\overline{N}}{\sigma}\left(x+k\widetilde{\boldsymbol{\lambda}}k_2 \quad r_N \quad \mathsf{E}U_1\right)\right)+\frac{C_3}{p\overline{N}}+(3T+14)e^{\ t} \quad \text{(by (B.5.3))}\\
&\quad \Phi\left(\frac{p\overline{N}}{\widehat{\sigma}}\left(x+k\widetilde{\boldsymbol{\lambda}}k_2 \quad r_N \quad \mathsf{E}U_1\right)\right)+\frac{C_3}{p\overline{N}}\\
&\qquad +C_6\left(\epsilon^2+\sqrt{\frac{t}{n}}\right)+(3T+14)e^{\ t} \quad \text{(w.p. } 1 \quad 2e^{\ t}\text{, by (A.2.38))}\\
&\quad \mathsf{P}\ \left(S^\flat_n \quad p\overline{m}\left(x+k\widetilde{\boldsymbol{\lambda}}k_2 \quad r_N \quad \mathsf{E}U_1\right)\right)+\frac{C_3}{p\overline{N}}+\frac{C_5}{p\overline{n}}\\
&\qquad +C_6\left(\epsilon^2+\sqrt{\frac{t}{n}}\right)+(3T+14)e^{\ t} \quad \text{(w.p. } 1 \quad 6e^{\ t}\text{, by (B.5.4))}\\
&\quad \mathsf{P}\ \left(\boldsymbol{\lambda}^{\mathsf{I}}(\widetilde{\boldsymbol{\theta}}^\flat_\tau \quad \widetilde{\boldsymbol{\theta}}_\tau) \quad p\overline{m}\left(x+k\widetilde{\boldsymbol{\lambda}}k_2 \quad r_N \quad \mathsf{E}U_1\right)+k\widetilde{\boldsymbol{\lambda}}k_2 \quad r_n\right)+\frac{C_3}{p\overline{N}}+\frac{C_5}{p\overline{n}}\\
&\qquad +C_6\left(\epsilon^2+\sqrt{\frac{t}{n}}\right)+(3T+19)e^{\ t} \quad \text{(w.p. } 1 \quad (3T+20)e^{\ t}\text{, by (B.5.2)).}
\end{aligned}$$

Denote

$$\Delta_t = \frac{C_3}{p\overline{N}}+\frac{C_5}{p\overline{n}}+C_6\left(\epsilon^2+\sqrt{\frac{t}{n}}\right)+(3T+19)e^{\ t}, \quad R_t = k\widetilde{\boldsymbol{\lambda}}k_2(r_N+r_n/p\overline{m}).$$

Then from the above derivation results, we know there exists some event E_t such that $\mathsf{P}(E_t)$ $1 \quad (3T+20)e^{\ t}$ and on this event, the following two inequalities hold simultaneously.

$$\begin{aligned}
&\mathsf{P}\left(\boldsymbol{\lambda}^{\mathsf{I}}(\widetilde{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \quad x\right) \quad \mathsf{P}\ \left(\frac{1}{p\overline{m}}\boldsymbol{\lambda}^{\mathsf{I}}(\widetilde{\boldsymbol{\theta}}^\flat_\tau \quad \widetilde{\boldsymbol{\theta}}_\tau) \quad x+R_t \quad \mathsf{E}U_1\right)+\Delta_t,\\
&\mathsf{P}\left(\boldsymbol{\lambda}^{\mathsf{I}}(\widetilde{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau) \quad x\right) \quad \mathsf{P}\ \left(\frac{1}{p\overline{m}}\boldsymbol{\lambda}^{\mathsf{I}}(\widetilde{\boldsymbol{\theta}}^\flat_\tau \quad \widetilde{\boldsymbol{\theta}}_\tau) \quad x \quad R_t \quad \mathsf{E}U_1\right) \quad \Delta_t.
\end{aligned}$$

From now on, we focus on the inference for the coefficient β_j (for $j = 1, \ldots, d$) and let $\boldsymbol{\lambda}^{\intercal} = (0, \ldots, 0, 1, 0, \ldots, 0)$(the $(j+1)$-th component is 1). We claim that on the event $\mathcal{E}_t$, for any $q \in (0, 1)$, we always have

$$\widetilde{c}_j(q \quad \Delta_t) \quad \widetilde{c}_j^{\flat}(q)/\sqrt{m} + R_t + \mathsf{E}U_1. \tag{B.6.2}$$

In fact, for any $\delta > 0$, on this event $\mathcal{E}_t$,

$$\mathsf{P}\left(\sqrt{\frac{1}{m}}(\widetilde{\beta}_j^{\flat} \quad \widetilde{\beta}_j) \quad \widetilde{c}_j(q \quad \Delta_t) \quad R_t \quad \mathsf{E}U_1 \quad \delta\right)$$
$$\mathsf{P}\left(\widetilde{\beta}_j \quad \beta_j \quad \widetilde{c}_j(q \quad \Delta_t) \quad \delta\right) + \Delta_t < q \quad \Delta_t + \Delta_t = q,$$

which implies that $\widetilde{c}_j(q \quad \Delta_t) \quad R_t \quad \mathsf{E}U_1 \quad \delta < \widetilde{c}_j^{\flat}(q)/\sqrt{m}$, for any $\delta > 0$, and thus $\widetilde{c}_j(q \quad \Delta_t)$ $\widetilde{c}_j^{\flat}(q)/\sqrt{m} + R_t + \mathsf{E}U_1$.

Step 2: In this step, we show that for any $C_D > 1$, any $q \in (1/2, 1)$,

$$R_t + |\mathsf{E}U_1| \quad (C_D \quad 1)\widetilde{c}_j^{\flat}(q)/\sqrt{m} + C_D\widetilde{D}_j \tag{B.6.3}$$

holds with high probability. In fact, by (A.2.29) and Lemma B.6.1, we have

$$|\mathsf{E}U_1| \quad \epsilon c_1\tau \left\| \frac{1}{c_F}\boldsymbol{\lambda}^{\intercal}\mathbf{S}^{1/2} \right\|_2 \quad \widetilde{D}_j + C\epsilon c_1\tau\left(\sqrt{\frac{d+t}{n}} + \epsilon\right) \quad \frac{C_D+1}{2}\widetilde{D}_j,$$

with probability at least $1 \quad (3T+8)e^{\quad t}$, as long as $n \quad C^{\prime}(d+t)$ and $\epsilon \quad c$ for some large enough $C^{\prime}$ and small enough c. The last inequality above follows the fact that by Lemma B.6.1, $\widetilde{D}_j \quad \epsilon c_1\tau(\|\widetilde{\boldsymbol{\lambda}}\|_2 \quad \delta_t)$ with $\delta_t = C(\sqrt{(d+t)/n} + \epsilon)$, and $\|\widetilde{\boldsymbol{\lambda}}\|_2 = \|c_F^{\quad 1}\boldsymbol{\lambda}^{\intercal}\mathbf{S}^{\quad 1/2}\|_2$ $c_4^{\quad 1}\sqrt{C_{\mathbf{S}}} \gtrsim \delta_t$, and thus $\widetilde{D}_j \gtrsim \epsilon\|\widetilde{\boldsymbol{\lambda}}\|_2 \gtrsim \epsilon\delta_t$.

We then bound R_t. Recall the definition of r_N and r_n. We know that

$$R_t \quad (\widetilde{C} + C^{\flat})\|\widetilde{\boldsymbol{\lambda}}\|_2\left\{\sqrt{\frac{d+t}{Nn}} + \epsilon\left(\sqrt{\frac{d+t}{n}} + \epsilon\right)\right\} := R_{t,1} + R_{t,2}$$

Clearly, we have $R_{t,2} \quad \frac{C_D \quad 1}{2}\widetilde{D}_j$ for large enough n and small enough ϵ. Then it remains to show that $R_{t,1} \quad (C_D \quad 1)\widetilde{c}_j^{\flat}(q)/\sqrt{m}$ with high probability. This is true because $R_{t,1} \quad 1/\sqrt{N} \quad \widetilde{c}_j^{\flat}(q)/\sqrt{m}$ when $n \gtrsim (d+t)^2$. Formally, by (B.5.2) and (B.5.4), we have

$$\mathsf{P}\left(\boldsymbol{\lambda}^{\intercal}(\widetilde{\boldsymbol{\theta}}_{\tau}^{\flat} \quad \widetilde{\boldsymbol{\theta}}_{\tau}) \quad x\right) \quad \Phi\left(\frac{\sqrt{n}}{\widehat{\sigma}}\left(x + \|\widetilde{\boldsymbol{\lambda}}\|_2 r_n\right)\right) + 5e^{\quad t} + \frac{C_5}{\sqrt{n}},$$

when conditioning on the event E_t. This indicates that

$$\widetilde{c}_j^{\flat}(q) \quad \left(\Phi^{-1}(q \quad \Delta_t^{\natural}) \quad k\widetilde{\lambda}k_2 r_n\right) \quad \frac{\widehat{\sigma}}{\sqrt{n}},$$

with $\Delta_t^{\natural} := 5e^{-t} + C_5/\sqrt{n}$. Picking $t = C^{\natural\natural}\log n$ for some large enough $C^{\natural\natural}$, then for any fixed $q \in (1/2, 1)$, we can have

$$\frac{1}{\sqrt{m}}\widetilde{c}_j^{\flat}(q) \quad \left[\Phi^{-1}\left(\frac{q}{2}+\frac{1}{4}\right) \quad k\widetilde{\lambda}k_2 r_n\right] \frac{\widehat{\sigma}}{\sqrt{N}} \quad \Phi^{-1}\left(\frac{q}{2}+\frac{1}{4}\right)\frac{\widehat{\sigma}}{2\sqrt{N}} \quad \frac{1}{C_D \quad 1}R_{t,1},$$

as long as $n \quad C^{\natural}(d+t)^2$ and $\epsilon \quad c$ for some large enough $C^{\natural}$ and small enough c. Combing the bounds for $R_{t,1}$, $R_{t,2}$ and $\|\mathbf{E}U_1\|$, we have shown that (B.6.3) holds when conditioning on the event E_t with $t = C^{\natural\natural}\log n$.

Step 3: Using the results (B.6.2) and (B.6.3) from the last two steps, we can have

$$\begin{aligned}\mathsf{P}\left(\widetilde{\beta}_j \quad \beta_j \quad C_D\widetilde{c}_j^{\flat}(q)/\sqrt{m} + C_D\widetilde{D}_j\right) \quad q \quad \mathsf{P}\left(\widetilde{\beta}_j \quad \beta_j \quad \widetilde{c}_j(q \quad \Delta_t)\right) \quad q \quad \mathsf{P}(E_t^c) \\ (q \quad \Delta_t) \quad q \quad \mathsf{P}(E_t^c) = \quad \Delta_t \quad \mathsf{P}(E_t^c).\end{aligned} \tag{B.6.4}$$

Regarding the other direction, one can similarly show that

$$\begin{aligned}\widetilde{c}_j^{\flat}(q)/\sqrt{m} \quad \widetilde{c}_j(q+\Delta_t) + R_t \quad \mathsf{E}U_1, \quad \text{(Step 1)} \\ R_t + \|\mathbf{E}U_1\| \qquad (C_D \quad 1)\widetilde{c}_j^{\flat}(q)/\sqrt{m} + C_D\widetilde{D}_j, \quad \text{(Step 2)}\end{aligned}$$

for any $q \in (0, 1/2)$. Therefore, $C_D\widetilde{c}_j^{\flat}(q)/\sqrt{m} \quad \widetilde{c}_j(q+\Delta_t) + C_D\widetilde{D}_j$, which implies

$$\begin{aligned}q \quad \mathsf{P}\left(\widetilde{\beta}_j \quad \beta_j \quad C_D\widetilde{c}_j^{\flat}(q)/\sqrt{m} \quad C_D\widetilde{D}_j\right) \quad q \quad \mathsf{P}(\widetilde{\beta}_j \quad \beta_j \quad \widetilde{c}_j(q+\Delta_t)) \quad \mathsf{P}(E_t^c) \\ q \quad \mathsf{P}(\widehat{\beta}_j \quad \beta_j \quad \widetilde{c}_j(q+\Delta_t) \quad \delta) \quad L(\delta) \quad \mathsf{P}(E_t^c). \\ \Delta_t \quad L(\delta) \quad \mathsf{P}(E_t^c),\end{aligned}$$

where $L(\delta) := \sup_{x \in \mathbb{R}} \mathsf{P}\left(\left|\widetilde{\beta}_j \quad \beta_j \quad x\right| \quad \delta\right)$ is the Lévy concentration function of the random variable $\widetilde{\beta}_j \quad \beta_j$. In order to avoid the usage of Lévy concentration function (which relies on the Bahadur representation and thus requires a more stringent condition on ϵ), we can actually show

$$R_t + \|\mathbf{E}U_1\| \qquad (C_D \quad 1)\widetilde{c}_j^{\flat}(q) + \widetilde{C}_D\widetilde{D}_j$$

for some $1 < \widetilde{C}_D < C_D$ in step 2. Therefore, we can instead have

$$\begin{aligned} & q \quad \mathsf{P}\left(\widehat{\beta}_j \quad \beta_j \quad C_D\widetilde{c}_j^{\flat}(q)/\sqrt{p}\overline{m} \quad C_D\widetilde{D}_j\right) \\ & q \quad \mathsf{P}(\widehat{\beta}_j \quad \beta_j \quad \widetilde{c}_j(q+\Delta_t) \quad (C_D \quad \widetilde{C}_D)\widetilde{D}_j) \quad \mathsf{P}(E_t^c) \\ & q \quad (q+\Delta_t) \quad \mathsf{P}(E_t^c) = \quad \Delta_t \quad \mathsf{P}(E_t^c). \end{aligned} \tag{B.6.5}$$

Let $q = 1 \quad \alpha/2$ in (B.6.4) and $q = \alpha/2$ in (B.6.5), respectively, and choose $t = C'' \log n$. This completes the proof.

□

Proof of Lemma B.6.1 We bound $|\widehat{c}_F \quad c_F|$ and $\|\widehat{\mathbf{S}}^{-1} \quad \mathbf{S}^{-1}\|_2$ subsequently, where $\| \quad \|_2$ is the spectral norm.

Step 1: Bound $|c_F \quad \widehat{c}_F|$. Let $R(\boldsymbol{\theta}) := n^{-1}\sum_{i\in I_1} \ell_\tau'(y_i \quad \bar{\boldsymbol{x}}_i^\top\boldsymbol{\theta})$. Then $\widehat{c}_F = R(\widetilde{\boldsymbol{\theta}}_\tau)$ and $c_F = \mathsf{E}_{\varepsilon \sim F}R(\boldsymbol{\theta}_\tau)$. Depending on whether the loss function is Huber's loss, we consider two cases.

Case 1: The loss function is Huber's loss and Condition 2.3.1 (ii) holds.

Under this case, by triangular inequality, we have

$$\begin{aligned} |\widehat{c}_F \quad c_F| & \quad \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} |R(\boldsymbol{\theta}) \quad \mathsf{E}R(\boldsymbol{\theta})| + |\mathsf{E}R(\boldsymbol{\theta}) \quad \mathsf{E}R(\boldsymbol{\theta}_\tau)| + |\mathsf{E}R(\boldsymbol{\theta}_\tau) \quad \mathsf{E}_{\varepsilon \sim F}R(\boldsymbol{\theta}_\tau)| \\ & := \Gamma_1 + \Gamma_2 + \Gamma_3, \end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}_\tau$ is assumed to be in $\boldsymbol{\Theta}_r := \{\boldsymbol{\theta} : \|\boldsymbol{\theta} \quad \boldsymbol{\theta}_\tau\|_{\mathbf{S}} \quad r\}$ for some $r > 0$. The last term Γ_3 is easy to bound as we notice that $\Gamma_3 = |\epsilon\mathsf{E}_{\boldsymbol{x}}\mathsf{E}_{\varepsilon|\boldsymbol{x}} \quad _{G_{\boldsymbol{x}}}\ell_\tau'(\varepsilon_i \quad \alpha_\tau)| \quad \epsilon$. To bound Γ_2, note that ℓ is Huber loss and $\ell_\tau'(u) = I(|u| \quad \tau)$. Therefore,

$$\begin{aligned} \Gamma_2 & \quad \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} |\mathsf{E}I(|\varepsilon \quad \alpha_\tau + \bar{\boldsymbol{x}}^\top(\boldsymbol{\theta}_\tau \quad \boldsymbol{\theta})| \quad \tau) \quad \mathsf{E}I(|\varepsilon \quad \alpha_\tau| \quad \tau)| \\ & \quad \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \mathsf{E}I\big(\big||\varepsilon \quad \alpha_\tau| \quad \tau\big| \quad |\bar{\boldsymbol{x}}^\top(\boldsymbol{\theta}_\tau \quad \boldsymbol{\theta})|\big) \\ & \quad \epsilon \quad 1 + (1 \quad \epsilon) \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \mathsf{E}_{\varepsilon \sim F}I(|\varepsilon \quad \alpha_\tau| \quad \tau| \quad |\bar{\boldsymbol{x}}^\top(\boldsymbol{\theta}_\tau \quad \boldsymbol{\theta})|) \\ & \quad \epsilon + 4C_F \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \mathsf{E}|\bar{\boldsymbol{x}}^\top(\boldsymbol{\theta}_\tau \quad \boldsymbol{\theta})| \quad \epsilon + 4C_F \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \|\boldsymbol{\theta}_\tau \quad \boldsymbol{\theta}\|_{\mathbf{S}} \quad \epsilon + 4C_F r. \end{aligned}$$

Here we use the anti-concentration inequality for distribution F under Condition 2.3.1 (ii).

Regarding Γ_1, as $\ell_\tau^{\prime\prime}$ is bounded by 1, then by McDiarmid's Inequality, we know that

$$\mathsf{P}(\Gamma_1 \quad \mathsf{E}\Gamma_1 > t) \quad \exp\left\{ \quad 2nt^2\right\}.$$

For the bound of $\mathsf{E}\Gamma_1$, it is easy to check that the VC dimension of the function class $\mathcal{F}_\tau := \left\{ f_{\boldsymbol{\theta}}(\varepsilon_i, \bar{\boldsymbol{x}}_i) = I(|\varepsilon_i \quad \alpha_\tau + \bar{\boldsymbol{x}}_i^\intercal(\boldsymbol{\theta}_\tau \quad \boldsymbol{\theta})| \quad \tau)|\boldsymbol{\theta} \in \Theta_r \right\}$ is no larger than Cd for some absolute constant C. Then by a standard VC dimension argument (see e.g. Theorem 8.3.23 in Vershynin (2018)), we have $\mathsf{E}\Gamma_1 \quad C\sqrt{d/n}$ for some (new) absolute constant $C^\prime$. Thus, $\Gamma_1 \quad C\sqrt{(d+t)/n}$ with probability at least $1 \quad e^{\ t}$, for some (new) absolute constant C. Combing the above bounds for $\Gamma_1, \Gamma_2, \Gamma_3$ (noting that $r \ . \ \sqrt{(d+t)/N} + \epsilon$ with probability at least $1 \quad 3(T+1)e^{\ t}$ by Corollary 3.2.1), we can obtain that

$$|\widehat{c}_F \quad c_F| \quad C\left(\sqrt{\frac{d+t}{n}} + \epsilon\right),$$

with probability at least $1 \quad (3T+4)e^{\ t}$, for some constant C independent of $(d, t, N, n, \epsilon, \boldsymbol{\lambda})$.

