UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays on robust statistical estimation and inference

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Mathematics

by

Myeonghun Yu

Committee in charge:

Professor Ery Arias-Castro, Chair
Professor Wen-Xin Zhou, Co-Chair
Professor Ronghui Xu
Professor Ying Zhu
Professor Danna Zhang

2024

The Dissertation of Myeonghun Yu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego
2024

DEDICATION

To my family.

EPIGRAPH

Man only likes to count his troubles;
he doesn't calculate his happiness.

Fyodor Dostoevsky

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor, Wenxin Zhou, for his invaluable guidance and support throughout my academic journey. His mentorship helped me navigate challenges and strengthened my dedication to academia. He continuously inspires me to grow as a scholar, and I eagerly anticipate expanding the field of statistics with him in the future.

Besides, I would like to express my sincere gratitude to my committee, including Professor Danna Zhang, Professor Ery Arias-Castro, Professor Ronghui Xu and Professor Ying Zhu. I really appreciate their participation as committee members, and their insightful comments and suggestions during my advancement talk and other discussions helped me to further improve the dissertation.

I am honored to have worked with several talented scholars. I owe my gratitude to Professor Kean Ming Tan, with whom I completed two papers. His warm care and encouragement during my visit to the University of Michigan were invaluable. I also thank the faculty, graduate students, and staff at UCSD and the University of Michigan for their help and suggestions.

My PhD is a direct result of the love and support from my parents and sisters. I am deeply thankful to my wife, Boram, for her unwavering support throughout these years. With our invaluable son, Ryan, we have experienced so much beyond this thesis, making this journey together a true blessing.

Chapter 1, in full, is a reprint of the material in the paper "Low-rank matrix recovery under heavy-tailed errors", Yu, Myeonghun, Sun, Qiang and Zhou, Wen-Xin. The paper has been published on *Bernoulli*, **30**, 2326–2345. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material in the paper "Gaussian differentially private robust mean estimation and inference", Yu, Myeonghun, Ren, Zhao and Zhou, Wen-Xin. The paper has been accepted by *Bernoulli*, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the

material. Yu, Myeonghun; Tan, Kean Ming; Huixia Judy Wang; Zhou, Wen-Xin. The dissertation author was the primary investigator and author of this material.

Chapter 4, in full, is currently being prepared for submission for publication of the material. Yu, Myeonghun; Wang, Yue; Xie, Siyu; Tan, Kean Ming; Zhou, Wen-Xin. The dissertation author was the primary investigator and author of this material.

VITA

2020	B.S. in Mathematical Sciences, Seoul National University
2020–2024	Graduate Teaching and Research Assistant, University of California San Diego
2024	Ph. D. in Mathematics, University of California San Diego

PUBLICATIONS

1. **M. Yu**, Z. Ren and W.-X. Zhou. (2023). Gaussian differentially private robust mean estimation and inference. *Bernoulli*, accepted.
2. **M. Yu**, K. M. Tan, H. J. Wang and W.-X. Zhou.(2023). Deep neural network expected shortfall regression with heavy-tailed data. *Annals of Statistics*, under review.
3. **M. Yu**, Q. Sun and W.-X. Zhou. (2024). Low-rank matrix recovery under heavy-tailed errors. *Bernoulli*, **30** 2326–2345.
4. **M. Yu**, Y. Wang, S. Xie, K. M. Tan and W.-X. Zhou.(2024). Estimation and inference for nonparametric expected shortfall regression over RKHS. *Journal of the American Statistical Association*, under review.

ABSTRACT OF THE DISSERTATION

Essays on robust statistical estimation and inference

by

Myeonghun Yu

Doctor of Philosophy in Mathematics

University of California San Diego, 2024

Professor Ery Arias-Castro, Chair
Professor Wen-Xin Zhou, Co-Chair

Due to the advancements of modern technologies, large-scale and high-dimensional data have been widely collected in almost every scientific disciplines. This introduces a several challenges including that the data are often accompanied by outliers due to possible measurement error, or many variables follow heavy-tailed distributions. To address these challenges, my thesis proposes methodologies in the setting of the mean estimation and matrix recovery when the data have asymmetric and heavy-tailed distributions. Additionally, I explore the characterization of tail behavior in random outcomes, focusing on expected shortfall, which is widely recognized as a measure of risk. I propose nonparametric approaches for estimating expected shortfall, aiming

to enhance its accuracy and applicability.

In Chapter 1, we propose a robust estimator to recover approximately low-rank matrices in the presence of heavy-tailed and asymmetric noises. Focusing on three archetypal applications including matrix compressed sensing, matrix completion and multitask learning, we provide sub-Gaussian-type deviation bounds when the noise variables only have bounded variances. Computationally, we propose a matrix version of the local adaptive majorize-minimization algorithm, which is much faster than the alternating direction method of multiplier used in previous work and is scalable to large datasets.

Chapter 2 studies the problem of robust and differentially private mean estimation and inference. We first provide a comprehensive analysis of the Huber mean estimator with increasing dimensions, including non-asymptotic deviation bound, Bahadur representation, and (uniform) Gaussian approximations. Then, we privatize the Huber mean estimator via noisy gradient descent, and construct private confidence intervals for the proposed estimator by incorporating a private and robust covariance estimator.

In Chapter 3, we consider the problem of nonparametric estimation of conditional expected shortfall functions. To mitigate the curse of dimensionality, we propose a two-step nonparametric ES estimator based on fully connected neural nets with the ReLU activation function. This approach (i) involves unobservable surrogate response variables that must be estimated from data in a preliminary step, and (ii) uses a properly chosen Huber loss to achieve exponential deviation bounds under heavy-tailed response distributions. Using a plugged-in nonparametric conditional quantile estimate, also trained on deep neural nets, we establish non-asymptotic high probability bounds for the final robust ES estimator, which are optimal as if the true quantile function were known without resorting to any type of sample splitting. We demonstrate the effectiveness of deep robust ES regression with both numerical experiments and an empirical study on the impact of El Niño on heavy precipitations, for which effective tail learning is imperative.

In Chapter 4, I introduce a two-step nonparametric ES estimator that involves a plugged-

in quantile function estimate without sample-splitting. We provide non-asymptotic estimation and Gaussian approximation error bounds, depending explicitly on the effective dimension, sample size, regularization parameters, and quantile estimation error. To construct pointwise confidence bands, we propose a fast multiplier bootstrap procedure and establish its validity. We demonstrate the finite-sample performance of the proposed methods through numerical experiments and an empirical study aimed at examining the heterogeneous effects of features on average and large medical expenses.

Chapter 1

Low-rank Matrix Recovery under Heavy-tailed Errors

1.1 Introduction

There has been a recent surge of interest in matrix recovery which aims to recover an unknown matrix from noisy observations. Matrix recovery has wide applications in practice, including collaborative filtering (Goldberg *et al.*, 1992), multitask regression (Argyriou, Evgeniou and Pontil, 2008), quantum state tomography (Gross *et al.*, 2010) and face recognition (Luan et al., 2014), to name a few. Statistically, it aims to estimate $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ based on n independently and identically distributed (i.i.d.) observations $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^n$ following the generative model

$$y_i = \mathrm{tr}(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\Theta}^*) + \varepsilon_i =: \langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle + \varepsilon_i,$$

where $\boldsymbol{X}_i \in \mathbb{R}^{d_1 \times d_2}$ is a random measurement matrix, and ε_i is an error variable satisfying $\mathbb{E}(\varepsilon_i | \boldsymbol{X}_i) = 0$ and $\mathbb{E}(\varepsilon_i^2 | \boldsymbol{X}_i) \leq \sigma_0^2$ for some $\sigma_0 > 0$. We consider matrix recovery in high dimensions, that is, $d_1 \times d_2$ can be much larger than the sample size n, making the problem ill-posed. It has been a common practice to assume that $\boldsymbol{\Theta}^*$ is (approximately) low-rank, and the resulting problem is referred to as low-rank matrix recovery.

The problem of low-rank matrix recovery can be naturally formulated as a nonconvex empirical risk minimization problem subject to a rank constraint. To find local optima of such a

rank-constrained program, commonly used methods are Riemannian gradient descent (Wei et al., 2016) and Burer-Monteiro type gradient descent (Burer and Monteiro, 2003; Chen, Liu and Li, 2020; Ma et al., 2020; Tong, Ma and Chi, 2021). The former views the set of rank-r matrices as a smooth manifold, while the latter relies on the matrix factorization $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U} \in \mathbb{R}^{d_1 \times r}$, $\boldsymbol{V} \in \mathbb{R}^{d_2 \times r}$, and $r = \mathrm{rank}(\boldsymbol{\Theta}^*)$ is assumed to be known. To relax the restrictive assumption that the true rank r is known *a priori*, Li, Ma and Zhang (2018) and Zhang, Fattahi and Zhang (2021) studied the gradient method for solving the reparameterized program in an over-parameterized regime where $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ with $\boldsymbol{U} \in \mathbb{R}^{d_1 \times r'}$ and $\boldsymbol{V} \in \mathbb{R}^{d_2 \times r'}$, and $r' \geq r$ is an upper bound of the true rank.

Another line of research resorts to convex relaxation in order to obtain computationally feasible solutions. Similar to Lasso (Tibshirani, 1996) in the context of sparse linear regression, convex low-rank matrix recovery methods are based on either constrained nuclear norm minimization or nuclear norm penalized least squares formulation. The nuclear norm of a matrix is defined as the sum of its singular values, and thus serves as a convex surrogate for its rank. We refer to Candès and Plan (2009), Candès and Recht (2009), Recht, Fazel and Parrilo (2010), Candès and Plan (2011), Rohde and Tsybakov (2011), Negahban and Wainwright (2011, 2012) and Klopp (2014) for an unavoidably incomplete list of notable works on exact and noisy low-rank matrix recovery through convex relaxation. In the context of multitask learning, Lounici et al. (2011) introduced an approach that utilizes the group Lasso penalty when only a small number of rows in the matrix $\boldsymbol{\Theta}^*$ are nonzero.

All the aforementioned methods, convex or nonconvex, are studied either in the noiseless setting or under a sub-Gaussian/sub-exponential assumption on the random error. However, both convex and nonconvex least squares estimators exhibit sub-optimal deviation bounds in the presence of heavy-tailed errors that only have a small number of finite moments. To make the estimator less sensitive to heavy-tailedness, a natural idea is to replace the ℓ_2-loss with a robust loss function, such as the ℓ_1-loss or the Huber loss (Huber, 1973). For example, Elsener and van de Geer (2018) proposed and studied nuclear norm penalized estimators using both

the ℓ_1-loss and the Huber loss; Tan, Sun and Witten (2023) considered robust sparse reduced rank regression by minimizing the empirical Huber loss plus a combination of the nuclear norm and entry-wise ℓ_1-norm penalties. For methods that rely on nonconvex optimization with robust losses, Shen et al. (2022) proposed a Riemannian sub-gradient method and proved the statistical properties of the iterates; Wang and Fan (2022) studied the statistical properties of vanilla gradient descent iterates for solving reparameterized (regularized) Huber loss minimization. In an over-parameterized regime, Ma and Fattahi (2023) showed that sub-gradient descent with the ℓ_1-loss function converges to the ground truth at a near-linear rate in the presence of arbitrarily large outliers.