Case 2: The loss function is not Huber's loss and Condition 2.3.1 (i) holds. Under this case, we utilize the Lipschitz property of $\ell^\prime(\ )$ and decompose $|\widehat{c}_F \quad c_F|$ as follows:

$$\begin{aligned} |\widehat{c}_F \quad c_F| \quad & \left|R(\widetilde{\boldsymbol{\theta}}_\tau) \quad R(\boldsymbol{\theta}_\tau)\right| + |R(\boldsymbol{\theta}_\tau) \quad \mathsf{E}R(\boldsymbol{\theta}_\tau)| + |\mathsf{E}R(\boldsymbol{\theta}_\tau) \quad \mathsf{E}_{\varepsilon \ _F} R(\boldsymbol{\theta}_\tau)| \\ & := \Gamma_1 + \Gamma_2 + \Gamma_3, \end{aligned}$$

As $\ell_\tau^{\prime\prime}(\ )$ is L/τ-Lipschitz, we have

$$\Gamma_1 \quad \frac{L}{\tau n}\sum_{i \in I_1} |\bar{\boldsymbol{x}}_i^\intercal(\widetilde{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau)| \quad \frac{L}{\tau}\|\widetilde{\boldsymbol{\theta}}_\tau \quad \boldsymbol{\theta}_\tau\|_{\mathsf{S}} M_1(\boldsymbol{z}), \quad \text{with} \quad M_1(\boldsymbol{z}) = \sup_{\boldsymbol{u} \in \mathsf{S}^d} \frac{1}{n}\sum_{i \in I_1} |\boldsymbol{u}^\intercal \boldsymbol{z}_i|.$$

Since $\{\boldsymbol{z}_i\}_{i \in I_1}$ are *i.i.d.* sub-Gaussian vectors, we know that $M_1(\boldsymbol{z}) \quad C\sup_{\boldsymbol{u} \in \mathsf{S}^d} \mathsf{E}|\boldsymbol{u}^\intercal \boldsymbol{z}| \quad C$ with probability at least $1 \quad e^{\ t}$ for some absolute constant C when $n \ \& \ (d+t)$. To bound Γ_2, just notice that $R(\boldsymbol{\theta}_\tau) = n^{\ 1}\sum_{i \in I_1} \ell_\tau^\prime(\varepsilon_i \quad \alpha_\tau)$ and $\ell_\tau^\prime(\ )$ is bounded by c_4 under Condition 2.3.1 (i), and thus by Bernstein's inequality, we have

$$\Gamma_2 \quad Cc_4\left(\sqrt{\frac{t}{n}} + \frac{t}{n}\right)$$

with probability at least 1 $2e^{-t}$ for some absolute constant C. Similar to Case (i), Γ_3 can be bounded by ϵc_4 as $\Gamma_3 = |\epsilon \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\varepsilon|\boldsymbol{x}} G_{\boldsymbol{x}} \ell_\tau''(\varepsilon_i - \alpha_\tau)| \leq \epsilon c_4$. Combing the above bounds with the bound for $\|\widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau\|_{\mathbf{S}}$ in Corollary 3.2.1), we have

$$|\widehat{c}_F - c_F| \leq C\left(\sqrt{\frac{d}{N}} + \sqrt{\frac{t}{n}} + \epsilon\right) \leq C\left(\sqrt{\frac{d+t}{n}} + \epsilon\right)$$

with probability at least $1 - (3T+6)e^{-t}$, for some constant C independent of $(d, t, N, n, \epsilon, \boldsymbol{\lambda})$.

Step 2: Bound $\|\widehat{\mathbf{S}}^{-1} - \mathbf{S}^{-1}\|_2$. As $\{\boldsymbol{z}_i g_i^n\}_{i=1}$ are i.i.d sub-Gaussian random vectors with parameter $\nu_{\boldsymbol{x}}$, then by a standard ϵ-net argument and Bernstein's inequality, we can get

$$\|\widehat{\mathbf{S}} - \mathbf{S}\|_2 \leq C_0 \nu_{\boldsymbol{x}}^2 \|\mathbf{S}\|_2 \sqrt{\frac{d+t}{n}},$$

with probability at least $1 - 2e^{-t}$, for some absolute constant C_0 (assuming that $n \geq d+t$). By Weyl's inequality, we know that

$$\lambda_{\min}(\widehat{\mathbf{S}}) \geq \lambda_{\min}(\mathbf{S}) - C_0 \nu_{\boldsymbol{x}}^2 \lambda_{\max}(\mathbf{S}) \sqrt{\frac{d+t}{n}} \geq \frac{1}{2}\lambda_{\min}(\mathbf{S}) > 0,$$

with probability at least $1 - 2e^{-t}$, as long as $n \geq 4C_0^2 C_{\mathbf{S}}^2 \nu_{\boldsymbol{x}}^4 (d+t)/c_{\mathbf{S}}^2$. Therefore, $\widehat{\mathbf{S}}$ is invertible with probability at least $1 - 2e^{-t}$. Noticing that $\widehat{\mathbf{S}}^{-1} - \mathbf{S}^{-1} = \widehat{\mathbf{S}}^{-1}(\mathbf{S} - \widehat{\mathbf{S}})\mathbf{S}^{-1}$, we can get

$$\|\widehat{\mathbf{S}}^{-1} - \mathbf{S}^{-1}\|_2 \leq \|\widehat{\mathbf{S}}^{-1}\|_2 \|\widehat{\mathbf{S}} - \mathbf{S}\|_2 \|\mathbf{S}^{-1}\|_2 \leq 2C_0 \nu_{\boldsymbol{x}}^2 \frac{\lambda_{\max}(\mathbf{S})}{\lambda_{\min}(\mathbf{S})^2} \sqrt{\frac{d+t}{n}}$$

with probability at least $1 - 2e^{-t}$.

Step 3: Bound $|\|\widetilde{\boldsymbol{\lambda}}\|_2 - \|\widehat{\widetilde{\boldsymbol{\lambda}}}\|_2|$, where $\widetilde{\boldsymbol{\lambda}} := (c_F^{-1}\boldsymbol{\lambda}^\intercal \mathbf{S}^{-1/2})^\intercal$, $\widehat{\widetilde{\boldsymbol{\lambda}}} := (\widehat{c}_F^{-1}\boldsymbol{\lambda}^\intercal \widehat{\mathbf{S}}^{-1/2})^\intercal$.

$$\begin{aligned}
\left|\|\widetilde{\boldsymbol{\lambda}}\|_2^2 - \|\widehat{\widetilde{\boldsymbol{\lambda}}}\|_2^2\right| &= \left|\boldsymbol{\lambda}^\intercal \left(c_F^{-2}\mathbf{S}^{-1} - \widehat{c}_F^{-2}\widehat{\mathbf{S}}^{-1}\right)\boldsymbol{\lambda}\right| \\
&\leq \|\boldsymbol{\lambda}\|_2^2 \left[\left|c_F^{-2} - \widehat{c}_F^{-2}\right| \|\mathbf{S}^{-1}\|_2 + \widehat{c}_F^{-2}\|\mathbf{S}^{-1} - \widehat{\mathbf{S}}^{-1}\|_2\right] \\
&\leq C\|\boldsymbol{\lambda}\|_2^2 \frac{(1+\kappa(\mathbf{S}))}{\lambda_{\min}(\mathbf{S})} \left(\sqrt{\frac{d+t}{n}} + \epsilon\right),
\end{aligned}$$

with probability at least $1 - (3T+8)e^{-t}$, where $\kappa(\mathbf{S}) = \lambda_{\max}(\mathbf{S})/\lambda_{\min}(\mathbf{S})$ is the condition number of $\mathbf{S}$, and C is an constant independent of $(d, t, N, n, \epsilon, \boldsymbol{\lambda})$. Here, we use the fact that

$c_4 \quad c_F \quad c_2\mathsf{P}_{\varepsilon\ F}(j\varepsilon \quad \alpha_\tau j \quad c_3\tau) > c_2\kappa_\tau > 0$. Noticing that $k\widetilde{\lambda}k_2 \quad c_F{}^1 k\lambda k_2/\sqrt{\lambda_{\mathsf{max}}(\mathbf{S})}$, we can obtain that

$$\left|k\widetilde{\lambda}k_2 \quad k\widehat{\widetilde{\lambda}}k_2\right| \quad \frac{1}{k\widetilde{\lambda}k_2}\left|k\widetilde{\lambda}k_2^2 \quad k\widehat{\widetilde{\lambda}}k_2^2\right| \quad Cc_4 k\lambda k_2 \frac{\sqrt{\kappa(\mathbf{S})}(1+\kappa(\mathbf{S}))}{\sqrt{\lambda_{\mathsf{min}}(\mathbf{S})}}\left(\sqrt{\frac{d+t}{n}}+\epsilon\right),$$

with probability at least $1 \quad (3T+8)e^{\ t}$. Notice that $\kappa(\mathbf{S}) \quad C_{\mathbf{S}}/c_{\mathbf{S}}$ and $\lambda_{\mathsf{min}}(\mathbf{S}) \quad c_{\mathbf{S}}$ and we are done.

□

Appendix C Supplement to Chapter 4

This section provides all the proofs for Chapter 4. We first prove two lemmas in Section C.1 on the general upper bounds of the bias and variance terms of kernel estimators, which will be used to prove the minimax rates. In Section C.2, we provide proofs for Theorems 4.2.1, 4.4.1, 4.4.2. Next, we move to prove theorems on the results of our adaptive procedures. Section C.3 provides a proof of Theorem 4.5.1 using a key lemma, Lemma C.3.1, in the analysis of our adaptation theory. Then in Section C.4, we prove Proposition 4.3.1, and Theorems 4.3.2, 4.3.3, two main results of our adaptive procedures with arbitrary contamination. In Section C.5, we further prove Theorem 4.3.4 with a novel iterative trick to illustrate how Lepski's method on R can be modified to give a selection procedure adaptive to the contamination fraction ϵ. In the end, in Section C.6, we provide a proof of Theorem 4.3.5 to show that adaptation to both β_0 and ϵ is impossible.

C.1 Two Lemmas about bias and variance of kernel estimators

Before we get into the proofs of the theorems for minimax rates, we summarize the general upper bounds for bias and variance terms of the kernel estimator in the following lemmas. Though the results of these two lemmas are well-known in the non-parametric density estimation community, we still include the proofs here for completeness. Recall that the notations for bias and variance terms of the kernel estimator are given in (4.2.4)-(4.2.5).

Lemma C.1.1 (Bias). *If $p_X \in \mathsf{N}_{p,d}(\beta, L)$, K is a kernel in $\mathsf{K}_\beta(L_K)$, then for any $p \in [1, \infty]$,*

$$\|B_h(p_X, t)\|_p \leq L L_K^d \sum_{j=1}^{d} h_j^{\beta_j}. \tag{C.1.1}$$

Lemma C.1.2 (Variance). *(i) For any number $p \in [1, 2]$ and any density p_X on R^d, we have*

$$E_{p_X} \|\xi_{h,n}(p_X, t)\|_p \leq \frac{C_p}{(nV_h)^{\frac{1}{q}}},$$

where $V_h = \prod_{j=1}^d h_j$, $C_p = 2^{1+\frac{1}{p}} \|K\|_p$, *and* $1/q = 1 - 1/p$.

(ii) *For any number* $p \in (2, \infty)$ *and any density on* $\mathbb{R}^d$ $p_X \in L_{p,d}(M)$, *we have*

$$E_{p_X} \|\xi_{h,n}(p_X, t)\|_p \leq \frac{C_p}{(nV_h)^{\frac{1}{2}}},$$

where $C_p = 8D_p(M - 1)^{\frac{p-2}{2(p-1)}} (\|K\|_p - \|K\|_2)$, *and* $D_p = \frac{15p}{\log p}$.

Proof of Lemma C.1.1.

$$\|B_h(p_X, t)\|_p = \left\| \int_{\mathbb{R}^d} K_h(t - x) p_X(x) dx - p_X(t) \right\|_p = \left\| \int_{\mathbb{R}^d} K(u) \Big(p_X(t - uh) - p_X(t) \Big) du \right\|_p.$$

Here, $p_X(t - uh) = p_X\big(t_1 - u_1 h_1, \ldots, t_d - u_d h_d\big)$ and the integral $\int_{\mathbb{R}^d}(\,) du = \int_{\mathbb{R}^d}(\,) du_1 \ldots du_d$. For $j = 1, \ldots, d$, let

$$\Delta_j(u, t) := p_X\Big((t_1 - u_1 h_1), \ldots, (t_{j-1} - u_{j-1} h_{j-1}), (t_j - u_j h_j), t_{j+1}, \ldots, t_d\Big) - p_X\Big((t_1 - u_1 h_1), \ldots, (t_{j-1} - u_{j-1} h_{j-1}), t_j, t_{j+1}, \ldots, t_d\Big).$$

Then $p_X(t - uh) - p_X(t) = \sum_{j=1}^d \Delta_j$. If $p_X \in \mathcal{N}_{p,d}(\beta, L)$, then by Taylor's expansion, we have

$$\begin{aligned}
\Delta_j(u, t) = &\sum_{k=1}^{l_j - 1} \frac{(-u_j h_j)^k}{k!} D_j^{(k)} p_X\Big((t_1 - u_1 h_1), \ldots, (t_{j-1} - u_{j-1} h_{j-1}), t_j, t_{j+1}, \ldots, t_d\Big) \\
&+ \frac{(-u_j h_j)^{l_j}}{(l_j - 1)!} \int_0^1 (1 - \tau)^{l_j - 1} D_j^{(l_j)} p_X\Big((t_1 - u_1 h_1), \ldots, (t_{j-1} - u_{j-1} h_{j-1}), \\
&(t_j - \tau u_j h_j), t_{j+1}, \ldots, t_d\Big) d\tau,
\end{aligned}$$

where $l_j = \lfloor \beta_j \rfloor$. For a kernel $K \in \mathcal{K}_\beta(L_K)$, we know

$$K(u) = \prod_{j=1}^d K_j(u_j), \quad \int_{\mathbb{R}} u_j^k K_j(u_j) du_j = 0, \quad \forall k = 1, \ldots, l_j.$$

Therefore, with the notation below,

$$\widetilde{\Delta}_j(u, t) := \sum_{k=1}^{l_j} \frac{(-u_j h_j)^k}{k!} D_j^{(k)} p_X\Big((t_1 - u_1 h_1), \ldots, (t_{j-1} - u_{j-1} h_{j-1}), t_j, t_{j+1}, \ldots, t_d\Big),$$

we must have

$$\int_{\mathsf{R}^d} K(u)\widetilde{\Delta}_j(u,t)du = 0. \tag{C.1.2}$$

Furthermore, denote

$$\begin{aligned}\Delta_{p_X,j}(\tau,u,t) :=& D_j^{(l_j)} p_X\Big((t_1 \quad u_1h_1),\ldots,(t_{j\ 1} \quad u_{j\ 1}h_{j\ 1}),(t_j \quad \tau u_jh_j),t_{j+1},\ldots,t_d\Big)\\ & D_j^{(l_j)} p_X\Big((t_1 \quad u_1h_1),\ldots,(t_{j\ 1} \quad u_{j\ 1}h_{j\ 1}),t_j,t_{j+1},\ldots,t_d\Big),\end{aligned}$$

then we have

$$\Delta_j(u,t) \quad \widetilde{\Delta}_j(u,t) = \frac{(\ u_jh_j)^{l_j}}{(l_j \quad 1)!}\int_0^1 (1 \quad \tau)^{l_j\ 1}\Delta_{p_X,j}(\tau,u,t)d\tau.$$

Applying twice Minkowski's integral inequality and using (C.1.2) and the fact that p_X 2 $\mathsf{N}_{p,d}(\beta,L)$, we have

$$\begin{aligned}&\left\|\int_{\mathsf{R}^d} K(u)\Delta_j(u,t)du\right\|_p = \left\|\int_{\mathsf{R}^d} K(u)(\Delta_j \quad \widetilde{\Delta}_j)(u,t)du\right\|_p\\
&= \left\{\int_{\mathsf{R}^d}\left|\int_{\mathsf{R}^d} K(u)(\Delta_j \quad \widetilde{\Delta}_j)(u,t)du\right|^p dt\right\}^{\frac{1}{p}}\\
&\quad \int_{\mathsf{R}^d}\left\{\int_{\mathsf{R}^d}\left|K(u)(\Delta_j \quad \widetilde{\Delta}_j)(u,t)\right|^p dt\right\}^{\frac{1}{p}}du\\
&\quad \int_{\mathsf{R}^d} \mathsf{j}K(u)\mathsf{j}\frac{\mathsf{j}u_jh_j\mathsf{j}^{l_j}}{(l_j \quad 1)!}\int_0^1 (1 \quad \tau)^{l_j\ 1}\left\{\int_{\mathsf{R}^d}\left|\Delta_{p_X,j}(\tau,u,t)\right|^p dt\right\}^{\frac{1}{p}}d\tau du\\
&\quad \int_{\mathsf{R}^d} \mathsf{j}K(u)\mathsf{j}\frac{\mathsf{j}u_jh_j\mathsf{j}^{l_j}}{l_j!}\Big(L\mathsf{j}u_jh_j\mathsf{j}^{\beta_j\ l_j}\Big)du\\
&\quad h_j^{\beta_j}L\int_{\mathsf{R}^d}\mathsf{j}K(u)\mathsf{j}\mathsf{j}u_j\mathsf{j}^{\beta_j}du \quad LL_K^d h_j^{\beta_j}.\end{aligned}$$

Noting that $p_X(t \quad uh) \quad p_X(t) = \sum_{j=1}^d \Delta_j(u,t)$, we get (C.1.1) by summing the above bound over $j = 1,\ldots,d$. □

Proof of Lemma C.1.2. **(*i*)** For simplicity, we write $E[\ ]$ to represent $E_{p_X}[\ ]$ in this proof. Let $\xi_i(t) = K_h(t \quad X_i) \quad E[K_h(t \quad X_i)]$, then $\xi_{h,n}(p_X, t) = \frac{1}{n}\sum_{i=1}^n \xi_i(t)$. It is known for centered independent random variables $\xi_1, ..., \xi_n$, the Bahr-Esseen inequality (4.3.8) holds. Therefore, we have

$$
\begin{aligned}
E\|\xi_{h,n}(p_X, t)\|_p &= \frac{1}{n} E\Big(\int_{\mathbb{R}^d} \Big| \sum_{i=1}^n \xi_i(t) \Big|^p dt \Big)^{\frac{1}{p}} \ (\ ) \ \frac{1}{n}\Big(\int_{\mathbb{R}^d} E\Big| \sum_{i=1}^n \xi_i(t) \Big|^p dt \Big)^{\frac{1}{p}} \\
&\ (\ ) \ \frac{2^{1/p}}{n}\Big(\int_{\mathbb{R}^d} \sum_{i=1}^n E\Big|\xi_i(t)\Big|^p dt \Big)^{\frac{1}{p}} \\
&= \frac{2^{1/p}}{n}\Big(\int_{\mathbb{R}^d} nE\Big| K_h(t \quad X_1) \quad EK_h(t \quad X_1) \Big|^p dt \Big)^{\frac{1}{p}},
\end{aligned}
$$

where $(\ )$ is due to Jensen's inequality, and $(\quad)$ follows from (4.3.8). By Jensen's inequality $(a+b)^p \quad 2^{p\ 1}(a^p + b^p), \forall a, b \quad 0, p \quad 1$, we know

$$
\begin{aligned}
E\Big| K_h(t \quad X_1) \quad EK_h(t \quad X_1) \Big|^p \quad & 2^{p\ 1}\Big(E\Big| K_h(t \quad X_1) \Big|^p + \Big| EK_h(t \quad X_1) \Big|^p \Big) \\
& 2^p E\Big| K_h(t \quad X_1) \Big|^p.
\end{aligned}
$$

Combining the above two inequalities, we get

$$
\begin{aligned}
E\|\xi_{h,n}(p_X, t)\|_p \quad & \frac{2^{1+\frac{1}{p}} n^{\frac{1}{p}}}{n}\Big(\int_{\mathbb{R}^d} E\Big| K_h(t \quad X_1) \Big|^p dt \Big)^{\frac{1}{p}} = \frac{2^{1+\frac{1}{p}}}{n^{\frac{1}{q}}}\Big(E\int_{\mathbb{R}^d} \Big| K_h(t \quad X_1) \Big|^p dt \Big)^{\frac{1}{p}} \\
&= \frac{2^{1+\frac{1}{p}}}{n^{\frac{1}{q}}}\Big(E\int_{\mathbb{R}^d} \Big| K_h(t) \Big|^p dt \Big)^{\frac{1}{p}} = \frac{2^{1+\frac{1}{p}}\|K\|_p}{(nV_h)^{\frac{1}{q}}}.
\end{aligned}
$$