In this paper, we propose a robust approach to recover an approximately low-rank matrix in a trace regression model with heavy-tailed and asymmetric error, which complements the extant literature on low-rank matrix recovery via convex relaxation. Borrowing ideas from robust (sparse) linear regressions (Fan, Li and Wang, 2017; Sun, Zhou and Fan, 2020), we adopt the Huber loss function with a diverging robustification parameter to achieve sub-Gaussian-type concentration bounds. We focus on three archetypal examples in matrix recovery: matrix compressed sensing, matrix completion and multitask regression. For each problem, we study the nonasymptotic deviation bounds of the nuclear norm penalized Huber estimator under both the Frobenius and nuclear norms, which match the minimax optimal rates. Our main contributions are as follows. First and foremost, we provide a comprehensive analysis of the nuclear norm penalized Huber regression estimator to gain robustness without compromising statistical efficiency. Our results either improve or complement those in Elsener and van de Geer (2018), Fan, Wang and Zhu (2021) and Tan, Sun and Witten (2023). For example, Elsener and van de Geer (2018) considered robust matrix completion under symmetric error distribution and also required a constant lower bound for the error density function. Tan, Sun and Witten (2023) examined the Huber-type estimator for sparse multitask regression but their analysis cannot be directly extended to the non-sparse setting. Secondly, we provide a unified algorithmic framework, which is a matrix variant of the local adaptive majorize-minimization (LAMM)

algorithm (Fan *et al.*, 2018), to solve the three problems (matrix sensing, matrix completion and multitask regression) all at once. By constructing an isotropic quadratic function that locally majorizes the empirical Huber loss, the solution to each proximal optimization problem has a closed form, which considerably facilitates the implementation. Compared to many other algorithms used in the literature, our algorithm is first-order and thus more scalable to large data sets.

1.1.1 Related work and paper organization

Our model setting is closely related to that in Fan, Wang and Zhu (2021), but the proposed robust estimators provably achieve sharper convergence rates than those obtained in Fan, Wang and Zhu (2021). More specifically, Fan, Wang and Zhu (2021) proposed a two-step procedure, which in step one applies shrinkage operators to the empirical average $(1/n)\sum_{i=1}^{n} y_i \boldsymbol{X}_i$. The truncation level on y_i's, which appears in the final convergence rate, depends on the variance of y_i and thus is not proportional to the noise scale. In contrast, by employing the adaptive Huber loss as in Sun, Zhou and Fan (2020) and the localized analysis developed by Fan *et al.* (2018), we show that the convergence rates of our estimators are proportional to the noise scale for matrix sensing and multitask regression; see Theorem 1.3.2, Theorem 1.3.4 and the subsequent remarks for details. Moreover, Fan, Wang and Zhu (2021) required ε_i to have bounded $(2k)$-th moment for some $k > 1$, while our estimators enjoy optimal rates as long as ε_i's have bounded variances. On the computational aspect, compared to the contractive Peaceman-Rachford splitting method and the alternating direction method of multiplier (ADMM) employed by Fan, Wang and Zhu (2021) to solve the nuclear norm penalized programs in step two, the proposed matrix variant of the LAMM algorithm is first-order and has a lower computational cost per iteration. See Section 1.2.2 for a more detailed comparison of computational complexity.

Although the proposed estimators satisfy exponential deviation bounds when the noise distribution is asymmetric and has finite variance, this advantage is accompanied by a trade-off: they sacrifice a considerable level of robustness when facing adversarial contamination of the data.

This is due to the use of a robustification parameter that increases with the sample size. In the case of adversarial contamination, recent studies have introduced robust estimators that showcase resistance to a small proportion of arbitrary outliers. For example, in sparse linear regression with Gaussian errors and adversarially corrupted labels, Dalalyan and Thompson (2019) demonstrated that the ℓ_1-penalized Huber's M-estimator attains the optimal rate of convergence, up to a logarithmic factor. Moreover, several recent studies (Chen *et al.*, 2013; Li, 2013; Klopp, Lounici and Tsybakov, 2017; Thompson, 2020) have specifically tackled the challenge of arbitrary outliers in the context of matrix sensing and matrix completion. For multitask regression, a robust multitask (reduced-rank) regression approach was introduced by She and Chen (2017) for simultaneous modeling and outlier detection. To address data contamination caused by arbitrary outliers, they formulated the problem as a regularized multivariate regression with a sparse mean-shift parametrization and developed a thresholding-based iterative procedure for optimization. It is worth noting that our methods and theory diverge from the conventional notion of robust statistics. While the aforementioned works assume sub-Gaussian or Gaussian noises, our work places emphasis on the distinct assumption of heavy-tailed errors rather than corruption by (arbitrary) outliers. The proposed methods and analysis therefore provide a useful complement to the current body of research on robust matrix completion and reduced-rank regression.