(*ii*) For $p \in (2, \infty)$, let $D_p = 15p/\log p$. By Rosenthal inequality (4.3.9), we have

$$
\begin{aligned}
E\|\xi_{h,n}(p_X, t)\|_p &= \frac{1}{n} E\Big(\int_{\mathbb{R}^d} \Big| \sum_{i=1}^n \xi_i(t) \Big|^p dt \Big)^{\frac{1}{p}} \quad \frac{1}{n}\Big(\int_{\mathbb{R}^d} E\Big| \sum_{i=1}^n \xi_i(t) \Big|^p dt \Big)^{\frac{1}{p}} \\
&\quad \frac{2D_p}{n}\Big(\int_{\mathbb{R}^d} \sum_{i=1}^n E\Big|\xi_i(t)\Big|^p dt \Big)^{\frac{1}{p}} + \frac{2D_p}{n}\Big(\int_{\mathbb{R}^d} \Big(\sum_{i=1}^n E\xi_i^2(t) \Big)^{p/2} dt \Big)^{\frac{1}{p}} \\
&:= \Gamma_1 + \Gamma_2.
\end{aligned}
$$

From the proof for (i), we know $\Gamma_1 \quad 4D_p\|K\|_p(nV_h)^{\ 1/q}$. For Γ_2, we have

$$\Gamma_2 \quad \frac{2D_p}{n^{1/2}}\left(\int_{\mathsf{R}^d}\left(EK_h^2(t \quad X_1)\right)^{p/2}dt\right)^{\frac{1}{p}} = \frac{2D_p}{n^{1/2}}\|K\|_h^2 \quad p_X\|_{p/2}^{1/2}$$
$$\frac{2D_p}{n^{1/2}}\|K\|_h^2\|_1^{1/2}\|p_X\|_{p/2}^{1/2} = \frac{2D_p}{(nV_h)^{1/2}}\|K\|\|_2\|p_X\|_{p/2}^{1/2}. \quad \text{(by Young's inequality)}$$

For a density function p_X, by Hölder's inequality, we know $\|p_X\|_{p/2} \quad \|p_X\|_p^{\frac{p}{p-1}}$. Therefore, for a density $p_X \in \mathcal{L}_{p,d}(M)$,

$$\Gamma_2 \quad \frac{2D_p}{(nV_h)^{1/2}}\|K\|\|_2 M^{\frac{p-2}{2(p-1)}},$$

$$E\|\xi_{h,n}(p_X,t)\|_p \quad \Gamma_1+\Gamma_2 \quad \frac{1}{2}C_p\left(\frac{1}{(nV_h)^{1/q}}+\frac{1}{(nV_h)^{1/2}}\right) \quad C_p\frac{1}{(nV_h)^{1/2}},$$

where $C_p = 8D_p(M-1)^{\frac{p-2}{2(p-1)}}(\|K\|_p-\|K\|_2)$. □

C.2 Proofs of Theorems 4.2.1, 4.4.1, 4.4.2

C.2.1 Proof of Theorem 4.2.1

Proof of Theorem 4.2.1. **(*i*) Upper bound (i.e. Proof of Theorem 4.2.2)** For simplicity, we write $P_\epsilon = P(\epsilon, f, G)$. In light of (4.2.6), we just need to appropriately bound the bias term, variance term and the contamination term respectively. Bias term: For any kernel $K \in \mathcal{K}_\beta(L_K)$ with $\beta \quad \beta_0$ (and thus $K \in \mathcal{K}_{\beta_0}(2L_K)$), by Lemma C.1.1, we have

$$\|B_h(f,t)\|_p \quad L_0(2L_K)^d\sum_{j=1}^d h_j^{\beta_{0,j}}.$$

Contamination term: For any kernel $K \in \mathcal{K}_\beta(L_K)$, a direct calculation shows that

$$E_{P_\epsilon}\frac{n \quad n_1}{n}\left(\|K_h\|_p+\|f\|_p\right) \quad \left(V_h^{-1/q}\|K\|_p+L_0\right)E_{P_\epsilon}\frac{n \quad n_1}{n} \quad \epsilon\left(V_h^{-1/q}L_K^d+L_0\right).$$

Variance term: By Lemma C.1.2, we know

$$E_{P_\epsilon}\left[\|\xi_{h,n_1}(f,t)\|_p|n_1\right] = E_f\left[\|\xi_{h,n_1}(f,t)\|_p|n_1\right] \quad C_p(n_1V_h)^{-1/(q-2)},$$

where $C_p = 2^{1+1/p} L_K^d$ for $p \in [1,2]$ and $C_p = 8D_p(L_0 - 1)^{\frac{p-2}{2(p-1)}} L_K^d$ with $D_p = \frac{15p}{\log p}$ for $p \in (2, \infty)$. Then it suffices to bound $E n_1^{1/(q-2)}$. Recall that $n_1 \sim Binomial(n, 1-\epsilon)$, by Bernstein inequality (e.g. Theorem 2.8.4 in Vershynin (2018)), we have

$$\frac{n_1}{n} \geq (1-\epsilon) - 2\sqrt{\frac{\log n}{n}}\sqrt{\epsilon(1-\epsilon)} - \frac{2}{3}\frac{\log n}{n} \geq \frac{1-\epsilon}{2}$$

with probability at least $1 - n^{-1}$, when $n/\log n \geq \max\{8/3(1-C_0), 64C_0/(1-C_0)\}$. Therefore,

$$E n_1^{1/(q-2)} \leq \left(n\frac{1-\epsilon}{2}\right)^{1/(q-2)} + n^{-1} \leq \left(\frac{2}{1-C_0}+1\right) n^{1/(q-2)},$$

for any $\epsilon \leq C_0 < 1$. Combining the above bounds, with $C = (L_0-1)(L_K^d-1)(2^d + 8D_p\frac{3-C_0}{2-C_0} + 2)$, we obtain that

$$\sup_{\substack{f\in\mathcal{P}_{p,d}(\beta_0,L_0)\\ G}} E_{P(\epsilon,f,G)}\|\widehat{f}_h - f\|_p \leq C\left\{\sum_{j=1}^d h_j^{\beta_{0,j}} + (nV_h)^{-1/(q-2)} + \epsilon V_h^{-1/q}\right\},$$

for any $\epsilon \leq C_0 < 1$, and any n such that $n/\log n \geq \max\{8/3(1-C_0), 64C_0/(1-C_0)\}$.

Finally, by choosing $h_j \asymp n^{-\frac{\beta_0}{\beta_{0,j}(2\beta_0+1)}} \wedge \epsilon^{\frac{q\beta_0}{\beta_{0,j}(q\beta_0+1)}}$, we achieve the upper bound $n^{-\frac{\beta_0}{2\beta_0+1}} \wedge \epsilon^{\frac{q\beta_0}{q\beta_0+1}}$.

***(ii)* Lower bound (i.e. Proof of Theorem 4.2.4)** We only show how to obtain the second term of the lower bound $\epsilon^{\frac{q\beta_0}{q\beta_0+1}}$. Choose a function $f_0 \in \mathcal{P}_{p,d}(\beta_0, L_0/2)$ bounded away from 0 (i.e. $\exists \gamma_0 > 0$ s.t. $f_0(x) \geq \gamma_0 > 0, \forall x \in \mathbb{R}^d$), and a function $\phi_0 : \mathbb{R} \to \mathbb{R}$, which is infinitely differentiable, has a compact support and satisfies $\int \phi_0 = 0$. (For example, $\phi_0(u) = ue^{\frac{1}{1-u^2}}\mathbf{1}\{|u| \leq 1\}$.) Let

$$f_1(x) = f_0(x) + \gamma V_h^{\beta_0+\frac{1}{p}} \prod_{j=1}^d \phi_0\Big(\frac{x_j}{h_j}\Big),$$

where $h_j = \epsilon^{\frac{q\beta_0}{\beta_{0,j}(q\beta_0+1)}}$, $V_h = \prod_{j=1}^d h_j = \epsilon^{\frac{q}{q\beta_0+1}}$. If $\bar{\beta}_0 \geq \frac{1}{p}$, we can choose γ sufficiently small (when $\gamma \leq \gamma_0/\|\phi_0\|_1^d$) such that f_1 is also a density function.

Denote $\phi(x) = \gamma V_h^{\beta_0+\frac{1}{p}} \prod_{j=1}^d \phi_0\Big(\frac{x_j}{h_j}\Big)$. If γ is sufficiently small, we have the following three facts about ϕ:

(a) For $k = 0, \ldots, \lfloor \beta_{0,j} \rfloor$, $j = 1, \ldots, d$,

$$\| D_j^{(k)} \phi \|_p = \gamma V_h^{\beta_0} h_j^{\ k} \| \phi_0 \|_p^d \cdot \| D_j^{(k)} \phi_0 \|_p \quad \gamma \| \phi_0 \|_p^d \cdot \| D_j^{(k)} \phi_0 \|_p \quad \frac{L_0}{2}.$$

Here, we have used the inequality $V_h^{\beta_0} h_j^{\ k} \quad V_h^{\beta_0} h_j^{\ \beta_{0,j}} = 1$.

(b) For $j = 1, \ldots, d$, denote $l_j = \lfloor \beta_{0,j} \rfloor$. Then we have

$$\begin{aligned}
&\left\{ \int_{\mathbb{R}^d} \left| D_j^{(l_j)} \phi(t_1, \ldots, t_j + z, \ldots, t_d) \quad D_j^{(l_j)} \phi(t_1, \ldots, t_j, \ldots, t_d) \right|^p \mathrm{d}t \right\}^{1/p} \\
&= \gamma V_h^{\beta_0 \quad \frac{1}{p}} \left(\prod_{k \in \mathcal{G}_j} h_k^{\frac{1}{p}} \| \phi_0 \|_p \right) \left\{ \int_{\mathbb{R}} \left| h_j^{\ l_j} \left(\phi_0^{(l_j)} \left(\frac{t_j + z}{h_j} \right) \quad \phi_0^{(l_j)} \left(\frac{t_j}{h_j} \right) \right) \right|^p \mathrm{d}t_j \right\}^{1/p} \\
&= \gamma V_h^{\beta_0} h_j^{\ l_j} \| \phi_0 \|_p^{d \ 1} \left\{ \int_{\mathbb{R}} \left| \phi_0^{(l_j)} \left(u + \frac{z}{h_j} \right) \quad \phi_0^{(l_j)}(u) \right|^p \mathrm{d}u \right\}^{1/p} \\
&\quad \gamma \| \phi_0 \|_p^{d \ 1} V_h^{\beta_0} h_j^{\ l_j} C_{p,l_j} \left| \frac{z}{h_j} \right|^{\beta_{0,j} \quad l_j} \\
&= \gamma \| \phi_0 \|_p^{d \ 1} V_h^{\beta_0} h_j^{\ \beta_{0,j}} C_{p,l_j} |z|^{\beta_{0,j} \quad l_j} = \gamma \| \phi_0 \|_p^{d \ 1} C_{p,l_j} |z|^{\beta_{0,j} \quad l_j} \quad \frac{L_0}{2} |z|^{\beta_{0,j} \quad l_j}.
\end{aligned}$$

Here, $C_{p,l_j} = \| \phi_0^{(l_j+1)} \|_p \_ 2 \| \phi_0^{(l_j)} \|_p$, and we have used the fact that

$$\begin{aligned}
&\| \phi_0^{(l_j)}(\ + t) \quad \phi_0^{(l_j)}(\ ) \|_p \quad \| \phi_0^{(l_j+1)} \|_p |t| \wedge 2 \| \phi_0^{(l_j)} \|_p \\
&\qquad (\| \phi_0^{(l_j+1)} \|_p _ 2 \| \phi_0^{(l_j)} \|_p) |t|^{\beta_j \quad l_j}.
\end{aligned}$$

(c)

$$\int \left| \frac{1 \quad \epsilon}{\epsilon} \phi \right| = \frac{1 \quad \epsilon}{\epsilon} \gamma V_h^{\beta_0 + \frac{1}{q}} \| \phi_0 \|_1^d = (1 \quad \epsilon) \gamma \| \phi_0 \|_1^d \quad 2.$$

Note that (a) and (b) imply $f_1 \in \mathcal{P}_{p,d}(\beta_0, L_0)$. In addition, (c) implies that $\mathrm{TV}(P_{f_0}, P_{f_1}) = 1/2 \|\phi\|_1 \quad \frac{\epsilon}{1 \quad \epsilon}$. Therefore, by Theorem 4.2.3, we obtain

$$\inf_{\widehat{f}} \sup_{\substack{f \in \mathcal{P}_{p,d}(\beta_0, L_0) \\ G}} E_{P(\epsilon, f, G)} \| \widehat{f} \quad f \|_p \gtrsim \omega(\epsilon, \Theta) \quad \|f_1 \quad f_0\|_p = \gamma V_h^{\beta_0} \| \phi_0 \|_p^d \quad \epsilon^{\frac{q \beta_0}{q \beta_0 + 1}},$$

which completes our proof of lower bound. □

C.2.2 Proofs of Theorems 4.4.1-4.4.2

Proofs of Theorems 4.4.1-4.4.2. **(*i*) Upper bound** We only show the upper bound for rate (4.4.1) in Theorem 4.4.1, where $g \in \mathcal{L}_{p,d}(L_1)$. Then the upper bound for rate (4.4.2) in Theorem 4.4.2 naturally holds as $\mathcal{P}_{p,d}(\beta_1, L_1)$ is a subclass of $\mathcal{L}_{p,d}(L_1)$. For simplicity, we write $p_\epsilon = p(\epsilon, f, g)$ in what follows. Clearly, we have

$$\begin{aligned} E_{p_\epsilon} \| \widehat{f}_h - f \|_p &\le E_{p_\epsilon} \| \widehat{f}_h - \widehat{f}_h \|_p + \| \widehat{f}_h - f \|_p \\ &:= var(\widehat{f}_h) + bias(\widehat{f}_h), \end{aligned}$$

where

$$\begin{aligned} bias(\widehat{f}_h) = \big\| E_{p_\epsilon} K_h(t - X) \big((1-\epsilon) f(t) + \epsilon g(t)\big) - \epsilon\big(f(t) - g(t)\big) \big\|_p \\ \le \| B_h((1-\epsilon) f + \epsilon g, t) \|_p + \epsilon \| f - g \|_p \\ (1-\epsilon) \| B_h(f, t) \|_p + \epsilon \| B_h(g, t) \|_p + \epsilon \| f - g \|_p. \end{aligned}$$

By assumptions, $f \in \mathcal{N}_{p,d}(\beta_0, L_0)$, K is a kernel in $\mathcal{K}_{\beta_0}(L_K)$, then Lemma C.1.1 tells us that for any $p \in [1, \infty)$,

$$\| B_h(f, t) \|_p \le L_0 L_K^d \sum_{j=1}^{d} h_j^{\beta_{0,j}}. \tag{C.2.1}$$

For $g \in \mathcal{L}_{p,d}(L_1)$, by Young's inequality, we can get

$$\begin{aligned} \epsilon \| B_h(g, t) \|_p &\le \epsilon\big(\| K_h - g \|_p + \| g \|_p \big) \le \epsilon\big(\| K_h \|_1 \| g \|_p + \| g \|_p \big) \\ &= \epsilon\big(\| K \|_1 + 1 \big) \| g \|_p \le \epsilon(L_K^d + 1) L_1. \end{aligned} \tag{C.2.2}$$

(C.2.1), (C.2.2) and the fact $\| f - g \|_p \le \| f \|_p + \| g \|_p \le L_0 + L_1$ give us an upper bound of the bias term:

$$bias(\widehat{f}_h) \le C\bigg(\sum_{j=1}^{d} h_j^{\beta_{0,j}} + \epsilon \bigg), \tag{C.2.3}$$

where constant $C = (L_0 + L_1)(L_K^d + 2)$. For the variance term, Lemma C.1.2 implies an upper bound for $p_X = p_\epsilon = (1 \quad \epsilon)f + \epsilon g$:

$$var(\widehat{f_h}) = E_{p_\epsilon} k \xi_{h,n}(p_\epsilon, t) k_p \quad \begin{cases} C_p^{\emptyset}(nV_h)^{1/q}, & p \in [1,2], \\ C_p^{\emptyset}(nV_h)^{1/2}, & p \in (2, \infty), \end{cases} \tag{C.2.4}$$

where $C_p^{\emptyset} = 2^{1+1/p} L_K^d$ for $p \in [1,2]$ and $C_p^{\emptyset} = 8 D_p L_K^d (L_0 + L_1 + 1)^{\frac{p-2}{2(p-1)}}$ for $p \in (2, \infty)$, $D_p = \frac{15p}{\log p}$.

(C.2.3) and (C.2.4) give us the upper bound for kernel estimator $\widehat{f_h}$:

$$E_{p(\epsilon,f,g)} k \widehat{f_h} \quad f k_p \ . \ \begin{cases} \sum_{j=1}^d h_j^{\beta_{0,j}} + \epsilon + (nV_h)^{1/q}, & p \in [1,2] \\ \sum_{j=1}^d h_j^{\beta_{0,j}} + \epsilon + (nV_h)^{1/2}, & p \in (2, \infty). \end{cases}$$

Choosing $h_j \quad n^{\frac{\beta_0}{\beta_{0,j}(q\beta_0+1)}}$ (for $1 \quad p \quad 2$) or $n^{\frac{\beta_0}{\beta_{0,j}(2\beta_0+1)}}$ (for $p > 2$), we get the upper bound $n^{\frac{\beta_0}{q\beta_0+1}} \_ \epsilon$ (for $1 \quad p \quad 2$) or $n^{\frac{\beta_0}{2\beta_0+1}}$ (for $p > 2$).

***(ii)* Lower bound** We only show the lower bound for rate (4.4.2) in Theorem 4.4.2, where $g \in P_{p,d}(\beta_1, L_1)$. Then the lower bound for rate (4.4.1) in Theorem 4.4.1 can be established immediately since $P_{p,d}(\beta_1, L_1)$ is a subclass of $L_{p,d}(L_1)$. Again, we just show the second term ϵ in (4.4.2), as the first term is the classical minimax rate. We consider the following functions:

$$f_1(x) = f_0(x) + \epsilon\gamma\phi(x),$$
$$g_1(x) = g_0(x) \quad (1 \quad \epsilon)\gamma\phi(x).$$

Here, we choose some functions $f_0 \in P_p(\beta_0, L_0/2)$, $g_0 \in P_p(\beta_1, L_1/2)$, both bounded away from zero; and some function $\phi(x) := \prod_{j=1}^d \phi_0(x_j)$, where $\phi_0 : \mathbb{R} \ ! \ \mathbb{R}$ is infinitely differentiable, has a compact support and satisfies $\int \phi_0 = 0$. (For example, $\phi_0(u) = ue^{\frac{1}{1-u^2}} 1\{|u| < 1\}$.) For any $\beta \in \mathbb{R}^+$, denote $l = \lfloor \beta \rfloor$. Then we have

$$k \phi_0^{(l)}(\cdot + t) \quad \phi_0^{(l)}(\cdot) k_p \ k \phi_0^{(l+1)} k_p |t| \wedge 2 k \phi_0^{(l)} k_p \quad (k \phi_0^{(l+1)} k_p \_ 2 k \phi_0^{(l)} k_p)|t|^\beta \ ^l. \tag{C.2.5}$$

Consequently, it is easy to verify $f_1 \in P_p(\beta_0, L_0)$, $g_1 \in P_p(\beta_1, L_1)$ with a sufficiently small γ.