The rest of the paper proceeds as follows. In Section 1.2, we first review the trace regression model with three prototypical applications. Next, we introduce the nuclear norm penalized robust matrix estimator via the use of adaptive Huber loss, followed by a unified algorithm that applies to all three settings. We provide non-asymptotic high probability bounds for the proposed estimators case-by-case in Section 1.3. Section 1.4 presents our numerical experiments, conducted to demonstrate the advantage of our methods over their non-robust counterparts and to corroborate the theoretical findings that the convergence rates are proportional to the noise scale under the matrix sensing and multitask regression settings. All the proofs are relegated to the Appendix in the Supplementary Material Yu, Sun and Zhou (2023).

NOTATION. For a matrix $\boldsymbol{A} = (A_{jk})_{1\le j\le d_1, 1\le k\le d_2} \in \mathbb{R}^{d_1\times d_2}$, its singular values are denoted as $\sigma_1(\boldsymbol{A}) \ge \sigma_2(\boldsymbol{A}) \ge \cdots \ge \sigma_{\min(d_1,d_2)}(\boldsymbol{A})$. Define its operator norm $\|\boldsymbol{A}\|_2 = \sigma_1(\boldsymbol{A})$, its Frobenius norm $\|\boldsymbol{A}\|_{\mathrm{F}} = \sum_{j=1}^{\min(d_1,d_2)} \sigma_j^2(\boldsymbol{A})$, its nuclear norm $\|\boldsymbol{A}\|_* = \sum_{j=1}^{\min(d_1,d_2)} \sigma_j(\boldsymbol{A})$ and its max norm $\|\boldsymbol{A}\|_\infty = \max_{1\le j\le d_1} \max_{1\le k\le d_2} |A_{jk}|$. For two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d_1\times d_2}$, let $\langle \boldsymbol{A}, \boldsymbol{B}\rangle$ be the matrix inner product defined as $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B})$. We use $\mathrm{vec}(\boldsymbol{A}) \in \mathbb{R}^{d_1 d_2}$ to denote the long vector obtained by stacking the columns of $\boldsymbol{A}$.

1.2 Robust matrix recovery via adaptive Huber loss

1.2.1 Model and methods

Suppose we have collected n i.i.d. data points $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^n$ generated according to the following heteroscedastic trace regression model

$$y_i = \langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle + \varepsilon_i, \tag{1.1}$$

where $\boldsymbol{X}_i \in \mathbb{R}^{d_1\times d_2}$'s are random measurement matrices, and ε_i's are additive random noise variables satisfying $\mathbb{E}(\varepsilon_i|\boldsymbol{X}_i) = 0$ and $\mathbb{E}(\varepsilon_i^2|\boldsymbol{X}_i) \le \sigma_0^2$. Based on the noisy observations $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^n$, we are interested in recovering the unknown matrix $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1\times d_2}$ that is either exactly or approximately low-rank. More specifically, assume for some $0 \le q \le 1$ and $\rho > 0$ that

$$\boldsymbol{\Theta}^* \in \mathscr{B}_q(\rho) := \left\{ \boldsymbol{\Theta} \in \mathbb{R}^{d_1\times d_2} : \sum_{j=1}^{\min(d_1,d_2)} \sigma_j(\boldsymbol{\Theta})^q \le \rho \right\}. \tag{1.2}$$

In particular, $\mathscr{B}_0(\rho) = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1\times d_2} : \mathrm{rank}(\boldsymbol{\Theta}) \le \rho\}$ denotes the set of matrices with rank at most ρ, and $\mathscr{B}_q(\rho)$ with $0 < q \le 1$ is set of approximately low-rank matrices. Throughout the rest of the paper, we assume without loss of generality that $d_1 \ge d_2$.

The difficulty of recovering $\boldsymbol{\Theta}^*$ varies depending on the random structures of the measurement matrices $\boldsymbol{X}_i$. Below we list three prototypical applications of model (1.1), which will be the main focus of this work.

(i) *Matrix sensing*: Matrix sensing often assumes that the entries of $\boldsymbol{X}_i \in \mathbb{R}^{d_1 \times d_2}$ are independently generated from the $\mathscr{N}(0,1)$ distribution. More generally, $\mathrm{vec}(\boldsymbol{X}_i)$'s are assumed to be zero-mean sub-Gaussian/sub-exponential random vectors.

(ii) *Matrix completion*: In matrix completion, $\boldsymbol{X}_i$ are randomly drawn from the set

$$\mathscr{X} = \{\boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\mathrm{T}}(d_2), 1 \le j \le d_1, 1 \le k \le d_2\},$$

where $\boldsymbol{e}_1(d), \ldots, \boldsymbol{e}_d(d)$ are the canonical basis vectors in $\mathbb{R}^d$.

(iii) *Multitask regression*: The multitask (reduced-rank) regression assumes

$$\boldsymbol{y}_i = (\boldsymbol{\Theta}^*)^{\mathrm{T}}\boldsymbol{x}_i + \boldsymbol{\varepsilon}_i, \;\; i = 1, \ldots, n, \tag{1.3}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{id_2})^{\mathrm{T}} \in \mathbb{R}^{d_2}$ are observed response vectors, $\boldsymbol{x}_i \in \mathbb{R}^{d_1}$ are covariate vectors, $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ is the target regression coefficient matrix, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id_2})^{\mathrm{T}} \in \mathbb{R}^{d_2}$ are independent zero-mean random noise vectors. For $i = 1, \ldots, n$ and $k = 1, \ldots, d_2$, define

$$y_{(i-1)d_2+k} = y_{ik}, \quad \boldsymbol{X}_{(i-1)d_2+k} = \boldsymbol{x}_i\boldsymbol{e}_k^{\mathrm{T}}(d_2) \;\text{ and }\; \varepsilon_{(i-1)d_2+k} = \varepsilon_{ik}. \tag{1.4}$$

Then the sample $\{(y_j, \boldsymbol{X}_j)\}_{j=1}^N$ with $N = nd_2$ satisfies model (1.1).