Given the constructed f_1 and g_1, we have

$$\inf_{\widehat{f}} \sup_{\substack{f \in P_p(\beta_0, L_0) \\ g \in P_p(\beta_1, L_1)}} E_{p(\epsilon,f,g)} \| \widehat{f} - f \|_p \geq \inf_{\widehat{f}} \frac{1}{2}(E_{p(\epsilon,f_0,g_0)} \| \widehat{f} - f_0 \|_p + E_{p(\epsilon,f_1,g_1)} \| \widehat{f} - f_1 \|_p)$$

$$\geq \frac{1}{2} \| f_0 - f_1 \|_p \geq \epsilon.$$

The second inequality holds because under our construction, $p(\epsilon, f_0, g_0) = p(\epsilon, f_1, g_1)$. □

C.3 Proof of Theorem 4.5.1

This proof is mainly based on the proofs of Theorems 1 and 2 in (Goldenshluger and Lepski, 2011a). We start with a key lemma from Lemmas 1 and 2 of (Goldenshluger and Lepski, 2011a) below. Later, we also need this lemma in the proofs of Theorems 4.3.2 and 4.3.3.

Lemma C.3.1 ((Goldenshluger and Lepski, 2011a)). *Assume $X_1, ..., X_n \overset{i.i.d.}{\sim} p_X$. Assume that the kernel K satisfies the conditions:*

(K1) K is Lipschitz: $|K(x) - K(y)| \leq L_K \|x - y\|, \forall x, y \in \mathbb{R}^d$.

(K2) K is compactly supported. Without loss of generality, we assume $supp(K) \subseteq [-1/2, 1/2]^d$.

(K3) $\exists k_1 < 1$ such that $\|K\|_\infty \leq k_1$.

Let $H = \bigotimes_{j=1}^d [h_j^{\min}, h_j^{\max}]$, $V_{\min} = \prod_{i=1}^d h_i^{\min}, V_{\max} = \prod_{i=1}^d h_i^{\max}$, $A_H = \prod_{j=1}^d [1 \vee \log(h_j^{\max}/h_j^{\min})]$, and $B_H = [1 \vee \log_2(V_{\max}/V_{\min})]$. Under the assumptions (K1) – (K3), the following two inequalities hold with $\delta_{n,s}$ and $\widetilde{\delta}_{n,s}$ specified later.

$$E_{p_X} \sup_{h \in H} [\|\xi_{h,n}(p_X, t)\|_s - d_s(K_h)]_+ \leq \delta_{n,s}, \tag{C.3.1}$$

$$E_{p_X} \sup_{(h,l) \in H \times H} [\|\xi_{h,l,n}(p_X, t)\|_s - d_s(K_h - K_l)]_+ \leq \widetilde{\delta}_{n,s}. \tag{C.3.2}$$

- *(i) For $s \in [1, 2)$, $n \geq 4^{2s/(2-s)}$, (C.3.1)-(C.3.2) hold with*

$$\delta_{n,s} := C_1 A_H^2 n^{1/s} \exp\left\{-\frac{2n^{2/s-1}}{37}\right\}, \quad \widetilde{\delta}_{n,s} := \widetilde{C_1} A_H^4 n^{1/s} \exp\left\{-\frac{2n^{2/s-1}}{37}\right\}.$$

- *(ii) For* $s = 2$*, let* $p_X \in \mathcal{D}_1 := \left\{ f : \mathbb{R} \to \mathbb{R} \,\middle|\, \|f\|_1 \leq L_1 < 1 \right\}$*. Assume* $8[L_1^2 V_{\max} + 4n^{-1/2}] \leq 1$*, then (C.3.1)-(C.3.2) hold with*

$$\delta_{n,2} := C_2 A_H^2 n^{1/2} \exp\left\{ -\frac{1}{16[L_1^2 V_{\max} + 4n^{-1/2}]} \right\},$$
$$\widetilde{\delta}_{n,2} := \widetilde{C_2} A_H^4 n^{1/2} \exp\left\{ -\frac{1}{16[L_1^2 V_{\max} + 4n^{-1/2}]} \right\}.$$

- *(iii) For* $s \in (2, 1)$*, assume that* $p_X \in \mathcal{D}_1$*,* $n \geq c_{0,1}$*,* $nV_{\min} > c_{0,2}$*,* $V_{\max} \geq 1/\sqrt{n}$*, then (C.3.1)-(C.3.2) hold with*

$$\delta_{n,s} := C_3 A_H^2 B_H n^{1/2} \exp\left\{ -\frac{C_4}{L_1 V_{\max}^{2/s}} \right\}, \quad \widetilde{\delta}_{n,s} := \widetilde{C_3} A_H^4 B_H n^{1/2} \exp\left\{ -\frac{\widetilde{C_4}}{L_1 V_{\max}^{2/s}} \right\}.$$

In addition, for any $H_1 \subset H$ *and* $H_2 \subset H$*,*

$$E_{p_X} \sup_{h \in H_1} \widehat{r}_s(K_h) \leq (1 + 8D_s) \sup_{h \in H_1} r_s(K_h, p_X) + \zeta_{n,s}, \tag{C.3.3}$$

$$E_{p_X} \sup_{(h,l) \in H_1 \times H_2} \widehat{r}_s(K_h - K_l) \leq (1 + 8D_s) \sup_{(h,l) \in H_1 \times H_2} r_s(K_h - K_l, p_X) + \widetilde{\zeta}_{n,s}, \tag{C.3.4}$$

where $\zeta_{n,s} := C_5 A_H^2 B_H n^{(s-2)/2s} \exp\left\{ -C_6 b_{n,s} \right\}$, $\widetilde{\zeta}_{n,s} := \widetilde{C_5} A_H^4 B_H n^{(s-2)/2s} \exp\left\{ -\widetilde{C_6} b_{n,s} \right\}$, $b_{n,s} := n^{4/s - 1}$ *if* $s \in (2,4)$ *and* $b_{n,s} := [L_1 V_{\max}^{4/s}]^{-1}$ *if* $s \in [4, 1)$*. The constants* $C_i, \widetilde{C}_i, i = 1, ..., 6$ *and* $c_{0,2}$ *depend on* L_K, k_1, d, s *only, while* $c_{0,1}$ *depends also on* L_1*.*

Proofs of Theorem 4.5.1. With the lemma above, we are ready to prove Theorem 4.5.1. Our strategy is to show

1

$$\left\| \widehat{f}_{\widehat{h}} - f \right\|_p \leq \inf_{h \in H} \left\{ 4\widehat{R}_{h,p} + \left\| \widehat{f}_h - f \right\|_p \right\},$$

where $\widehat{R}_{h,p}$ and $\widehat{h}$ are defined in (4.5.1).

2 For $p \in [1,2)$,

$$\inf_{h \in H} \sup_{\substack{f \in P_p(\beta_0, L_0) \\ g \in P_p(\beta_1, L_1)}} E_{p(\epsilon,f,g)} \widehat{R}_{h,p} \asymp n^{-\frac{\beta_0}{q\beta_0+1}} \asymp \epsilon. \tag{C.3.5}$$

For $p \in [2, \infty)$,

$$\inf_{h \in H} \sup_{\substack{f \in P_p(\beta_0, L_0) \\ g \in P_p(\beta_1, L_1)}} E_{p(\epsilon,f,g)} \widehat{R}_{h,p} \asymp n^{-\frac{\beta_0}{2\beta_0+1}} \asymp \epsilon. \tag{C.3.6}$$

In fact, 1 is a direct result from the definition of $\widehat{R}_{h,p}$ and $\widehat{h}$. For any $l \in H$, we have

$$\left\| \widehat{f}_{\widehat{h}} - f \right\|_p^p \le \left\| \widehat{f}_{\widehat{h}} - \widehat{f}_{\widehat{h},l} \right\|_p^p + \left\| \widehat{f}_{\widehat{h},l} - \widehat{f}_l \right\|_p^p + \left\| \widehat{f}_l - f \right\|_p^p.$$

For any $l \in H$, we have

$$\begin{aligned} \left\| \widehat{f}_{\widehat{h}} - \widehat{f}_{\widehat{h},l} \right\|_p^p &\le \left[\left\| \widehat{f}_{l,\widehat{h}} - \widehat{f}_{\widehat{h}} \right\|_p^p - m_p(l,\widehat{h}) \right]_+ + m_p(l,\widehat{h}) \\ &\le \widehat{R}_{l,p} + m_p(\widehat{h}) \wedge \widehat{R}_{l,p} + \widehat{R}_{\widehat{h},p} \wedge 2\widehat{R}_{l,p}, \end{aligned}$$

and

$$\begin{aligned} \left\| \widehat{f}_{\widehat{h},l} - \widehat{f}_l \right\|_p^p &\le \left[\left\| \widehat{f}_{\widehat{h},l} - \widehat{f}_l \right\|_p^p - m_p(\widehat{h},l) \right]_+ + m_p(\widehat{h},l) \\ &\le \widehat{R}_{\widehat{h},p} + m_p(l) \wedge \widehat{R}_{\widehat{h},p} + \widehat{R}_{l,p} \wedge 2\widehat{R}_{l,p}. \end{aligned}$$

Therefore, we get

$$\left\| \widehat{f}_{\widehat{h}} - f \right\|_p^p \le 4\widehat{R}_{l,p} + \left\| \widehat{f}_l - f \right\|_p^p, \quad \forall l \in H.$$

For 2 , we can obtain an upper bound for $E\widehat{R}_{h,p}$ using the upper bounds for bias and variance parts of the kernel estimator with kernel K_h and $K_h - K_l$ in Lemmas C.1.1, C.1.2 and C.3.1.

Recall that we use p_ϵ to denote $p(\epsilon, f, g)$ for simplicity. For any $h \in H$, by the bias-variance decomposition, we can get

$$\begin{aligned} E_{p_\epsilon} \widehat{R}_{h,p} &\le \sup_{l \in H} \|B_{h,l}(p_\epsilon, t) - B_l(p_\epsilon, t)\|_p + E_{p_\epsilon} \sup_{l \in H} \left[\|\xi_{l,n}(p_\epsilon, t)\|_p - d_p(K_l) \right]_+ \\ &\quad + E_{p_\epsilon} \sup_{l \in H} \left[\|\xi_{h,l,n}(p_\epsilon, t)\|_p - d_p(K_h - K_l) \right]_+ + E_{p_\epsilon} m_p(h). \end{aligned}$$

For the first term above, by Proposition 4.3.1, we have

$$\sup_{l \in H} \|B_{h,l}(p_\epsilon, t) \quad B_l(p_\epsilon, t)\|_p \quad \|K\|_1 \|B_h(p_\epsilon, \ )\|_p . \tag{C.3.7}$$

For any $h \in H$, $f \in P_p(\beta_0, L_0)$, $g \in L_{p,d}(L_1)$, kernel $K \in K_\beta(L_K)$ with $\beta \quad \beta_0$ (and thus $K \in K_{\beta_0}(2L_K)$), we have

$$\|B_h(p_\epsilon, t)\|_p \quad (1 \quad \epsilon) \|B_h(f, t)\|_p + \epsilon \|B_h(g, t)\|_p \quad C \left(\sum_{j=1}^{d} h_j^{\beta_{0,j}} + \epsilon \right),$$

where C is some constant only depending on L_0, L_1, L_K. (e.g. C can be $(L_0 + L_1)(2^d L_K^d + 1)$; see (C.2.1)-(C.2.3).)

To control the second and third terms, by Lemma C.3.1 (it is easy to check all the assumptions are satisfied), we have

$$E_{p_\epsilon} \sup_{l \in H} \left[\|\xi_{l,n}(p_\epsilon, t)\|_p \quad d_p(K_l) \right]_+ \quad \delta_{n,p}, \quad E_{p_\epsilon} \sup_{l \in H} \left[\|\xi_{h,l,n}(p_\epsilon, t)\|_p \quad d_p(K_h \quad K_l) \right]_+ \quad \widetilde{\delta}_{n,p}$$

for all large enough n and all $p \in [1, \infty)$.

Next we bound the last term $E_{p_\epsilon} m_p(h)$. For $p \in [1, 2]$, we know for any $h \in H$,

$$\begin{aligned} E_{p_\epsilon} m_p(h) = m_p(h) &= r_p(K_h) + \sup_{l \in H} r_p(K_h \quad K_l) = C_p n^{-1/q} \left(\|K_h\|_p + \sup_{l \in H} \|K_h \quad K_l\|_p \right) \\ &\quad C_p n^{-1/q} \left(\|K_h\|_p + \sup_{l \in H} \|K_l\|_1 \|K_h\|_p \right) \\ &= C_p n^{-1/q} (1 + \|K\|_1) \|K_h\|_p \quad C (nV_h)^{-1/q} \end{aligned} \tag{C.3.8}$$

for some constant $C = C_p L_K^d (1 + L_K^d)$. Therefore, for any $p \in [1, 2]$, $h \in H$, $f \in P_p(\beta_0, L_0)$, $g \in P_p(\beta_1, L_1)$ (or just $L_{p,d}(L_1)$), we have

$$E_{p_\epsilon} \widehat{R}_{h,p} \quad C \left(\sum_{j=1}^{d} h_j^{\beta_{0,j}} + (nV_h)^{-1/q} + \epsilon \right) + \delta_{n,p} + \widetilde{\delta}_{n,p} \ . \ \sum_{j=1}^{d} h_j^{\beta_{0,j}} + (nV_h)^{-1/q} + \epsilon$$

as the terms $\delta_{n,p}$ and $\widetilde{\delta}_{n,p}$ can be negligible if we assume $V_{\max} \quad c_1/(\log n)^{p/2}$ for a sufficiently small constant c_1 in Theorem 4.5.1 (ii) (when $p = 2$). Therefore, based on the bounds of four terms in the upper bound of $E_{p_\epsilon} \widehat{R}_{h,p}$, choosing $h_j = h_j \quad n^{-\frac{\beta_0}{\beta_{0,j}(q\beta_0 + 1)}}$ (noting that the

oracle bandwidth h is contained in H), we get (C.3.5).

For $p \in (2, 1)$, by (C.3.3)-(C.3.4) in Lemma C.3.1, we get

$$E_{p_\epsilon} m_p(h) \;.\; r_p(K_h, p_\epsilon) + \sup_{l \in H} r_p(K_h \;\; K_l, p_\epsilon) + \zeta_{n,s} + \widetilde{\zeta}_{n,p}.$$

Again, the terms $\zeta_{n,p}$ and $\widetilde{\zeta}_{n,p}$ are negligible if we assume $V_{\max} \;\; c_1/(\log n)^{p/2}$ for a sufficiently small constant c_1. Some standard calculation shows for any $h \in H$,

$$\begin{aligned} r_p(K_h, p_\epsilon) \;.\; & n^{-1/2} \|K_h^2\| \;\; p_\epsilon \|_{p/2}^{1/2} \_ n^{-1/q} \|K_h\| \|_p \_ n^{-1/2} \|K_h\| \|_2 \\ & n^{-1/2} \|K_h\| \|_2 \|p_\epsilon\|_{p/2}^{1/2} \_ (nV_h)^{-1/q} \|K\| \|_p \_ (nV_h)^{-1/2} \|K\| \|_2 \\ & (nV_h)^{-1/2}. \end{aligned} \tag{C.3.9}$$

Here we have used Minkowski's integral inequality (or Young's convolution inequality) to get $\|f_1 \;\; f_2\|_s \;\; \|f_1\|_1 \|f_2\|_s$ for any $s \in [1, 1)$. Similarly, for any $h, l \in H$, we can get

$$\sup_{l \in H} r_p(K_h \;\; K_l, p_\epsilon) \;.\; \sup_{l \in H} n^{-1/2} (V_h \_ V_l)^{-1/2} \;\; (nV_h)^{-1/2} \tag{C.3.10}$$

as long as we notice that $\|K_h \;\; K_l\|_s \;\; \|K_l\|_1 \|K_h\|_s = \|K\|_1 \|K_h\|_s$ for any $s \in [1, 1)$ and any $h, l \in H$. Combining the above inequalities, we finally get

$$E_{p_\epsilon} \widehat{R}_{h,p} \;.\; \sum_{j=1}^{d} h_j^{\beta_{0,j}} + (nV_h)^{-1/2} + \epsilon.$$

Choosing $h_j = h_j \;\; n^{-\frac{\beta_0}{\beta_{0,j}(2\beta_0+1)}}$ (notice that the oracle bandwidth h is contained in H), we get (C.3.6). Combining the results in 1 and 2 , we complete the proof of Theorem 4.5.1. □

C.4 Proofs of Proposition 4.3.1, Theorem 4.3.2 and Theorem 4.3.3

C.4.1 Proof of Proposition 4.3.1

Proof of Proposition 4.3.1. Standard calculation shows that

$$\begin{aligned}
&\|E_f[K_h \quad K_l](X \quad t) \quad E_f K_l(X \quad t)\|_p \\
=&\|(K_h \quad K_l) \quad f \quad K_l \quad f\|_p \overset{(i)}{=} \|K_l \quad (K_h \quad f) \quad K_l \quad f\|_p \\
\overset{(ii)}{\leq}& \|K_l\|_1 \|K_h \quad f \quad f\|_p = \|K\|_1 \|E_f K_h(t \quad X) \quad f(t)\|_p ,
\end{aligned}$$

where (i) is due to Fubini's theorem and (ii) is due to Young's inequality. □

C.4.2 Proofs of Theorem 4.3.2 and Theorem 4.3.3

The proof of these two theorems follows the same strategy as the proof of Theorem 4.5.1:

1 First we prove

$$\left\| \widehat{f}_{\widehat{h}} \quad f \right\|_p \quad \inf_{h \in H} \left\{ 4\widehat{R}^{(i)}_{h,p} + \left\| \widehat{f}_h \quad f \right\|_p \right\}, \quad i = 1, 2. \tag{C.4.1}$$

2 For $i = 1, 2$, we show

$$\inf_{h \in H} \sup_{\substack{f \in P_{p,d}(\beta_0, L_0) \\ G}} E_{P(\epsilon, f, G)} \widehat{R}^{(i)}_{h,p} \; . \; n^{\frac{\beta_0}{q\beta_0 + 1}} \_ \epsilon^{\frac{q\beta_0}{q\beta_0 + 1}}, \quad \text{for } p \in [1, 2). \tag{C.4.2}$$

$$\inf_{h \in H} \sup_{\substack{f \in P_{p,d}(\beta_0, L_0) \\ G}} E_{P(\epsilon, f, G)} \widehat{R}^{(i)}_{h,p} \; . \; n^{\frac{\beta_0}{2\beta_0 + 1}} \_ \epsilon^{\frac{q\beta_0}{q\beta_0 + 1}}, \quad \text{for } p \in [2, \infty). \tag{C.4.3}$$

Proof of Theorem 4.3.2. As we discussed, it suffices to show (C.4.1)–(C.4.3) hold for $\widehat{R}^{(1)}_{h,p}$. To prove (C.4.1), we first notice that for any $l \in H$, we have the following decomposition of $k\widehat{f}_{\widehat{h}} \quad fk_p$:

$$\left\|\widehat{f}_{\widehat{h}} \quad f\right\|_p \quad \left\|\widehat{f}_{\widehat{h}} \quad \widehat{f}_{\widehat{h},l}\right\|_p + \left\|\widehat{f}_{\widehat{h},l} \quad \widehat{f}_l\right\|_p + \left\|\widehat{f}_l \quad f\right\|_p.$$

By the definition of $\widehat{R}^{(1)}_{h,p}$ (4.3.1) and $\widehat{h}$ (4.3.2), it is easy to check that for any $l \in H$,

$$\begin{aligned}\left\|\widehat{f}_{\widehat{h}} \quad \widehat{f}_{\widehat{h},l}\right\|_p \quad & \left[\left\|\widehat{f}_{l,\widehat{h}} \quad \widehat{f}_{\widehat{h}}\right\|_p \quad 2m_p(l,\widehat{h}) \quad (2+128D_p)m_{\epsilon,p}(l,\widehat{h})\right]_+ \\ & + 2m_p(l,\widehat{h}) + (2+128D_p)m_{\epsilon,p}(l,\widehat{h}) \\ & \widehat{R}^{(1)}_{l,p} + 2m_p(\widehat{h}) + (2+128D_p)m_{\epsilon,p}(\widehat{h}) \quad \widehat{R}^{(1)}_{l,p} + \widehat{R}^{(1)}_{\widehat{h},p} \quad 2\widehat{R}^{(1)}_{l,p},\end{aligned}$$

and

$$\begin{aligned}\left\|\widehat{f}_{\widehat{h},l} \quad \widehat{f}_l\right\|_p \quad & \left[\left\|\widehat{f}_{\widehat{h},l} \quad \widehat{f}_l\right\|_p \quad 2m_p(\widehat{h},l) \quad (2+128D_p)m_{\epsilon,p}(\widehat{h},l)\right]_+ \\ & + 2m_p(\widehat{h},l) + (2+128D_p)m_{\epsilon,p}(\widehat{h},l) \\ & \widehat{R}^{(1)}_{\widehat{h},p} + 2m_p(l) + (2+128D_p)m_{\epsilon,p}(l) \quad \widehat{R}^{(1)}_{\widehat{h},p} + \widehat{R}^{(1)}_{l,p} \quad 2\widehat{R}^{(1)}_{l,p}.\end{aligned}$$

Therefore, we obtain

$$\left\|\widehat{f}_{\widehat{h}} \quad f\right\|_p \quad 4\widehat{R}^{(1)}_{l,p} + \left\|\widehat{f}_l \quad f\right\|_p, \quad 8l \in H.$$

To obtain the desired results in 2 , we consider Huber's contamination model in the form (4.1.2). Our overall idea is to decompose the kernel estimator into two parts: the part consisting of the "clean" observations generated from the density f, and the part consisting of the contaminated observations generated from G. We try to get the desired results for the first part by applying Lemmas C.1.1 - C.3.1, and give the second part a bound related

to the contamination proportion ϵ. Following this idea, we first decompose $\|\widehat{f}_{h,l} - \widehat{f}_l\|_p$ as follows:

$$\begin{aligned}\left\|\widehat{f}_{h,l} - \widehat{f}_l\right\|_p \le & \frac{n_1}{n}\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\Big([K_h - K_l](t - X_i) - E_f[K_h - K_l](t - X)\Big)\right\|_p \\ & + \frac{n_1}{n}\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\Big(K_l(t - X_i) - E_f K_l(t - X)\Big)\right\|_p \\ & + \frac{n_1}{n}\left\|E_f[K_h - K_l](t - X) - E_f K_l(t - X)\right\|_p \\ & + \left\|\frac{1}{n}\sum_{i=n_1+1}^{n}[K_h - K_l](t - X_i)\right\|_p + \left\|\frac{1}{n}\sum_{i=n_1+1}^{n} K_l(t - X_i)\right\|_p \\ & \|\xi_{h,l,n_1}(f,t)\|_p + \|\xi_{l,n_1}(f,t)\|_p + \|B_{h,l}(f,t) - B_l(f,t)\|_p \\ & + \frac{n - n_1}{n}\big(\|K_h - K_l\|_p + \|K_l\|_p\big)\end{aligned} \tag{C.4.4}$$

for any $h, l \in H$, any $p \in [1, 1)$. Applying Bernstein's inequality (e.g. Theorem 2.8.4 in Vershynin (2018)) to $\frac{n-n_1}{n}$, we get

$$P\left(\frac{n - n_1}{n} - 2\epsilon\right) - \exp\Big(-\frac{1}{4}n\epsilon\Big). \tag{C.4.5}$$

Knowing that $P(\frac{n-n_1}{n} - 2\epsilon)$ is "small", we consider

$$E_{P_\epsilon}\widehat{R}^{(1)}_{h,p} = E_{P_\epsilon}\left(\widehat{R}^{(1)}_{h,p} : \frac{n - n_1}{n} < 2\epsilon\right) + E_{P_\epsilon}\left(\widehat{R}^{(1)}_{h,p} : \frac{n - n_1}{n} - 2\epsilon\right) \tag{C.4.6}$$

and mainly focus on the first term. Here for simplicity, we write $P_\epsilon = P(\epsilon, f, G)$ and will use the notation $\bar{E}(\ ) = E\big(\ : \frac{n-n_1}{n} < 2\epsilon\big) := E\|\ \|\mathbb{1}(\frac{n-n_1}{n} < 2\epsilon)\|$ in what follows. Conditional on the event $\|\frac{n-n_1}{n} < 2\epsilon\|$, we have

$$\frac{n - n_1}{n}\big(\|K_h - K_l\|_p + \|K_l\|_p\big) - 2\epsilon\big((V_h - V_l)^{1/q} + V_l^{1/q}\big)(\|K\|_1 + 1)\|K\|_p = 2m_{\epsilon,p}(h,l). \tag{C.4.7}$$

Therefore, by (C.4.4), (C.4.7) and the definition of $\widehat{R}^{(1)}_{h,p}$ (4.3.1), we get, for any $h \in H$,

$$\begin{aligned}E_{P_\epsilon}\Big(\widehat{R}^{(1)}_{h,p} : \frac{n - n_1}{n} < 2\epsilon\Big) - & \sup_{l \in H}\|B_{h,l}(f,t) - B_l(f,t)\|_p + \bar{E}_{P_\epsilon}\sup_{l \in H}\Big[\|\xi_{l,n_1}(f,t)\|_p \\ & 2d_p(K_l) + \|\xi_{h,l,n_1}(f,t)\|_p\big)\|_p - 2d_p(K_h - K_l) - 128D_p m_{\epsilon,p}(h,l)\Big]_+ \\ & + 2\bar{E}_{P_\epsilon} m_p(h) + (2 + 128D_p)m_{\epsilon,p}(h).\end{aligned} \tag{C.4.8}$$

It is easy to check that

$$m_{\epsilon,p}(h) = \sup_{l \in H} m_{\epsilon,p}(l,h) \quad 2(L_K^d+1)^2 \epsilon V_h^{1/q}, \quad \forall h \in H, \tag{C.4.9}$$

$$\sup_{l \in H} \|B_{h,l}(f,t) \quad B_l(f,t)\|_p \quad \|K\|_1 \|B_h(f,t)\|_p \quad 2^d L_0 L_K^{2d} \sum_{j=1}^{d} h_j^{\beta_{0,j}}, \quad \forall h \in H, \tag{C.4.10}$$

for any $f \in \mathcal{N}_{p,d}(\beta_0, L_0)$ and any $K \in \mathcal{K}_\beta(L_K)$ with $\beta_0 \quad \beta$ (see Proposition 4.3.1 and Lemma C.1.1, and notice that $K \in \mathcal{K}_{\beta_0}(2L_K)$).

Deriving the bounds for the remaining terms in (C.4.8) is more involved , as we cannot directly apply the results in Lemma C.3.1 to $\xi_{l,n_1}(f,t)$ and $\xi_{h,l,n_1}(f,t)$. We use (C.3.1) as an example to elaborate this point. Recall that the term $d_p(K_h)$ in (C.3.1) (plug-in $s = p$) , equals $\widehat{r}_p(K_h)$ for $p \in (2, \infty)$, which, by its definition (4.3.4), utilizes the whole data set $\{X_1, \ldots, X_n\}$ (including the contaminated data) and thus does not align with $\xi_{l,n_1}(f,t)$ (defined for only the clean data generated from P_f) in our setting. We take into account the existence of contamination and generalize Lemma C.3.1 to get the following lemma. Its proof is left at the end of this proof.

Lemma C.4.1. *Under the conditions of Theorem 4.3.2, for any $p \in [1, \infty)$, any $h \in H$, we have*

(i)

$$\bar{E}_{P_\epsilon} \sup_{l \in H} \Big[\|\xi_{l,n_1}(f,t)\|_p \quad 2d_p(K_l) + \|\xi_{h,l,n_1}(f,t)\|_p \quad 2d_p(K_h \quad K_l) \\ 128 D_p m_{\epsilon,p}(h,l) \Big]_+ \quad . \quad \delta_{n,p} + \widetilde{\delta}_{n,p},$$

where $\delta_{n,p}$, $\widetilde{\delta}_{n,p}$ are defined in Lemma C.3.1.

(ii)

$$\bar{E}_{P_\epsilon} m_p(h) \quad . \quad \begin{cases} (nV_h)^{-1/q}, & p \in [1,2] \\ (nV_h)^{-1/2} + \epsilon V_h^{-1/q}, & p \in (2, \infty). \end{cases}$$

Combining the bounds (C.4.8)-(C.4.10) with Lemma C.4.1, we obtain that for any $p \in [1, \infty)$, $h \in \mathcal{H}$,

$$E_{P_\epsilon}\left(\widehat{R}^{(1)}_{h,p} : \frac{n - n_1}{n} < 2\epsilon\right) \lesssim \sum_{j=1}^{d} h_j^{\beta_{0,j}} + \epsilon V_h^{-1/q} + (nV_h)^{-1/(q-2)},$$

since the terms $\delta_{n,p}$ and $\widetilde{\delta}_{n,p}$ are negligible if we assume $V_{\min} \geq c_1/(\log n)^{p/2}$ for a sufficiently small constant c_1 for $p \in (2, \infty)$.

On the other hand, on the event $\{\frac{n - n_1}{n} \leq 2\epsilon\}$, we may just derive some general upper bounds of $\widehat{R}^{(1)}_{h,p}$ for any $h \in \mathcal{H}$ since we know that $P(\frac{n - n_1}{n} \leq 2\epsilon)$ is "small". It is easy to check that for any $p \in [1, \infty)$, any $h, l \in \mathcal{H}$, the following inequalities hold.

$$\begin{aligned}
&\|\widehat{f}_l\|_p = \left\|\frac{1}{n}\sum_{i=1}^{n} K_l(t - X_i)\right\|_p \leq \|K_l\|_p \leq V_l^{-1/q}\|K\|_p,\\
&\|\widehat{f}_{h,l}\|_p \leq \|K_h - K_l\|_p \leq (V_h - V_l)^{-1/q}\|K\|_1\|K\|_p,\\
&d_p(K_h) \lesssim n^{-1}\left\|\sum_{i=1}^{n} K_h^2(t - X_i)\right\|_{p/2}^{1/2} - n^{-1/q}\|K_h\|_p - n^{-1/2}\|K_h\|_2,\\
&\qquad n^{-1/2}\|K_h\|_p - n^{-1/q}\|K_h\|_p - n^{-1/2}\|K_h\|_2 \leq n^{-1/2}V_h^{-1/q} - (nV_h)^{-1/(q-2)},\\
&d_p(K_h - K_l) \lesssim n^{-1/2}(V_h - V_l)^{-1/q} - [n(V_h - V_l)]^{-1/(q-2)},\\
&m_p(h) = \sup_{l\in\mathcal{H}} d_p(K_h - K_l) + d_p(K_h) \lesssim n^{-1/2}V_h^{-1/q} - (nV_h)^{-1/(q-2)}.
\end{aligned}$$

Therefore, for any $p \in [1, \infty)$, any $h \in \mathcal{H}$, we have

$$\begin{aligned}
\widehat{R}^{(1)}_{h,p} &\leq \sup_{l\in\mathcal{H}}\left(\left\|\widehat{f}_{h,l}\right\|_p + \left\|\widehat{f}_l\right\|_p\right) + 2m_p(h) + (2 + 128D_p)m_{\epsilon,p}(h)\\
&\lesssim \sup_{l\in\mathcal{H}} V_l^{-1/q} + n^{-1/2}V_h^{-1/q} + (nV_h)^{-1/(q-2)} + \epsilon V_h^{-1/q} \lesssim n^{1/q},
\end{aligned}$$

where we have used $V_{\min} \gtrsim 1/n$. Then by (C.4.6), we have for any $\epsilon \geq \frac{8\log n}{qn}$,

$$\begin{aligned}
E_{P_\epsilon}\widehat{R}^{(1)}_{h,p} &\lesssim \sum_{j=1}^{d} h_j^{\beta_{0,j}} + \epsilon V_h^{-1/q} + (nV_h)^{-1/(q-2)} + n^{1/q}\exp\left\{-\frac{1}{4}n\epsilon\right\}\\
&\lesssim \sum_{j=1}^{d} h_j^{\beta_{0,j}} + \epsilon V_h^{-1/q} + (nV_h)^{-1/(q-2)}, \quad \forall h \in \mathcal{H}.
\end{aligned}$$

Choosing h h where h_j $n^{\frac{\beta_0}{\beta_{0,j}((q-2)\beta_0+1)}} \asymp \epsilon^{\frac{q\beta_0}{\beta_{0,j}(q\beta_0+1)}}$, we achieve the desired bounds (C.4.2) and (C.4.3) for $\widehat{R}^{(1)}_{h,p}$. Here, it remains to check if the oracle bandwidth h is contained in H. In fact, one can check the condition $h \in H$ is naturally satisfied for $p \in [1,2)$. But for $p \in [2, \infty)$, we need to further assume

$$\epsilon^{\frac{q}{q\beta_0+1}} \cdot (\log n)^{p/2}, \quad i.e. \quad \epsilon \cdot (\log n)^{\frac{p(\beta_0+1)}{2}\frac{1}{q}},$$

to guarantee V_h $n^{\frac{1}{(q-2)\beta_0+1}} \asymp \epsilon^{\frac{q}{q\beta_0+1}} \cdot V_{\max} \cdot (\log n)^{p/2}$ and thus $h \in H$.

For the case when $\epsilon < \frac{8\log n}{qn}$, one can show that there always exists some distribution $\widetilde{G}$ such that

$$(1-\epsilon)P_f + \epsilon G = (1-\frac{8}{q}\frac{\log n}{n})P_f + \frac{8}{q}\frac{\log n}{n}\widetilde{G}.$$

(e.g. pick $\widetilde{G} := \frac{c(n)-\epsilon}{c(n)}P_f + \frac{\epsilon}{c(n)}G$, $c(n) = \frac{8\log n}{qn}$.) Therefore, we can equivalently think of the contamination proportion as $\frac{8}{q}\frac{\log n}{n}$ (with a different contamination distribution $\widetilde{G}$). Similarly, we can get

$$\begin{aligned}\inf_{h\in H}\sup_{f\in P_{p,d}(\beta_0,L_0)\atop G} E_{P(\epsilon,f,G)}\widehat{R}^{(1)}_{h,p} \cdot n^{\frac{\beta_0}{(q-2)\beta_0+1}} &\asymp \left(\frac{8\log n}{qn}\right)^{\frac{q\beta_0}{q\beta_0+1}} \\ n^{\frac{\beta_0}{(q-2)\beta_0+1}} \cdot n^{\frac{\beta_0}{(q-2)\beta_0+1}} &\asymp \epsilon^{\frac{q\beta_0}{q\beta_0+1}},\end{aligned} \tag{C.4.11}$$

which completes our proof. □

Proof of Lemma C.4.1. (i) We define the following quantities with n replaced by n_1 in (4.3.3)-(4.3.5). Let

$$\bar{r}_s(U) := C_s n_1^{1/s-1}\|U\|_s, s \in [1,2], \tag{C.4.12}$$

where $C_s = 128$ for $s \in [1,2)$ and $C_2 = 100/3$. When $s \in (2, \infty)$, we set

$$\begin{aligned}\bar{r}_s(U, p_X) := 32D_s\Bigg\{n_1^{1/2}\Bigg(\int\Bigg[\int U^2(t-x)p_X(x)dx\Bigg]^{s/2}dt\Bigg)^{1/s} \\ + 2n_1^{1/s-1}\|U\|_s\Bigg\} - 32n_1^{1/2}\|U\|_2,\end{aligned} \tag{C.4.13}$$

$$\bar{\widehat{r}}_s(U) := 32D_s\Big\{n_1{}^{1/2}\Big(\int\Big[\frac{1}{n_1}\sum_{i=1}^{n_1}U^2(t \quad X_i)\Big]^{s/2}dt\Big)^{1/s} + 2n_1^{1/s\ \ 1}\,k\,U\,k_s\Big\}\_ 32n_1{}^{1/2}\,k\,U\,k_2\ . \tag{C.4.14}$$

Define

$$\bar{d}_s(U) := \begin{cases}\bar{r}_s(U), & s\ 2\ [1,2]\\ \bar{\widehat{r}}_s(U), & s\ 2\ (2,\ 1\ ).\end{cases} \tag{C.4.15}$$

Conditional on the event $f\frac{n\ \ n_1}{n} < 2\epsilon g$, we have $n_1 > (1 \quad 2\epsilon)n \quad n/2$ assuming $\epsilon \quad 1/4$. (One can check that our proof is valid as long as $\epsilon \quad C < 1$, though some very mild modification might be needed.) Then for $p\ 2\ [1,2]$, by Lemma C.3.1, we have

$$\bar{E}_{P_\epsilon}\sup_{l2H}\Big[k\xi_{l,n_1}(f,t)k_p \quad 2d_p(K_l)\Big]_+ \quad \bar{E}_f\sup_{l2H}\Big[k\xi_{l,n_1}(f,t)k_p \quad \bar{d}_p(K_l)\Big]_+ \quad \bar{E}\delta_{n_1,p} \quad \delta_{n,p},$$

given the fact that $\bar{d}_p(K_l) < 2d_p(K_l)$ (for $p\ 2\ [1,2]$) and $\delta_{n_1,p} \quad \delta_{n,p}$ if we know $n/2 < n_1 \quad n$. Similarly, we have

$$\bar{E}_{P_\epsilon}\sup_{l2H}\Big[k\xi_{h,l,n_1}(f,t)k_p \quad 2d_p(K_h \quad K_l)\Big]_+\ .\ \widetilde{\delta}_{n,p}.$$

For $p\ 2\ (2,\ 1\ )$, conditional on $f\frac{n\ \ n_1}{n} < 2\epsilon g$, for any $l\ 2\ H$, we have

$$\begin{aligned}
2d_p(K_l) = 2\widehat{r}_p(K_l) &= 32D_p\Big\{2n^{\ 1}\Big\|\sum_{i=1}^n K_l^2(t \quad X_i)\Big\|_{p/2}^{1/2} + 2\ \ 2n^{1/s\ \ 1}\,k\,K_l\,k_p\Big\}_ 32\ \ 2n^{\ 1/2}\,k\,K_l\,k_2\\
&\quad 32D_p\Big\{n_1{}^1\Big\|\sum_{i=1}^n K_l^2(t \quad X_i)\Big\|_{p/2}^{1/2} + 2n_1^{1/s\ \ 1}\,k\,K_l\,k_p\Big\}_ 32n_1{}^{1/2}k\,K_l\,k_2\\
&\quad \bar{\widehat{r}}_p(K_l) \quad 32D_p n_1{}^1\Big\|\sum_{i=n_1+1}^n K_l^2(t \quad X_i)\Big\|_{p/2}^{1/2}\\
&\quad \bar{\widehat{r}}_p(K_l) \quad 32D_p(n/2)\ \ ^1(n \quad n_1)\,\big\|K_l^2\big\|_{p/2}^{1/2} \quad (\text{as } p\frac{}{n \quad n_1} \quad (n \quad n_1))\\
&\quad \bar{\widehat{r}}_p(K_l) \quad 128D_p\epsilon kK_lk_p \quad \bar{\widehat{r}}_p(K_l) \quad 128D_p\epsilon(V_l)\ \ ^{1/q}kKk_p.
\end{aligned}$$

Similarly, for any $h, l \in H$, we have

$$2d_p(K_h \quad K_l) \quad \bar{\widehat{r}}_p(K_h \quad K_l) \quad 128D_p\epsilon k K_h \quad K_l k_p$$
$$\bar{\widehat{r}}_p(K_l) \quad 128D_p\epsilon(V_l \_ V_h)^{1/q} k K k_1 k K k_p.$$

Therefore, for any $h \in H$, we obtain

$$\bar{E}_{P_\epsilon} \sup_{l \in H} \left[k\xi_{l,n_1}(f,t) k_p \quad 2d_p(K_l) + k\xi_{h,l,n_1}(f,t) k_p \quad 2d_p(K_h \quad K_l) \quad 128D_p m_{\epsilon,p}(h,l) \right]_+$$
$$\bar{E}_f \sup_{l \in H} \left[k\xi_{l,n_1}(f,t) k_p \quad \bar{\widehat{r}}_p(K_l) \right]_+ + \bar{E}_f \sup_{l \in H} \left[k\xi_{h,l,n_1}(f,t) k_p \quad \bar{\widehat{r}}_p(K_h \quad K_l) \right]_+$$
$$\bar{E}\delta_{n_1,p} + \bar{E}\widetilde{\delta}_{n_1,p} \quad \delta_{n,p} + \widetilde{\delta}_{n,p}.$$