For matrix sensing and matrix completion with noisy measurements, a popular approach is the the following convex relaxation approach (Candès and Plan, 2009)

$$\widehat{\boldsymbol{\Theta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\Theta} \in \mathscr{C}} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - \langle \boldsymbol{X}_i, \boldsymbol{\Theta} \rangle)^2 + \lambda \|\boldsymbol{\Theta}\|_* \right\}, \tag{1.5}$$

where $\mathscr{C}$ is a convex feasible set of $\mathbb{R}^{d_1 \times d_2}$ and $\lambda > 0$ is a regularization parameter. When $\mathscr{C} = \mathbb{R}^{d_1 \times d_2}$, $\widehat{\boldsymbol{\Theta}}_\lambda$ is the matrix analog of the Lasso estimator for linear regression (Tibshirani,

1996). The statistical properties of $\widehat{\boldsymbol{\Theta}}_\lambda$ in (1.5), mainly nonasymptotic deviation bounds under various matrix norms, have been studied in the literature when the additive noises ε_i are either Gaussian or sub-Gaussian. The performance of such a least-square-type estimator may break down quickly when the noise distribution is heavier-tailed. This is because outliers occur more frequently and the square loss is very sensitive to outliers. The impact of heavy-tailed errors on low-rank matrix recovery can be alleviated by replacing the ℓ_2-loss with a more robust loss function, typified by the ℓ_1-loss and the Huber loss (Elsener and van de Geer, 2018). When the error distribution is not only heavy-tailed but also asymmetric around zero, the use of ℓ_1-loss or Huber loss with a fixed tuning parameter induces a bias that remains non-negligible as the number of measurements grows. For a better trade-off between robustness and bias, in the following we propose to use adaptive Huber loss (Fan, Li and Wang, 2017; Sun, Zhou and Fan, 2020) for robust low-rank matrix recovery, with a focus on the above three prototypical applications.

For matrix sensing and completion problems, i.e. applications (i) and (ii), we define the empirical loss function to be

$$\widehat{L}_\tau(\boldsymbol{\Theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell_\tau(y_i - \langle \boldsymbol{X}_i, \boldsymbol{\Theta}\rangle), \;\; \boldsymbol{\Theta} \in \mathbb{R}^{d_1\times d_2}, \tag{1.6}$$

where $\ell_\tau(u) = \min\{u^2/2, \tau|u| - \tau^2/2\}$ denotes the adaptive Huber loss parameterized by $\tau = \tau_n > 0$, referred to as the robustification parameter in Sun, Zhou and Fan (2020). For any pre-specified convex subset $\mathscr{C}$ of $\mathbb{R}^{d_1\times d_2}$, we consider the following nuclear norm penalized robust regression estimator

$$\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} \in \operatorname*{argmin}_{\boldsymbol{\Theta}\in\mathscr{C}} \big\{\widehat{L}_\tau(\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_*\big\}, \tag{1.7}$$

where $\tau = \tau_n > 0$ and $\lambda = \lambda_n > 0$ are the robustification and regularization parameters respectively.

For the multitask regression problem – Application (iii), recall that the vector-valued observations $\{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}_{i=1}^n$ can be written as $\{(y_j, \boldsymbol{X}_j)\}_{j=1}^N$ ($N = nd_2$) via (1.4) so model (1.1) can be used. The classical reduced-rank regression method is based on solving the rank-constrained problem (Izenman, 1975)

$$\min_{\mathrm{rank}(\boldsymbol{\Theta})\le r} \left\{ \sum_{i=1}^n \|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2^2 = \sum_{j=1}^N (y_j - \langle \boldsymbol{X}_j, \boldsymbol{\Theta}\rangle)^2 \right\},$$

for which an analytic solution is available. To robustify this classical procedure, similarly to the formulation (1.7) one may naively apply the Huber loss to each residual $y_j - \langle \boldsymbol{X}_j, \boldsymbol{\Theta}\rangle$. This, however, is no longer plausible because $\boldsymbol{X}_j$'s are now dependent random matrices. Moreover, since we do not impose independence on the entries of $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id_2})^{\mathrm{T}}$, $\boldsymbol{\varepsilon}_j$'s defined in (1.4) may also be highly correlated. We propose to replace the ℓ_2-loss on $\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2$ with the Huber loss, leading to $\min_{\mathrm{rank}(\boldsymbol{\Theta})\le r} \sum_{i=1}^n \ell_\tau(\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2)$, which is a highly nonconvex problem. Similarly to (1.7), we resort to convex relaxation and consider the following nuclear norm penalized estimator

$$\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} \in \operatorname*{argmin}_{\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}} \left\{ \frac{1}{n}\sum_{i=1}^n \ell_\tau\big(\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2\big) + \lambda\|\boldsymbol{\Theta}\|_* \right\}. \tag{1.8}$$

In Section 1.3, we characterize the nonasymptotic statistical accuracy for the robust low-rank estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ defined in (1.7) and (1.8) when the noise variables only have bounded variances. The key is to seek a suitable choice of τ and λ jointly to trade off among robustness, bias and approximation errors.