(ii) For $p \in [1,2]$, similar to (C.3.8), we know for any $h \in H$,

$$\bar{E}_{P_\epsilon} m_p(h) = m_p(h) \ . \ (nV_h)^{1/q}.$$

For $p \in (2, \infty)$ and any $h \in H$,

$$\bar{E}_{P_\epsilon} d_p(K_h) = \bar{E}_{P_\epsilon} \widehat{r}_p(K_h) \ . \ n^{-1} \bar{E}_{P_\epsilon} \left\| \sum_{i=1}^n K_h^2(t \quad X_i) \right\|_{p/2}^{1/2} + n^{-1/q} k K_h k_p + n^{-1/2} k K_h k_2$$
$$\bar{E}_f n_1^{-1/2} \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} K_h^2(t \quad X_i) \right\|_{p/2}^{1/2} + n^{-1} \bar{E}_G \left\| \sum_{i=n_1+1}^{n} K_h^2(t \quad X_i) \right\|_{p/2}^{1/2}$$
$$+ (nV_h)^{-1/q} k K k_p + (nV_h)^{-1/2} k K k_2$$
$$\ . \ \bar{E}_f \bar{\widehat{r}}_p(K_h) + \bar{E}\frac{n \quad n_1}{n} k K_h^2 k_{p/2}^{1/2} + (nV_h)^{-1/2}$$
$$\overset{(a)}{\ .} \ r_p(K_h, f) + \zeta_{n,p} + \epsilon k K_h k_p + (nV_h)^{-1/2}$$
$$\overset{(b)}{\ .} \ (nV_h)^{-1/2} + \epsilon V_h^{-1/q}.$$

Here, (a) is due to the fact that $\bar{E}_f \bar{\widehat{r}}_p(K_h) \;.\; r_p(K_h, f) + \zeta_{n,p}$ by (C.3.3) from Lemma C.3.1; (b) follows from $r_p(K_h, f) \;.\; (nV_h)^{1/2}$ by (C.3.9). Similarly, for any $h \in H$, we have

$$\begin{aligned}
\bar{E}_{P_\epsilon} \sup_{l \in H} d_p(K_h - K_l) &\;.\; \bar{E}_f \sup_{l \in H} \bar{\widehat{r}}_p(K_h - K_l) + \sup_{l \in H} \epsilon \|K_h - K_l\|_{p/2}^2 \|K_{p/2}^{1/2} + \sup_{l \in H} n^{-1/2}(V_h - V_l)^{1/2} \\
&\overset{(c)}{\;.\;} \sup_{l \in H} r_p(K_h - K_l, f) + \widetilde{\zeta}_{n,p} + \sup_{l \in H} \epsilon \|K_h K_h - K_l K_l\|_p + \sup_{l \in H} n^{-1/2}(V_h - V_l)^{1/2} \\
&\overset{(d)}{\;.\;} \sup_{l \in H} n^{-1/2}(V_h - V_l)^{1/2} + \sup_{l \in H} \epsilon (V_h - V_l)^{1/q} \\
&\;.\; (nV_h)^{-1/2} + \epsilon V_h^{1/q}.
\end{aligned}$$

Here, (c) follows from $\bar{E}_f \sup_{l \in H} \bar{\widehat{r}}_p(K_h - K_l) \;.\; \sup_{l \in H} r_p(K_h - K_l, f) + \widetilde{\zeta}_{n,p}$ by (C.3.4) in Lemma C.3.1; (d) holds because $r_p(K_h - K_l, f) \;.\; n^{-1/2}(V_h - V_l)^{1/2}$ by (C.3.9)–(C.3.10). Therefore, we finally get

$$\bar{E}_{P_\epsilon} m_p(h) = \bar{E}_{P_\epsilon} d_p(K_h) + \bar{E}_{P_\epsilon} \sup_{l \in H} d_p(K_h - K_l) \;.\; (nV_h)^{-1/2} + \epsilon V_h^{1/q}, \quad \forall h \in H, \quad \text{(C.4.16)}$$

which completes our proof of the lemma. □

Proof of Theorem 4.3.3. As mentioned in Section C.4.2, it suffices to show (C.4.1)-(C.4.3) hold for $\widehat{R}_{h,p}^{(2)}$.

1 . By the triangular inequality, we have

$$\left\|\widehat{f}_{\widehat{h}} - f\right\|_p \le \left\|\widehat{f}_{\widehat{h}} - \widehat{f}_{\widehat{h},l}\right\|_p + \left\|\widehat{f}_{\widehat{h},l} - \widehat{f}_l\right\|_p + \left\|\widehat{f}_l - f\right\|_p, \quad \forall l \in H.$$

By the definitions of $\widehat{R}_{h,p}^{(2)}$ and $\widehat{h}$, we obtain that for any $l \in H$,

$$\left\|\widehat{f}_{\widehat{h}} - \widehat{f}_{\widehat{h},l}\right\|_p \le \left[\left\|\widehat{f}_{\widehat{h},l} - \widehat{f}_{\widehat{h}}\right\|_p - m_b(l)\right]_+ + m_b(l) \le \widehat{R}_{\widehat{h},p}^{(2)} + m_b(l) \le \widehat{R}_{\widehat{h},p}^{(2)} + \widehat{R}_{l,p}^{(2)} \le 2\widehat{R}_{l,p}^{(2)},$$

and

$$\left\|\widehat{f}_{\widehat{h},l} - \widehat{f}_l\right\|_p \le \left[\left\|\widehat{f}_{l,\widehat{h}} - \widehat{f}_l\right\|_p - m_b(\widehat{h})\right]_+ + m_b(\widehat{h}) \le \widehat{R}_{l,p}^{(2)} + m_b(\widehat{h}) \le \widehat{R}_{l,p}^{(2)} + \widehat{R}_{\widehat{h},p}^{(2)} \le 2\widehat{R}_{l,p}^{(2)}.$$

Combining the above inequalities, we achieve (C.4.1) for $\widehat{R}_{h,p}^{(2)}$.

2 . Similar to (C.4.4), for any $p \in [1, 1)$, any $h, l \in H$, we have the following decomposition of $K \widehat{f}_{h,l} - \widehat{f}_h K_p$:

$$
\begin{aligned}
\left\|\widehat{f}_{h,l} \quad \widehat{f}_h\right\|_p \quad & \frac{n_1}{n}\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\Big([K_h \quad K_l](t \quad X_i) \quad E_f[K_h \quad K_l](t \quad X)\Big)\right\|_p \\
& + \frac{n_1}{n}\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\Big(K_h(t \quad X_i) \quad E_f K_h(t \quad X)\Big)\right\|_p \\
& + \frac{n_1}{n}\left\|E_f[K_h \quad K_l](t \quad X) \quad E_f K_h(t \quad X)\right\|_p \\
& + \left\|\frac{1}{n}\sum_{i=n_1+1}^{n}[K_h \quad K_l](t \quad X_i)\right\|_p + \left\|\frac{1}{n}\sum_{i=n_1+1}^{n}K_h(t \quad X_i)\right\|_p \\
& k\xi_{h,l,n_1}(f,t)k_p + k\xi_{h,n_1}(f,t)k_p + kB_{h,l}(f,t) \quad B_h(f,t)k_p \\
& + \frac{n \quad n_1}{n}\big(\, k\, K_h \quad K_l\, k_p + k\, K_h\, k_p\,\big).
\end{aligned}
\tag{C.4.17}
$$

By Proposition 4.3.1 and the result of Lemma C.1.1, we have

$$
\begin{aligned}
kB_{h,l}(f,t) \quad B_h(f,t)k_p \quad & kKk_1\, kB_l(f,t)k_p \\
& kKk_1 L_0(2L_K)^d \sum_{j=1}^{d} l_j^{\beta_{0,j}} \quad m_b(l), \quad 8h, l \ 2\ H,
\end{aligned}
$$

for any $f \ 2\ N_{p,d}(\beta_0, L_0)$ and $K \ 2\ K_\beta(L_K)$ with $\beta_0 \quad \beta$ (and thus $K \ 2\ K_{\beta_0}(2L_K)$). Therefore, by the definition of $\widehat{R}_{h,p}^{(2)}$, we have

$$
\begin{aligned}
E_{P_\epsilon}\widehat{R}_{h,p}^{(2)} \quad & E_{P_\epsilon}\sup_{l2H}[k\xi_{h,l,n_1}(f,t)k_p + k\xi_{h,n_1}(f,t)k_p] \\
& + E_{P_\epsilon}\sup_{l2H}\frac{n \quad n_1}{n}\big(\, k\, K_h \quad K_l\, k_p + k\, K_h\, k_p\,\big) + m_b(h), \quad 8h \ 2\ H.
\end{aligned}
$$

Notice that for any $p \ 2\ [1, \ 1\ )$, $h \ 2\ H$, it holds

$$
\begin{aligned}
E_{P_\epsilon}\sup_{l2H}\frac{n \quad n_1}{n}\big(\, k\, K_h \quad K_l\, k_p + k\, K_h\, k_p\,\big) &= \epsilon\big(\sup_{l2H} k\, K_h \quad K_l\, k_p + k\, K_h\, k_p\,\big) \\
& \epsilon\big(\sup_{l2H} k\, K_l\, k_1 k\, K_h\, k_p + k\, K_h\, k_p\,\big) \ . \ \epsilon V_h^{\ 1/q}.
\end{aligned}
$$

In Lemma C.1.2, let $p_X = f$, then for any $p \ 2\ [1, \ 1\ )$, $h \ 2\ H$, we have

$$
E_{P_\epsilon}k\xi_{h,n_1}(f,t)k_p = E\left\{E_f\left[k\xi_{h,n_1}(f,t)k_p\,\middle|\,n_1\right]\right\} \ . \ (nV_h)^{\ 1/(q\_2)}.
$$

It remains to obtain an upper bound of $E_{P_\epsilon}\sup_{l2H}k\xi_{h,l,n_1}(f,t)k_p$. Using the results of Lemma C.3.1 and the concentration inequality on n_1 (C.4.5), we can get the following lemma.

Lemma C.4.2. *For any $h \in H$, any $\epsilon \in [0, 1/4]$, we have*

$$E_{P_\epsilon} \sup_{l \in H} |\xi_{h,l,n_1}(f,t)| \lesssim (nV_h)^{1/(q-2)} + n^{1/q} \exp\left\{-\frac{1}{4}n\epsilon\right\}.$$

Combining the above inequalities, we obtain that when $\epsilon > \frac{8\log n}{qn}$,

$$E_{P_\epsilon} |\widehat{R}^{(2)}_{h,p}| \lesssim \sum_{j=1}^{d} h_j^{\beta_{0,j}} + \epsilon V_h^{1/q} + (nV_h)^{1/(q-2)} + n^{1/q} \exp\left\{-\frac{1}{4}n\epsilon\right\}$$
$$\lesssim \sum_{j=1}^{d} h_j^{\beta_{0,j}} + \epsilon V_h^{1/q} + (nV_h)^{1/(q-2)}, \quad \forall h \in H.$$

Choosing $h \asymp h$ ($h \in H$ is guaranteed by $\epsilon \gtrsim (\log n)^{\frac{p(\beta_0+1)-1}{2}}$) where $h_j \asymp n^{-\frac{\beta_0}{\beta_{0,j}((q-2)\beta_0+1)}} \wedge \epsilon^{\frac{q\beta_0}{\beta_{0,j}(q\beta_0+1)}}$, we get (C.4.2) and (C.4.3) for $\widehat{R}^{(2)}_{h,p}$.

When $\epsilon \le \frac{8\log n}{qn}$, similar to what we did in the proof of Theorem 4.3.2, we can find some distribution $\widetilde{G}$ such that

$$(1-\epsilon)P_f + \epsilon G = \left(1 - \frac{8}{q}\frac{\log n}{n}\right)P_f + \frac{8}{q}\frac{\log n}{n}\widetilde{G}$$

(e.g. pick $\widetilde{G} := \frac{c(n)-\epsilon}{c(n)}P_f + \frac{\epsilon}{c(n)}G$, $c(n) = \frac{8\log n}{qn}$). Therefore, we may equivalently think of the contamination proportion as $\frac{8}{q}\frac{\log n}{n}$ (with a different contamination distribution $\widetilde{G}$), and still get (C.4.2) and (C.4.3) for $\widehat{R}^{(2)}_{h,p}$ (see (C.4.11)). □

Proof of Lemma C.4.2. We consider

$$E_{P_\epsilon} \sup_{l \in H} |\xi_{h,l,n_1}(f,t)| = E_{P_\epsilon}\left(\sup_{l \in H} |\xi_{h,l,n_1}(f,t)| : \frac{n-n_1}{n} < 2\epsilon\right)$$
$$+ E_{P_\epsilon}\left(\sup_{l \in H} |\xi_{h,l,n_1}(f,t)| : \frac{n-n_1}{n} \ge 2\epsilon\right)$$

We remind the readers of the notation $\bar{E}(\cdot) = E\big(\cdot : \frac{n-n_1}{n} < 2\epsilon\big) := Ef(\cdot)\mathbb{1}(\frac{n-n_1}{n} < 2\epsilon)g$. Recall the definition of $\bar{r}_s(U)$, $\bar{r}_s(U, p_X)$, $\bar{\widehat{r}}_s(U)$, $\bar{d}_s(U)$ in (C.4.12)-(C.4.15) for all $s \in [1, 1)$.

Conditional on the event $\left\{\frac{n - n_1}{n} < 2\epsilon\right\}\Big($ $\left\{n/2$ n_1 $n\right\}\Big)$, by Lemma C.3.1 (i) and (ii), for any $p \in [1,2]$, any $h \in \mathcal{H}$, we have

$$\begin{aligned}\bar{E}_{P_\epsilon}\sup_{l\in\mathcal{H}}\|\xi_{h,l,n_1}(f,t)\|_p &\leq \bar{E}_f\sup_{l\in\mathcal{H}}\left[\|\xi_{h,l,n_1}(f,t)\|_p - \bar{r}_p(K_h - K_l)\right]_+ + \sup_{l\in\mathcal{H}}\bar{r}_p(K_h - K_l)\\ &\leq \bar{E}\widetilde{\delta}_{n_1,p} + \sup_{l\in\mathcal{H}} C_p(n/2)^{-1/q}\|K_l\|_1\|K_h\|_p\\ &\lesssim \widetilde{\delta}_{n,p} + (nV_h)^{-1/q}.\end{aligned}$$

For any $p \in (2, \infty)$, any $h \in \mathcal{H}$, we have

$$\begin{aligned}\bar{E}_{P_\epsilon}\sup_{l\in\mathcal{H}}\|\xi_{h,l,n_1}(f,t)\|_p &\leq \bar{E}_f\sup_{l\in\mathcal{H}}\left[\|\xi_{h,l,n_1}(f,t)\|_p - \hat{\bar{r}}_p(K_h - K_l)\right]_+ + \bar{E}_f\sup_{l\in\mathcal{H}}\hat{\bar{r}}_p(K_h - K_l)\\ &\lesssim \bar{E}\widetilde{\delta}_{n_1,p} + \bar{E}\sup_{l\in\mathcal{H}}\bar{r}_p(K_h - K_l, f) + \bar{E}\widetilde{\zeta}_{n_1,p} \quad \text{(by (C.3.2) and (C.3.4))}\\ &\lesssim \widetilde{\delta}_{n,p} + \sup_{l\in\mathcal{H}} r_p(K_h - K_l, f) + \widetilde{\zeta}_{n,p}\\ &\lesssim \widetilde{\delta}_{n,p} + (nV_h)^{-1/2} + \widetilde{\zeta}_{n,p}. \quad \text{(by (C.3.10))}\end{aligned}$$

For $p \in [2, \infty)$, if we set $V_{\max} \asymp c_1/(\log n)^{p/2}$ for a sufficiently small constant c_1, then the terms $\widetilde{\delta}_{n,p}$ and $\widetilde{\zeta}_{n,p}$ will be dominated by $(nV_h)^{-1/(q-2)}$. Therefore, we have

$$\bar{E}_{P_\epsilon}\sup_{l\in\mathcal{H}}\|\xi_{h,l,n_1}(f,t)\|_p \lesssim (nV_h)^{-1/(q-2)}, \quad \forall h \in \mathcal{H}.$$

Notice that

$$\begin{aligned}\sup_{l\in\mathcal{H}}\|\xi_{h,l,n_1}(f,t)\|_p &= \sup_{l\in\mathcal{H}}\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\Big([K_h - K_l](t - X_i) - E_f[K_h - K_l](t - X)\Big)\right\|_p\\ &\leq \sup_{l\in\mathcal{H}}\|K_h - K_l\|_p + \sup_{l\in\mathcal{H}}\|E_f[K_h - K_l](t - X)\|_p\\ &\leq \sup_{l\in\mathcal{H}}\|K_h - K_l\|_p + \sup_{l\in\mathcal{H}}\|K_h - K_l\|_p\|f\|_1 \quad \text{(by Young's inequality)}\\ &\leq 2\sup_{l\in\mathcal{H}}\|K_l\|_1\|K_h\|_p \lesssim V_h^{-1/q} \lesssim n^{1/q}, \quad \forall h \in \mathcal{H}.\end{aligned}$$

The above inequalities and the concentration inequality on n_1 (C.4.5) yield the result in Lemma C.4.2. □

C.5 Proof of Theorem 4.3.4

Proof of Theorem 4.3.4. Let h be the oracle bandwidth. We know $h_j \asymp n^{-\frac{\beta_0}{\beta_{0,j}((q-2)\beta_0+1)}} \asymp \epsilon^{\frac{q\beta_0}{\beta_{0,j}(q\beta_0+1)}}$ for $j = 1, \ldots, d$, and $V_h \asymp n^{-\frac{1}{(q-2)\beta_0+1}} \asymp \epsilon^{\frac{q}{q\beta_0+1}}$. Then under the assumption $\epsilon \gtrsim (\log n)^{-\frac{p(\beta_0+1)-1}{2}}$, we can guarantee that $h \in \mathcal{H}$, which is defined in (4.3.14). Our strategy is to prove $\widehat{h} \asymp h$ (i.e. $V_{\widehat{h}} \asymp V_h$) with high probability, then conditional on event $\{\widehat{h} \asymp h\}$, it is easy to show $E\|\widehat{f}_{\widehat{h}} - f\|_p \lesssim E\|\widehat{f}_h - f\|_p$ by the definition of $\widehat{h}$ (4.3.15). The proof of $P(\widehat{h} > h)$ being "small" relies essentially on the concentration inequalities in Lemma C.3.1. In order to use the results in Lemma C.3.1 (requiring the existence of density), we consider Huber's contamination model in the form (4.1.2).

We will use $\bar{P}$ and $\bar{E}$ to denote probability and expectation conditional on the event $\{\frac{n-n_1}{n} < 2\epsilon\}$. That is, $\bar{E}(\cdot) = E(\cdot)\mathbf{1}(\frac{n-n_1}{n} < 2\epsilon)\}$ (as defined before), $\bar{P}(\cdot) = E\mathbf{1}(\cdot)\mathbf{1}(\frac{n-n_1}{n} < 2\epsilon)\}$.

For any $p \in [1, \infty)$, any bandwidth $h > 0$, we always have the following decomposition of $\|\widehat{f}_h - f\|_p$:

$$\begin{aligned}\|\widehat{f}_h - f\|_p &\le \left\|\frac{1}{n}\sum_{i=1}^{n_1}\Big(K_h(t-X_i) - E_f K_h(t-X)\Big)\right\|_p + \left\|\frac{n_1}{n}\Big(E_f K_h(t-X) - f(t)\Big)\right\|_p \\ &\quad + \left\|\frac{n_1}{n}f - f\right\|_p + \left\|\frac{1}{n}\sum_{i=n_1+1}^{n} K_h(t-X_i)\right\|_p \\ &\le \|\xi_{h,n_1}(f,t)\|_p + \|B_h(f,t)\|_p + \frac{n-n_1}{n}\big(\|f\|_p + \|K_h\|_p\big)\end{aligned}$$

By Lemma C.1.1, for any kernel $K \in \mathcal{K}_\beta(L_K)$ with $\beta \ge \beta_0$ (and thus $K \in \mathcal{K}_{\beta_0}(2L_K)$), we know

$$\|B_h(f,t)\|_p \le L_0(2L_K)^d \sum_{j=1}^{d} h_j^{\beta_{0,j}} = L_0 2^d L_K^d d V_h^{\beta_0}, \quad \forall h \in \mathcal{H}.$$

Conditional on the event $\{\frac{n-n_1}{n} < 2\epsilon\}$, we have

$$\frac{n-n_1}{n}\big(\|f\|_p + \|K_h\|_p\big) \le 2\epsilon\left(L_0 + L_K^d V_h^{-1/q}\right).$$

Then for any $h, l \in H$, $h \neq l$, we have

$$\|\widehat{f}_h - \widehat{f}_l\|_p \le \|\widehat{f}_h - f\|_p + \|\widehat{f}_l - f\|_p \le \|\xi_{h,n_1}(f,t)\|_p + \|\xi_{l,n_1}(f,t)\|_p + C_0(V_l^{\beta_0} + \epsilon V_h^{1/q}),$$

where $C_0 = 8d(2^d(L_0 - 1)(L_K^d - 1)$. By the definition of $\widehat{h}$, we know

$$\bar{P}(\widehat{h} > h) \le \bar{P}\Big(\exists l \neq h, l \in H \text{ s.t. } \|\widehat{f}_h - \widehat{f}_l\|_p > c_0 V_l^{\beta_0}\Big).$$

Noting that $\epsilon V_h^{1/q} \le V_h^{\beta_0} \le V_l^{\beta_0}$ for any $l \le h$, we have

$$\bar{P}(\widehat{h} > h) \le \sum_{l<h, l\in H} \bar{P}\Big(\|\xi_{l,n_1}(f,t)\|_p + \|\xi_{h,n_1}(f,t)\|_p > \frac{c_0}{2} V_l^{\beta_0}\Big),$$

whenever $C_0 \le c_0/4$. Notice that $(nV_l)^{1/(q-2)} \le (nV_h)^{1/(q-2)} \le V_h^{\beta_0} \le V_l^{\beta_0}$ for any $l \le h$; hence, we have

$$\bar{P}(\widehat{h} > h) \le 2|H|\bar{P}_{\max}, \quad \bar{P}_{\max} := \max_{l<h, l\in H} \bar{P}\Big(\|\xi_{l,n_1}(f,t)\|_p > \frac{c_0}{4}(nV_l)^{1/(q-2)}\Big).$$

It is not straightforward to use Lemma C.3.1 to bound $\bar{P}_{\max}$ as it involves $\widehat{r}_p(K_l)$ for $p \in (2, 1)$, which, by its definition (4.3.4), relies on the whole data set $\{X_1, \ldots, X_n\}$ (including the contaminated data) and thus does not align with $\xi_{l,n_1}(f,t)$ (defined for only the clean data generated from P_f) in our setting. We use some iterative trick to get rid of $\widehat{r}_p(K_l)$ and show the final bound of $\bar{P}_{\max}$ in the following lemma. Its proof is left at the end.

Lemma C.5.1. *Under the conditions of Theorem 4.3.4, we have* $\bar{P}_{\max} \lesssim n^{-2/q}$.

Using this lemma, we get the following bound conditional on the event $\{\frac{n-n_1}{n} < 2\epsilon\}$:

$$\begin{aligned}
\bar{E}\|\widehat{f}_{\widehat{h}} - f\|_p &\le \bar{E}(\|\widehat{f}_{\widehat{h}} - f\|_p : \widehat{h} \le h) + \bar{E}(\|\widehat{f}_{\widehat{h}} - f\|_p : \widehat{h} > h) \\
&\overset{(i)}{\le} \bar{E}(\|\widehat{f}_{\widehat{h}} - \widehat{f}_h\|_p : \widehat{h} \le h) + \bar{E}(\|\widehat{f}_h - f\|_p : \widehat{h} \le h) + \\
&\quad \bar{E}\{(\|K_{\widehat{h}}\|_p + \|f\|_p)\mathbf{1}(\widehat{h} > h)\} \\
&\overset{(ii)}{\le} c_0 V_h^{\beta_0} + E\|\widehat{f}_h - f\|_p + (L_K^d n^{1/q} + L_0)\bar{P}(\widehat{h} > h) \\
&\lesssim V_h^{\beta_0} + \epsilon V_h^{1/q} + (nV_h)^{1/(q-2)} + n^{1/q}|H|\bar{P}_{\max} \\
&\overset{(iii)}{\lesssim} n^{-\frac{\beta_0}{(q-2)\beta_0+1}} \vee \epsilon^{\frac{q\beta_0}{q\beta_0+1}}.
\end{aligned}$$

In (i) and (ii), we have used the fact that $\|\widehat{f}_{\widehat{h}} - f\|_p \le \|K_{\widehat{h}} * f - f\|_p + \|\widehat{f}_{\widehat{h}} - K_{\widehat{h}} * f\|_p \le L_K^d V_{\widehat{h}}^{1/q} + L_0 L_K^d n^{1/q} + L_0$ for $\widehat{h} \in H$. (iii) is due to that $n^{1/q}|H|\bar{P}_{\max} \cdot \log n \cdot n^{1/q} \cdot n^{\frac{\beta_0}{(q-2)\beta_0+1}}$. Finally, we get

$$\begin{aligned}
E\|\widehat{f}_{\widehat{h}} - f\|_p &\le E\left(\|\widehat{f}_{\widehat{h}} - f\|_p : \frac{n - n_1}{n} < 2\epsilon\right) + E\left(\|\widehat{f}_{\widehat{h}} - f\|_p : \frac{n - n_1}{n} \ge 2\epsilon\right) \\
&\lesssim n^{\frac{\beta_0}{(q-2)\beta_0+1}} - \epsilon^{\frac{q\beta_0}{q\beta_0+1}} + (L_K^d n^{-1/q} + L_0) P\left(\frac{n - n_1}{n} \ge 2\epsilon\right) \\
&\lesssim n^{\frac{\beta_0}{(q-2)\beta_0+1}} - \epsilon^{\frac{q\beta_0}{q\beta_0+1}} + n^{1/q}\exp\left(-\frac{1}{4}n\epsilon\right) \quad \text{(by (C.4.5))} \\
&\lesssim n^{\frac{\beta_0}{(q-2)\beta_0+1}} - \epsilon^{\frac{q\beta_0}{q\beta_0+1}}
\end{aligned}$$

for any $\epsilon > \frac{8\log n}{qn}$. If $\epsilon \le \frac{8\log n}{qn}$, similar to what we did at the end of the proofs of Theorems 4.3.2 and 4.3.3, we can find some distribution $\widetilde{G}$ such that

$$(1-\epsilon)P_f + \epsilon G = (1 - \frac{8}{q}\frac{\log n}{n})P_f + \frac{8}{q}\frac{\log n}{n}\widetilde{G}.$$

(e.g. pick $\widetilde{G} := \frac{c(n)-\epsilon}{c(n)}P_f + \frac{\epsilon}{c(n)}G$, $c(n) = \frac{8\log n}{qn}$.) Thus we may equivalently treat the contamination proportion as $\frac{8}{q}\frac{\log n}{n}$ and still get the above result (similar to (C.4.11)). □

Proof of Lemma C.5.1. Recall the definition of $\bar{r}_s(U)$, $\bar{r}_s(U, p_X)$, $\bar{\widehat{r}}_s(U)$, $\bar{d}_s(U)$ in (C.4.12)-(C.4.15) for all $s \in [1, 1)$. In this proof, all the inequalities are conditional on $\{\frac{n-n_1}{n} < 2\epsilon\} \cap \{n/2 \le n_1 \le n\}$ (assuming that $\epsilon \le 1/4$) and we will use the condition $n/2 \le n_1 \le n$ from place to place.

(i) For any $p \in [1,2]$, any $l \in H$, we know $\bar{r}_p(K_l) \le 128 n_1^{-1/q}\|K_l\|_p \le 256 L_K^d (nV_l)^{-1/q}$, conditional on $\{\frac{n-n_1}{n} < 2\epsilon\} \cap \{n/2 \le n_1 \le n\}$. If $256 L_K^d < c_0/8$, then

$$\begin{aligned}
\bar{P}_{\max} &\le \max_{l=h,\, l\in H} \bar{P}\left(\|\xi_{l,n_1}(f,t)\|_p - \bar{r}_p(K_l) > \frac{c_0}{8}(nV_l)^{-1/q}\right) \\
&\le \max_{l=h,\, l\in H} \frac{8}{c_0}(nV_l)^{1/q}\bar{E}_f \sup_{l\in H}[\|\xi_{l,n_1}(f,t)\|_p - \bar{r}_p(K_l)]_+ \le n^{1/q}\bar{E}\delta_{n_1,p} \le n^{1/q}\delta_{n,p} \cdot n^{-2/q}.
\end{aligned}$$

(ii) For $p \in (2, 1)$, we will first assume $p \in (2,4]$ and illustrate our main idea through the proof for this simple case and then use the same idea to deal with the general case where $p \in (2^m, 2^{m+1}]$ for any $m \in \mathbb{N}$.

(a) $p \in (2,4]$. From the definition (C.4.13) of $\bar{r}_s(U, p_X)$, we know for any $s \in (2, \infty)$, function U and density p_X, the following inequality holds

$$\begin{aligned}\bar{r}_s(U, p_X) &\le 256 D_s \left\{ n^{-1/2} \|U\|^2 \|p_X\|_{s/2}^{1/2} - n^{1/s-1} \|U\|_s - n^{-1/2} \|U\|_2 \right\} \\ &\le 256 D_s \left\{ n^{-1/2} \|U\|^2 \|1\|^{1/2} \|p_X\|_{s/2}^{1/2} - n^{1/s-1} \|U\|_s - n^{-1/2} \|U\|_2 \right\} \\ &\le 256 D_s (\|p_X\|_{s/2}^{1/2} - 1) \left\{ n^{1/s-1} \|U\|_s - n^{-1/2} \|U\|_2 \right\}. \end{aligned} \tag{C.5.1}$$

Then for $\bar{r}_p(K_l, f)$, since $\|f\|_{p/2}^{1/2} \le \|f\|_p^{\frac{p-2}{2(p-1)}} \le L_0 - 1$, we know

$$\bar{r}_p(K_l, f) \le C_p (nV_l)^{-1/2},$$

where $C_p = 256 D_p L_K^d (L_0 - 1)$. If $C_p \le c_0/8$, then for any $l \in H$,

$$\begin{aligned}\bar{P}\left(\|\xi_{l,n_1}(f,t)\|_p > \frac{c_0}{4}(nV_l)^{-1/2} \right) &\le \bar{P}\left(\|\xi_{l,n_1}(f,t)\|_p \cdot \bar{\hat{r}}_p(K_l) > \frac{c_0}{16}(nV_l)^{-1/2} \right) \\ &\quad + \bar{P}\left(\bar{\hat{r}}_p(K_l) \le \bar{r}_p(K_l, f) > \frac{c_0}{16}(nV_l)^{-1/2} \right) \\ &:= \Gamma_{l,0} + \Gamma_{l,1}. \end{aligned} \tag{C.5.2}$$

By Lemma C.3.1 (iii), we know for any $p \in (2, \infty)$,

$$\Gamma_{l,0} \le \frac{16}{c_0}(nV_l)^{1/2} \bar{E}_f \sup_{l \in H} [\|\xi_{l,n_1}(f,t)\|_p \cdot \bar{\hat{r}}_p(K_l)]_+ \le n^{1/2} \bar{E} \delta_{n_1,p} \le n^{1/2} \delta_{n,p}. \tag{C.5.3}$$

For the second term $\Gamma_{l,1}$, by the definition of $\bar{\hat{r}}_p(K_l), \bar{r}_p(K_l, f)$, we have

$$\begin{aligned}\bar{\hat{r}}_p(K_l) \le \bar{r}_p(K_l, f) &\le 32 D_p n_1^{-1/2} \left| \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} K_l^2(t, X_i) \right\|_{p/2}^{1/2} - \left\| E_f K_l^2(t, X) \right\|_{p/2}^{1/2} \right| \\ &\quad 64 D_p n^{-1/2} \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[K_l^2(t, X_i) - E_f K_l^2(t, X) \right] \right\|_{p/2}^{1/2}. \end{aligned} \tag{C.5.4}$$

Define

$$\xi_{l,n_1}^{(1)}(f,t) := \frac{1}{n_1} \sum_{i=1}^{n_1} \left[K_l^2(t, X_i) - E_f K_l^2(t, X) \right].$$

Then for any $p \in (2, \infty)$, we have

$$\Gamma_{l,1} \le \bar{P}\left(\|\xi_{l,n_1}^{(1)}(f,t)\|_{p/2} > c_0^{(1)} V_l^{-1} \right), \tag{C.5.5}$$

where $c_0^{(1)} := \left(\frac{c_0}{2^{10}D_p}\right)^2$.

Notice that $\xi_{l,n_1}^{(1)}(f,t)$ represents the variance part for the kernel estimator $\frac{1}{n_1}\sum_{i=1}^{n_1} K_l^2(X_i \quad t)$ with the kernel K^2. (We possibly need to multiply K^2 by a normalization constant $C = 1/\int K^2$ to make it a real kernel. But one can easily check that this normalization constant won't affect any following result since we essentially only care about the asymptotic rates in the following bounds. Therefore, for simplicity, we still use K^2 to denote the kernel CK^2.) Also, when $p \in (2,4]$, $p/2 \in (1,2]$; hence, we can use Lemma C.3.1 (i) or (ii) directly (without involving $\bar{\hat{r}}_{p/2}(K_l^2)$) to get a bound related to $\| \xi_{l,n_1}^{(1)}(f,t) \|_{p/2}$. It is easy to check the kernel K^2 still satisfies the assumptions (K1) and (K2) in Lemma C.3.1 (with some new parameters L_K, k_1 though). Therefore, for any $l \in H$, we have

$$\bar{P}\left(\|\xi_{l,n_1}^{(1)}(f,t)\|_{p/2} > c_0^{(1)} V_l^{\ 1}\right) \quad \bar{P}\left(\|\xi_{l,n_1}^{(1)}(f,t)\|_{p/2} \quad \bar{r}_{p/2}(K_l^2) > \frac{c_0^{(1)}}{2} V_l^{\ 1}\right)$$
$$\frac{2V_l}{c_0^{(1)}} \bar{E} \sup_{l \in H} \left[\|\xi_{l,n_1}^{(1)}(f,t)\|_{p/2} \quad \bar{r}_{p/2}(K_l^2)\right]_+$$
$$. \ \bar{E}\delta_{n_1,p/2} \quad \delta_{n,p/2}.$$

Here we have used the fact that for $p \in (2,4]$,

$$\bar{r}_{p/2}(K_l^2) \quad 128 n_1^{2/p\ 1} \|K_l^2\|_{p/2} \quad 256 L_K^{2d} n^{2/p\ 1} V_l^{2/p\ 2} \quad \frac{c_0^{(1)}}{2} V_l^{\ 1},$$

whenever $256 L_K^{2d} \quad \frac{c_0^{(1)}}{2}$ and $nV_l \quad 1$. Therefore, we get

$$\bar{P}_{\max} \quad \max_{l \ h \ , l \in H} \{\Gamma_{l,0} + \Gamma_{l,1}\} \ . \ n^{1/2}\delta_{n,p} + \delta_{n,p/2} \ . \ n^{\ 2/q},$$

as $V_{\max} \quad c_1/(\log n)^{p/2}$ and we can make $\delta_{n,p} = o(n^{\ 3})$ by setting c_1 sufficiently small. For $p \in (2,4)$, $\delta_{n,p/2} \quad \exp\{ \ Cn^{4/p\ 1}\}$ decays exponentially with n. For $p = 4$, $V_{\max}$ $c_1/(\log n)^{p/2} \quad c_1/\log n$ and we can make $\delta_{n,p/2} = \delta_{n,2} = o(n^{\ 2/q})$ by setting c_1 sufficiently small.

(b) More generally, for any $p \in (2, 1)$, we may assume that $p \in (2^m, 2^{m+1}]$ for some integer $m \quad 1$. We will show

1 For any $l \in H$, any $p \in (2^m, 1)$, we have

$$\bar{P}\Big(\| \xi_{l,n_1}(f,t) \|_p > \frac{c_0}{4}(nV_l)^{-1/2} \Big) \cdot n^{1/2}\delta_{n,p} + \delta_{n,p/2} + \ldots + \delta_{n,p/2^{m-1}} + \bar{P}\Big(\| \xi^{(m)}_{l,n_1}(f,t) \|_{p/2^m} > c_0^{(m)} \frac{n^{2^{m-1}-1}}{V_l^{2^{m-1}-1}} \Big), \tag{C.5.6}$$

where

$$\xi^{(k)}_{l,n_1}(f,t) := \frac{1}{n_1}\sum_{i=1}^{n_1} [K_l^{2^k}(t \quad X_i) \quad E_f K_l^{2^k}(t \quad X)],$$

and $c_0^{(k+1)} := \left(\frac{c_0^{(k)}}{256 D_{p/2^k}} \right)^2$, for $k \quad 1$ and $c_0^{(1)} := \left(\frac{c_0}{2^{10} D_p} \right)^2$ as defined before.

2 For any $l \in H$, any $p \in (2^m, 2^{m+1}]$, we have

$$\bar{P}\Big(\| \xi^{(m)}_{l,n_1}(f,t) \|_{p/2^m} > c_0^{(m)} \frac{n^{2^{m-1}-1}}{V_l^{2^{m-1}-1}} \Big) \cdot \delta_{n,p/2^m}.$$

Note that with 1 and 2 , one can immediately obtain the following inequality with $V_{\max}$ $c_1/(\log n)^{p/2}$ for a sufficiently small c_1,

$$\bar{P}_{\max} \cdot n^{1/2}\delta_{n,p} + \delta_{n,p/2} + \ldots + \delta_{n,p/2^{m-1}} + \delta_{n,p/2^m} \cdot n^{-2/q}.$$

Thus, it suffices to show 1 and 2 .