1.2.2 Algorithms

To solve the nuclear norm penalized optimization programs (1.7) and (1.8), in this section we present a unified algorithm by extending the local adaptive majorize-minimization (LAMM) principle proposed in Fan *et al.* (2018) to matrix settings. Recall that the proposed nuclear norm

penalized Huber regression estimators have a general form

$$\widehat{\mathbf{\Theta}}_{\tau,\lambda} \in \underset{\mathbf{\Theta}\in\mathscr{C}}{\operatorname{argmin}}\{\widehat{L}_\tau(\mathbf{\Theta}) + \lambda\|\mathbf{\Theta}\|_*\},$$

where $\widehat{L}_\tau(\mathbf{\Theta})$ is the empirical loss in (1.6) or (1.8), and $\mathscr{C}$ is taken to be $\mathbb{R}^{d_1\times d_2}$ or $\{\mathbf{\Theta}\in\mathbb{R}^{d_1\times d_2} : \|\mathbf{\Theta}\|_\infty \le \alpha_0\}$ for some $\alpha_0 > 0$. The main idea of the LAMM principle is to construct an isotropic quadratic function that locally majorizes the objective function at each iteration. In the matrix setting, given the previous iterate $\mathbf{\Theta}^{(k-1)}$ at the k-th iteration, define the quadratic function

$$F(\mathbf{\Theta};\mathbf{\Theta}^{(k-1)},\phi_k) = \widehat{L}_\tau(\mathbf{\Theta}^{(k-1)}) + \langle\nabla\widehat{L}_\tau(\mathbf{\Theta}^{(k-1)}),\mathbf{\Theta}-\mathbf{\Theta}^{(k-1)}\rangle + \frac{\phi_k}{2}\|\mathbf{\Theta}-\mathbf{\Theta}^{(k-1)}\|_{\mathrm{F}}^2,$$

satisfying $F(\mathbf{\Theta}^{(k-1)};\mathbf{\Theta}^{(k-1)},\phi_k) = \widehat{L}_\tau(\mathbf{\Theta}^{(k-1)})$, where $\phi_k > 0$ is a quadratic parameter. Next, define the k-th iterate as

$$\mathbf{\Theta}^{(k)} \in \underset{\mathbf{\Theta}\in\mathscr{C}}{\operatorname{argmin}}\{F(\mathbf{\Theta};\mathbf{\Theta}^{(k-1)},\phi_k) + \lambda\|\mathbf{\Theta}\|_*\}. \tag{1.9}$$

The parameter ϕ_k needs to be sufficiently large so that $\widehat{L}_\tau(\mathbf{\Theta}^{(k)}) \le F(\mathbf{\Theta}^{(k)};\mathbf{\Theta}^{(k-1)},\phi_k)$, which further implies

$$\begin{aligned}
\widehat{L}(\mathbf{\Theta}^{(k)}) + \lambda\|\mathbf{\Theta}^{(k)}\|_* &\le F(\mathbf{\Theta}^{(k)};\mathbf{\Theta}^{(k-1)},\phi_k) + \lambda\|\mathbf{\Theta}^{(k)}\|_* \\
&\le F(\mathbf{\Theta}^{(k-1)};\mathbf{\Theta}^{(k-1)},\phi_k) + \lambda\|\mathbf{\Theta}^{(k-1)}\|_* \\
&= \widehat{L}_\tau(\mathbf{\Theta}^{(k-1)}) + \lambda\|\mathbf{\Theta}^{(k-1)}\|_*,
\end{aligned}$$

where the second inequality is due to the optimality of $\mathbf{\Theta}^{(k)}$. This ensures the descent of the objective function at each iteration. To choose a sufficiently large ϕ_k, we start from a small value, say $\phi_0 = 0.01$, and inflate it by a factor $\gamma > 1$, say $\gamma = 2$, until the local majorization requirement $\widehat{L}_\tau(\mathbf{\Theta}^{(k)}) \le F(\mathbf{\Theta}^{(k)};\mathbf{\Theta}^{(k-1)},\phi_k)$ is met. Since $F(\mathbf{\Theta};\mathbf{\Theta}^{(k-1)},\phi_k) \ge \widehat{L}_\tau(\mathbf{\Theta})$ when ϕ_k is no less than the largest eigenvalue of $\nabla^2\widehat{L}_\tau(\mathbf{\Theta}^{(k-1)})$, the iteration will stop after sufficiently many

steps. Repeat the above steps until convergence (e.g., $\|\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Theta}^{(k-1)}\|_{\mathrm{F}} \leq \varepsilon$ for a sufficiently small $\varepsilon > 0$) or until the maximum number of iterations is reached.

The main benefit of minimizing a penalized isotropic quadratic objective function is that the minimizer often has a closed form. For any matrix $\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}$ with rank r, consider the singular value decomposition (SVD) $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are, respectively, $d_1 \times r$ and $d_2 \times r$ matrices with orthonormal columns, and $\boldsymbol{\Sigma} = \mathrm{diag}(\{\sigma_i\}_{1 \leq i \leq r})$ is an $r \times r$ diagonal matrix with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. For $\lambda > 0$, define the soft-thresholding operator $S(\boldsymbol{\Theta}, \lambda) = \boldsymbol{U} \cdot \mathrm{diag}(\{\max(\sigma_i - \lambda, 0)\}_{1 \leq i \leq r}) \cdot \boldsymbol{V}^{\mathrm{T}}$. By Theorem 2.1 in Cai, Candès and Shen (2010), $\boldsymbol{\Theta}^{(k)}$ given in (1.9) with $\mathscr{C} = \mathbb{R}^{d_1 \times d_2}$ admits the closed-form expression

$$\boldsymbol{\Theta}^{(k)} = T_{\lambda, \phi_k}(\boldsymbol{\Theta}^{(k-1)}) := S(\boldsymbol{\Theta}^{(k-1)} - \phi_k^{-1} \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)}), \phi_k^{-1}\lambda).$$

For a general convex subset $\mathscr{C} \subseteq \mathbb{R}^{d_1 \times d_2}$, we can update $\boldsymbol{\Theta}^{(k)}$ as

$$\boldsymbol{\Theta}^{(k)} = \Pi_{\mathscr{C}}(T_{\lambda, \phi_k}(\boldsymbol{\Theta}^{(k-1)})),$$

where $\Pi_{\mathscr{C}}$ denotes Euclidean projection onto the subspace $\mathscr{C}$. When $\mathscr{C} = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Theta}\|_\infty \leq \alpha_0\}$, for example, $\Pi_{\mathscr{C}}(\boldsymbol{\Theta}) = (\max\{\min(\boldsymbol{\Theta}_{jk}, \alpha_0), -\alpha_0\})_{1 \leq j \leq d_1, 1 \leq k \leq d_2}$. We summarize the key steps in Algorithm 1.

As a unified algorithm, Algorithm 1 applies to all three problems considered in this paper, matrix sensing, matrix completion and multitask regression. In terms of complexity, at each iteration $\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)})$ and $T_{\lambda, \phi_k}(\boldsymbol{\Theta}^{(k-1)})$ can be computed in $O(nd_1d_2)$ and $O(d_1d_2^2)$ operations (assuming $d_1 \geq d_2$), respectively (Trefethen and Bau III, 1997). On the other hand, Fan, Wang and Zhu (2021) employed the contractive Peaceman-Rachford splitting method for matrix sensing and multitask regression, and an ADMM-based algorithm for matrix completion. In addition to the operations described above, each ADMM iterate also involves computing the inverse of $2\mathbb{X}^{\mathrm{T}}\mathbb{X}/n + \mathbf{I}_{d_1d_2}$, where $\mathbb{X}$ is an $n \times d_1d_2$ matrix whose i-th row is $\mathrm{vec}(\boldsymbol{X}_i)$, and $\mathbf{I}_k$

Algorithm 1. LAMM algorithm for regularized adaptive Huber trace regression

Algorithm: $\{\boldsymbol{\Theta}^{(k)}\}_{k=1}^{\infty} \leftarrow \mathrm{LAMM}(\lambda, \boldsymbol{\Theta}^{(0)}, \phi_0, \gamma, \varepsilon)$

Input: $\lambda, \boldsymbol{\Theta}^{(0)}, \phi_0, \gamma, \varepsilon$

for $k = 1, 2, \ldots$ until $\|\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Theta}^{(k-1)}\|_2 \le \varepsilon$ **do**
  repeat
    $\boldsymbol{\Theta}^{(k)} \leftarrow S(\boldsymbol{\Theta}^{(k-1)} - \phi_k^{-1} \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)}), \phi_k^{-1}\lambda)$
    $\boldsymbol{\Theta}^{(k)} \leftarrow \Pi_{\mathscr{C}}(\boldsymbol{\Theta}^{(k)})$
    if $F(\boldsymbol{\Theta}^{(k)}; \boldsymbol{\Theta}^{(k-1)}, \phi_k) < \widehat{L}_\tau(\boldsymbol{\Theta}^{(k)})$ **then**
      $\phi_k \leftarrow \gamma \cdot \phi_k$
    end if
  until $F(\boldsymbol{\Theta}^{(k)}; \phi_k, \boldsymbol{\Theta}^{(k-1)}) \ge \widehat{L}_\tau(\boldsymbol{\Theta}^{(k)})$
  return $\boldsymbol{\Theta}^{(k)}$
end for

Output: $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(T)}$

denotes the $k \times k$ identity matrix. By applying the Sherman–Morrison–Woodbury formula, this step can be implemented in $O(\min\{n, d_1 d_2\}^3)$ operations. Still, the computational complexity and storage cost (per iteration) of ADMM are much higher than the LAMM algorithm in the context of matrix completion, especially for large-scale datasets.

1.3 Theoretical guarantees

In this section, we establish the finite-sample statistical properties of the robust estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ for matrix sensing, matrix completion and multitask regression. Throughout, the noise variables ε_i in (1.1) and $\boldsymbol{\varepsilon}_i$ in (1.3) are assumed to have bounded variance only, and we do not require independence between ε_i and $\boldsymbol{X}_i$ or $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{x}_i$.

1.3.1 Matrix sensing

In the case of matrix compressed sensing, we set $\mathscr{C} = \mathbb{R}^{d_1 \times d_2}$ in (1.7), and impose the following assumptions.

(A1) $\boldsymbol{\Theta}^* \in \mathscr{B}_q(\rho)$ for some $0 \le q \le 1$ and $\rho > 0$.

(A2) The random matrix $\boldsymbol{X}_i \in \mathbb{R}^{d_1 \times d_2}$ satisfies (i) $\mathbb{E}\boldsymbol{X}_i = \boldsymbol{0}$, and (ii) $\mathrm{vec}(\boldsymbol{X}_i) \in \mathbb{R}^{d_1 d_2}$ is sub-

exponential, that is, there exists a constant $\nu_0 \geq 1$ such that for any $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$ and $u \geq 0$,

$$\mathbb{P}\big(|\mathrm{vec}(\boldsymbol{A})^{\mathrm{T}}\mathrm{vec}(\boldsymbol{X}_i)| \geq \nu_0 \|\boldsymbol{A}\|_{\mathrm{F}} \cdot u\big) \leq 2e^{-u}.$$

Moreover, there exists a constant $c_l > 0$ such that $\lambda_{\min}\big(\mathbb{E}\mathrm{vec}(\boldsymbol{X}_i)\mathrm{vec}(\boldsymbol{X}_i)^{\mathrm{T}}\big) \geq c_l$.