Proof of 1 : For $m = 1$, by (C.5.2), (C.5.3) and (C.5.5) in (a), we actually have shown that for any $p \in (2, 1)$,

$$\bar{P}\Big(\| \xi_{l,n_1}(f,t) \|_p > \frac{c_0}{4}(nV_l)^{-1/2} \Big) \quad \Gamma_{l,0} + \Gamma_{l,1} \cdot n^{1/2}\delta_{n,p} + \bar{P}\Big(\| \xi^{(1)}_{l,n_1}(f,t) \|_{p/2} > c_0^{(1)} V_l^{-1} \Big).$$

Assume that (C.5.6) holds for any $p \in (2^k, 1)$ (the case $m = k$ and $k \quad 1$). Then if we can show for any $p \in (2^{k+1}, 1)$ (the case $m = k+1$),

$$\bar{P}\Big(\| \xi^{(k)}_{l,n_1}(f,t) \|_{p/2^k} > c_0^{(k)} \frac{n^{2^{k-1}-1}}{V_l^{2^{k-1}-1}} \Big) \cdot \delta_{n,p/2^k} + \bar{P}\Big(\| \xi^{(k+1)}_{l,n_1}(f,t) \|_{p/2^{k+1}} > c_0^{(k+1)} \frac{n^{2^k-1}}{V_l^{2^k}} \Big), \tag{C.5.7}$$

then by induction on m, (C.5.6) holds for any $p \in (2^m, 1]$ and any integer $m \geq 1$. Thus, it suffices to show (C.5.7) holds for any $p \in (2^{k+1}, 1]$. Noticing that $p/2^k \in (2, 1]$, by (C.5.1), we have

$$\begin{aligned}
\bar{r}_{p/2^k}\left(K_l^{2^k}, f\right) &\leq 256 D_{p/2^k}(\|f\|_{p/2^{k+1}}^{1/2} \vee 1)\left\{n^{\frac{2^k}{p}-1}\left\|K_l^{2^k}\right\|_{p/2^k} \vee n^{-1/2}\left\|K_l^{2^k}\right\|_2\right\} \\
&\leq 256 D_{p/2^k}(L_0 \vee 1)\left\{(nV_l)^{\frac{2^k}{p}-1} V_l^{1-2^k}\|K\|_p^{2^k} \vee (nV_l)^{-1/2} V_l^{1-2^k}\|K\|_{2^{k+1}}^{2^k}\right\} \\
&\leq 256 D_{p/2^k}(L_0 \vee 1)(L_K \vee 1)^{dp}(nV_l)^{-1/2} V_l^{1-2^k}.
\end{aligned}$$

Here, we have used the fact $\|f\|_{s/2} \leq \|f\|_s^{\frac{s-2}{s-1}} \leq \|f\|_s \vee 1$ for any $s > 2$ and any density f with $\|f\|_s < \infty$. With $256 D_{p/2^k}(L_0 \vee 1)(L_K \vee 1)^{dp} \leq c_0^{(k)}/2$ and $nV_l \geq 1$, we have

$$\begin{aligned}
\bar{P}\left(\left\|\xi_{l,n_1}^{(k)}(f,t)\right\|_{p/2^k} > c_0^{(k)} \frac{n^{2^k-1}}{V_l^{2^k-1}}\right) &\leq \bar{P}\left(\|\xi_{l,n_1}^{(k)}(f,t)\|_{p/2^k} \geq \bar{\hat{r}}_{p/2^k}\left(K_l^{2^k}\right) > \frac{c_0^{(k)}}{4}\frac{n^{2^k-1}}{V_l^{2^k-1}}\right) \\
&\quad + \bar{P}\left(\bar{\hat{r}}_{p/2^k}\left(K_l^{2^k}\right) \geq \bar{r}_{p/2^k}\left(K_l^{2^k}, f\right) > \frac{c_0^{(k)}}{4}\frac{n^{2^k-1}}{V_l^{2^k-1}}\right) \\
&:= \Gamma_{l,k} + \Gamma_{l,k+1}.
\end{aligned}$$

By Lemma C.3.1 (iii), we know for any $p \in (2^{k+1}, 1]$,

$$\Gamma_{l,k} \leq \frac{4}{c_0^{(k)}}\frac{V_l^{2^k-1}}{n^{2^k-1-1}}\bar{E}_f \sup_{l \in H}\left[\|\xi_{l,n_1}^{(k)}(f,t)\|_{p/2^k} - \bar{\hat{r}}_{p/2^k}\left(K_l^{2^k}\right)\right]_+ \cdot \bar{E}\delta_{n_1,p/2^k} \leq \delta_{n,p/2^k}. \quad \text{(C.5.8)}$$

For $\Gamma_{l,k+1}$, similar to (C.5.4), we have

$$\begin{aligned}
&\bar{\hat{r}}_{p/2^k}\left(K_l^{2^k}\right) - \bar{r}_{p/2^k}\left(K_l^{2^k}, f\right) \\
&\leq 64 D_{p/2^k} n^{-1/2}\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\left[K_l^{2^{k+1}}(t - X_i) - E_f K_l^{2^{k+1}}(t - X)\right]\right\|_{p/2^{k+1}}^{1/2} \\
&= 64 D_{p/2^k} n^{-1/2}\left\|\xi_{l,n_1}^{(k+1)}(f,t)\right\|_{p/2^{k+1}}^{1/2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\Gamma_{l,k+1} &\leq \bar{P}\left(64 D_{p/2^k} n^{-1/2}\left\|\xi_{l,n_1}^{(k+1)}(f,t)\right\|_{p/2^{k+1}}^{1/2} > \frac{c_0^{(k)}}{4}\frac{n^{2^k-1}}{V_l^{2^k-1}}\right) \\
&\quad \bar{P}\left(\left\|\xi_{l,n_1}^{(k+1)}(f,t)\right\|_{p/2^{k+1}} > c_0^{(k+1)}\frac{n^{2^k-1}}{V_l^{2^k}}\right),
\end{aligned}$$

where $c_0^{(k+1)} := \left(\frac{c_0^{(k)}}{256 D_{p/2^k}}\right)^2$.

Proof of 2 : For $p \in (2^m, 2^{m+1}]$, we know $p/2^m \in (1,2]$, and thus we may introduce $\bar{r}_{p/2^m}\left(K_l^{2^m}\right)$ and then use the results in Lemma C.3.1 (i) or (ii). Notice that

$$\bar{r}_{p/2^m}\left(K_l^{2^m}\right) \quad 128 n_1^{\frac{2^m}{p} \quad 1} \left\| K_l^{2^m} \right\|_{p/2^m} \quad 256 n^{\frac{2^m}{p} \quad 1} k K_l k_p^{2^m}$$
$$256 (L_K \_ 1)^{dp} (n V_l)^{\frac{2^m}{p} \quad 1} V_l^{1 \quad 2^m} \quad 256 (L_K \_ 1)^{dp} V_l^{1 \quad 2^m}.$$

Therefore, with $256(L_K \_ 1)^{dp} \quad c_0^{(m)}/2$, by Lemma C.3.1 (i) or (ii), we have

$$\bar{P}\left(\left\| \xi_{l,n_1}^{(m)}(f,t) \right\|_{p/2^m} > c_0^{(m)} \frac{n^{2^m \quad 1} \quad 1}{V_l^{2^m \quad 1}} \right)$$
$$\bar{P}\left(\left\| \xi_{l,n_1}^{(m)}(f,t) \right\|_{p/2^m} \quad \bar{r}_{p/2^m}\left(K_l^{2^m}\right) > \frac{c_0^{(m)}}{2} \frac{n^{2^m \quad 1} \quad 1}{V_l^{2^m \quad 1}} \right)$$
$$\frac{2}{c_0^{(m)}} \frac{V_l^{2^m \quad 1}}{n^{2^m \quad 1} \quad 1} \bar{E}_f \sup_{l 2 H} \left[\left\| \xi_{l,n_1}^{(m)}(f,t) \right\|_{p/2^m} \quad \bar{r}_{p/2^m}\left(K_l^{2^m}\right) \right]_+$$
$$. \ \bar{E} \delta_{n_1, p/2^m} \quad \delta_{n, p/2^m},$$

which completes the proof of the lemma. □

C.6 Proof of Theorem 4.3.5

Proof of Theorem 4.3.5. We prove it by contradiction. Assume that for some $j \in \{1, \ldots, d\}$ and some positive functions $R_1(\ )$ and $R_2(\ )$, there exists an estimator $\widehat{f}$ that is $(R_1(\ ), R_2(\ ))$-rate adaptive with respect to $\{\beta_{0,j}, \epsilon\}$. For simplicity, we assume $j = 1$. From the Definition 4.3.1, we know that there exist three constants C_0, C_1, C_2 such that for any $\beta_{0,1} \quad C_1$, $\epsilon \quad C_2$, $n \quad 1$, we have

$$\sup_{f \in P_{p,d}^G(\beta_0, L_0)} E_{P(\epsilon,f,G)} \| \widehat{f} \quad f \|_p \quad C_0 \left(n^{\quad R_1(\beta_0)} \_ \epsilon^{R_2(\beta_0)} \right). \tag{C.6.1}$$

For any $\beta_{0,1}, \widetilde{\beta}_{0,1} \quad C_1$, we pick a function $f_0 \in P_p(\beta_0, L_0/2) \setminus P_p(\widetilde{\beta}_0, L_0/2)$, where $\widetilde{\beta}_0 = (\widetilde{\beta}_{0,1}, \beta_{0,2}, \ldots, \beta_{0,d})'$. For example, we can choose $f_0(x) = \gamma_0^d \prod_{j=1}^d \phi_0(\gamma_0 x_j)$ where ϕ_0

is some infinitely differentiable density function on R with a compact support and γ_0 is a sufficiently small number.

Let

$$g_0(x) = \frac{1}{\epsilon}\gamma_0^d V_h^{\widetilde{\beta}_0} \ {}^{\frac{1}{p}} \prod_{j=1}^{d} \phi_0\Big(\frac{\gamma_0 x_j}{h_j}\Big),$$

where $h_j = \epsilon^{\frac{q\widetilde{\beta}_0}{\widetilde{\beta}_{0,j}(q\widetilde{\beta}_0+1)}}$ and $V_h = \prod_{j=1}^{d} h_j = \epsilon^{\frac{q}{q\widetilde{\beta}_0+1}}$.

It is easy to check that g_0 is a density function on R^d and ϵg_0 2 $\mathsf{N}_{p,d}(\widetilde{\beta}_0, L_0/2)$ (similar to part (a) or (b) in the proof of Theorem 4.2.4) if γ_0 is sufficiently small. Let

$$\widetilde{f}_0 = (1 \quad \epsilon) f_0 + \epsilon g_0.$$

Then $\widetilde{f}_0$ is a density function in $\mathsf{P}_p(\widetilde{\beta}_0, L_0)$. By our construction, we actually get

$$P(\epsilon, f_0, G_0) = P(0, \widetilde{f}_0, \widetilde{G}_0),$$

where $G_0 = P_{g_0}$ and $\widetilde{G}_0$ is an arbitrary distribution on R^d. Notice that f_0 2 $\mathsf{P}_p(\beta_0, L_0)$ and $\widetilde{f}_0$ 2 $\mathsf{P}_p(\widetilde{\beta}_0, L_0)$. Consequently, we have

$$\sup_{\substack{f 2 \mathsf{P}_{p,d}(\beta_0, L_0) \\ G}} E_{P(\epsilon,f,G)} \mathsf{k}\, \widehat{f} \quad f\, \mathsf{k}_p + \sup_{\substack{f 2 \mathsf{P}_{p,d}(\widetilde{\beta}_0, L_0) \\ G}} E_{P(0,f,G)} \mathsf{k}\, \widehat{f} \quad f\, \mathsf{k}_p$$

$$E_{P(\epsilon,f_0,G_0)} \mathsf{k}\, \widehat{f} \quad f_0\, \mathsf{k}_p + E_{P(0,\widetilde{f}_0,\widetilde{G}_0)} \mathsf{k}\, \widehat{f} \quad \widetilde{f}_0\, \mathsf{k}_p$$

$$\mathsf{k} f_0 \quad \widetilde{f}_0 \mathsf{k}_p \quad \mathsf{k}\epsilon g_0 \mathsf{k}_p \quad \mathsf{k}\epsilon f_0 \mathsf{k}_p = \gamma_0^{d/q} \mathsf{k}\phi_0 \mathsf{k}_p^d (\epsilon^{\frac{q\widetilde{\beta}_0}{q\widetilde{\beta}_0+1}} \quad \epsilon) \quad c_0 \epsilon^{\frac{q\widetilde{\beta}_0}{q\widetilde{\beta}_0+1}}$$

for all ϵ small enough (say all ϵ C_3) and $c_0 = \frac{1}{2}\gamma_0^{d/q} \mathsf{k}\phi_0 \mathsf{k}_p^d$ is an independent constant. Then by (C.6.1), we must have

$$c_0 \epsilon^{\frac{q\widetilde{\beta}_0}{q\widetilde{\beta}_0+1}} \quad 2C_0 \left(n^{\ R_1(\beta_0)} \_ \epsilon^{R_2(\beta_0)} \_ n^{\ R_1(\widetilde{\beta}_0)} \right)$$

for any $\beta_{0,1}, \widetilde{\beta}_{0,1}$ C_1, n 1, ϵ $C_2 \wedge C_3$. But this is impossible, as we notice that

$$\frac{q\bar{\widetilde{\beta}}_0}{q\bar{\widetilde{\beta}}_0 + 1} = \frac{q}{q + \sum_j {}_2\, 1/\beta_{0,j} + 1/\widetilde{\beta}_{0,1}}$$

and thus, for some given β_0, we can always choose a small enough $\widetilde{\beta}_{0,1}$ and a small enough ϵ such that $c_0 \epsilon^{\frac{q\widetilde{\beta}_0}{q\widetilde{\beta}_0+1}} > 2C_0 \epsilon^{R_2(\beta_0)}$. Then for some large enough n, we also have $c_0 \epsilon^{\frac{q\widetilde{\beta}_0}{q\widetilde{\beta}_0+1}} > 2C_0 n^{\ R_1(\beta_0)} \_ n^{\ R_1(\widetilde{\beta}_0)}$ for some fixed β_0,$\widetilde{\beta}_0$ and ϵ. □

Bibliography

Radoslaw Adamczak, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, 2011.

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. *arXiv preprint arXiv:2007.01394*, 2020.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3):1352 – 1382, 2018.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Lawrence D Brown and Mark G Low. A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics*, 24(6):2524–2535, 1996.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.

Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for huber's ϵ-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.

Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

Xi Chen and Wen-Xin Zhou. Robust inference via multiplier bootstrap. *The Annals of Statistics*, 48(3):1665 – 1691, 2020.

Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, pages 1655–1684, 2014.

Adam Cheng, David Kessler, Ralph Mackinnon, Todd P Chang, Vinay M Nadkarni, Elizabeth A Hunt, Jordan Duval-Arnould, Yiqun Lin, Martin Pusic, and Marc Auerbach. Conducting multicenter research in healthcare simulation: Lessons learned from the inspire network. *Advances in Simulation*, 2(1):1–14, 2017.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM, 2019.

James C Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, Jeffrey John Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, et al. Spanner: Google's globally distributed database. *ACM Transactions on Computer Systems (TOCS)*, 31(3):1–22, 2013.

Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. In *Advances in Neural Information Processing Systems*, pages 13188–13198, 2019.

Luc Devroye. Nonparametric density estimation. *The L_1 View*, 1985.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 655–664. IEEE, 2016.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

David L Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22 (1):238–270, 1994.

David L Donoho and Richard C Liu. Geometrizing rates of convergence, iii. *The Annals of Statistics*, 19(2):668–701, 1991.

Simon S Du, Yining Wang, Sivaraman Balakrishnan, Pradeep Ravikumar, and Aarti Singh. Robust nonparametric regression under huber's ϵ-contamination model. *arXiv preprint arXiv:1805.10406*, 2018.

John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46 (2):814 – 841, 2018.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

Samuel H Fuller and Lynette I Millett. *The future of computing performance: game over or next level?* National Academy Press, 2011.

Chao Gao. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170, 2020.

Alexander Goldenshluger and Oleg Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008.

Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011a.

Alexander Goldenshluger and Oleg Lepski. Uniform bounds for norms of sums of independent random functions. *The Annals of Probability*, 39(6):2318–2384, 2011b.

Alexander Goldenshluger and Oleg Lepski. On adaptive minimax density estimation on R^d. *Probability Theory and Related Fields*, 159(3-4):479–543, 2014.

AV Goldenshluger and OV Lepski. General selection rule from a family of linear estimators. *Theory of Probability & Its Applications*, 57(2):209–226, 2013.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on in uence functions*, volume 196. John Wiley & Sons, 2011.

Frank Rudolf Hampel. *Contributions to the theory of robust estimation*. University of California, Berkeley, 1968.

Rafael Hasminskii and Ildar Ibragimov. On Density Estimation in the View of Kolmogorov's Ideas in Approximation Theory. *The Annals of Statistics*, 18(3):999 – 1010, 1990.

Cheng Huang and Xiaoming Huo. A distributed one-step estimator. *Mathematical Programming*, 174(1):41–76, 2019.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

IA Ibragimov and RZ Khasminski. More on estimation of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.(LOMI)*, 108(194):198, 1981.

Il'dar Abdullovich Ibragimov and Rafail Zalmanovich Khas' minskii. On estimate of the density function. *Zapiski Nauchnykh Seminarov POMI*, 98:61–85, 1980.

W. B. Johnson. Best Constants in Moment Inequalities for Linear Combinations of Independent and Exchangeable Random Variables. *The Annals of Probability*, 13(1):234 – 253, 1985.

Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Ravi Kannan, Santosh Vempala, and David Woodruff. Principal component analysis and higher correlations for distributed data. In *Conference on Learning Theory*, pages 1040–1057. PMLR, 2014.

Gérard Kerkyacharian, Oleg Lepski, and Dominique Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probability Theory and Related Fields*, 121(2):137–170, 2001.

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430, 2018.

Pravesh K Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *arXiv preprint arXiv:1711.11581*, 2017.

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, 2017.

Oleg V Lepski and Vladimir G Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 1997.

OV Lepski and A Goldenshluger. Structural adaptation via lp-norm oracle inequalities. *Probability Theory and Related Fields*, 126(1-2):47–71, 2009.

OV Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

OV Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

OV Lepskii. Asymptotically minimax adaptive estimation. ii. schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37(3):433–448, 1993.

Haoyang Liu and Chao Gao. Density estimation with contamination: minimax rates and theory of adaptation. *Electronic Journal of Statistics*, 13(2):3613–3653, 2019.

Lester Mackey, Ameet Talwalkar, and Michael I. Jordan. Distributed matrix completion and robust factorization. *Journal of Machine Learning Research*, 16(28):913–960, 2015.

D. M. Masaon. Risk bounds for kernel density estimators. *Journal of Mathematical Sciences*, 163(3):238, 2009.

Sergei Mikhailovich Nikol'skii. *Approximation of functions of several variables and imbedding theorems*, volume 205. Springer Science & Business Media, 2012.

Xiaoou Pan and Wen-Xin Zhou. Multiplier bootstrap for quantile regression: non-asymptotic theory under random design. *Information and Inference: A Journal of the IMA*, 10(3): 813–861, 2021.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, Pradeep Ravikumar, et al. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B*, 82(3):601–627, 2020.

Elvezio M Ronchetti and Peter J Huber. *Robust statistics*. John Wiley & Sons, 2009.

Jonathan D Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John wiley & sons, 2005.

Peter J Rousseeuw and Katrien Van Driessen. Computing lts regression for large data sets. *Data mining and knowledge discovery*, 12(1):29–45, 2006.

Takeyuki Sasai and Hironori Fujisawa. Robust estimation with lasso when outputs are adversarially contaminated. *arXiv preprint arXiv:2004.05990*, 2020.

Srijan Sengupta, Stanislav Volgushev, and Xiaofeng Shao. A subsampled double bootstrap for massive data. *Journal of the American Statistical Association*, 111(515):1222–1232, 2016.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008, 2014.

Chengchun Shi, Wenbin Lu, and Rui Song. A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association*, 113(524):1698–1709, 2018.

Ellen Sidransky, Michael A Nalls, Jan O Aasly, Judith Aharon-Peretz, Grazia Annesi, Egberto R Barbosa, Anat Bar-Shira, Daniela Berg, Jose Bras, Alexis Brice, et al. Multicenter analysis of glucocerebrosidase mutations in parkinson's disease. *New England Journal of Medicine*, 361(17):1651–1661, 2009.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

Vladimir Spokoiny. Bernstein-von mises theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430*, 2013.

Vladimir Spokoiny, Mayya Zhilova, et al. Bootstrap confidence sets under model misspecification. *The Annals of Statistics*, 43(6):2653–2675, 2015.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

Il'ya Sergeevich Tyurin. On the convergence rate in lyapunov's theorem. *Theory of Probability & Its Applications*, 55(2):253–270, 2011.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Bengt von Bahr, Carl-Gustav Esseen, et al. Inequalities for the rth absolute moment of a sum of random variables, 1 5 r 5 2. *The Annals of Mathematical Statistics*, 36(1): 299–303, 1965.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *International Conference on Machine Learning*, pages 3636–3645. PMLR, 2017.

Xiangyu Wang and David B Dunson. Parallelizing mcmc via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Yang Yu, Shih-Kang Chao, and Guang Cheng. Simultaneous inference for massive data: distributed bootstrap. In *International Conference on Machine Learning*, pages 10892–10901. PMLR, 2020.

Yuchen Zhang and Lin Xiao. Communication-efficient distributed optimization of self-concordant empirical loss. In *Large-Scale and Distributed Optimization*, pages 289–341. Springer, 2018.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *The Annals of statistics*, 44(4):1400, 2016.

Wen-Xin Zhou, Koushiki Bose, Jianqing Fan, and Han Liu. A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *The Annals of statistics*, 46(5):1904, 2018.