(A3) The noise variables ε_i are such that $\mathbb{E}(\varepsilon_i|\boldsymbol{X}_i) = 0$ and $\mathbb{E}(\varepsilon_i^2|\boldsymbol{X}_i) \leq \sigma_0^2$ (almost surely) for some constant $\sigma_0 > 0$.

Remark 1.3.1. The parameter ν_0 is often referred to as the sub-exponential parameter. For various well-behaved distributions on $\mathbb{R}^{d_1 \times d_2}$, the associated sub-exponential parameters are independent of the dimensions d_1 and d_2. As prototypical examples, the distributions listed below satisfy Condition (A2) with dimension-free parameters ν_0 and c_l.

(i) (Multivariate normal) $\mathrm{vec}(\boldsymbol{X}_i)$ follow s $\mathcal{N}(\boldsymbol{0}, \Sigma)$ with a positive-definite $\Sigma \in \mathbb{R}^{(d_1 d_2)\times(d_1 d_2)}$.

(ii) (Uniform distribution on the Euclidean sphere) $\boldsymbol{X}_i$ follows the uniform distribution on the sphere centered at the origin with radius $(d_1 d_2)^{1/2}$, namely, $\{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{X}\|_{\mathrm{F}} = (d_1 d_2)^{1/2}\}$.

(iii) (Uniform distribution on the ℓ_1-ball) $\boldsymbol{X}_i$ follows the uniform distribution on the ℓ_1-norm ball centered at the origin with radius $r \asymp d_1 d_2$, that is, $\{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2} : \sum_{j=1}^{d_1}\sum_{k=1}^{d_2} |X_{jk}| \leq r\}$.

Here, we note that the multivariate distributions in (i) and (ii) are not only sub-exponential but also sub-Gaussian.

To derive the convergence rate of $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ under either the Frobenius norm or the nuclear norm, we first define a probability event that concerns the local restricted strong convexity (RSC) of the empirical loss $\widehat{L}_\tau(\cdot)$. For $s, l > 0$, define the Frobenius norm ball and trace norm cone

$$\mathbb{B}(s) = \{\boldsymbol{\Delta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Delta}\|_{\mathrm{F}} \leq s\} \text{ and } \mathbb{C}(l) = \{\boldsymbol{\Delta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Delta}\|_* \leq l\|\boldsymbol{\Delta}\|_{\mathrm{F}}\}, \tag{1.10}$$

respectively.

Definition 1.3.1 (Local restricted strong convexity)**.** Given radius parameters $s, l > 0$ and a curvature parameter $\kappa > 0$, define the event

$$\mathscr{E}(s,l,\kappa) = \left\{ \inf_{\boldsymbol{\Theta} \in \boldsymbol{\Theta}^* + \mathbb{B}(s) \cap \mathbb{C}(l)} \frac{\langle \nabla \widehat{L}_\tau(\boldsymbol{\Theta}) - \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*), \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \rangle}{\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2} \geq \kappa \right\}, \tag{1.11}$$

which concerns the local restricted strong convexity of the empirical loss function.

We first provide a deterministic result on the convergence rate of $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$: for any choice of λ such that $\|\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2 \leq \lambda/2$, and conditioned on event $\mathscr{E}(s,l,\kappa)$ with suitably chosen (s,l), we are guaranteed that

$$\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \lesssim \sqrt{\rho}\,(\lambda/\kappa)^{1-q/2} \quad \text{and} \quad \|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_* \lesssim \rho(\lambda/\kappa)^{1-q}.$$

Theorem 1.3.1. Assume Condition (A1) holds. Let (λ, s, l, κ) satisfy $\lambda \geq 2\|\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2$,

$$s \geq 9.15\sqrt{\rho}\,(\lambda/\kappa)^{1-q/2} \quad \text{and} \quad l \geq 6.1\sqrt{\rho}\,(\kappa/\lambda)^{q/2}. \tag{1.12}$$

Conditioned on the event $\mathscr{E}(s,l,\kappa)$, the error matrix $\widehat{\boldsymbol{\Delta}} := \widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*$ satisfies

$$\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}} \leq 9.15\sqrt{\rho}\left(\frac{\lambda}{\kappa}\right)^{1-q/2} \quad \text{and} \quad \|\widehat{\boldsymbol{\Delta}}\|_* \leq 56\rho\left(\frac{\lambda}{\kappa}\right)^{1-q}.$$

In the following two propositions, we first derive an upper bound of $\|\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2$, and then establish the local RSC property of the empirical Huber loss function $\widehat{L}_\tau(\cdot)$. Together, these results show that with properly chosen λ, τ that depend on (n,d,s,l) along with the distributional parameters in Conditions (A2) and (A3), the event $\{\lambda \geq 2\|\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2\} \cap \mathscr{E}(s,l,c_l/4)$ occurs with high probability.

Proposition 1.3.1. Assume Conditions (A2) and (A3) hold. For any $\sigma \geq \sigma_0$ and $z > 0$, the

empirical Huber loss $\widehat{L}_\tau(\cdot)$ with $\tau = \sigma\sqrt{n/(3d+z)}$ satisfies

$$\|\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2 \leq 10\nu_0 \cdot \sigma\sqrt{\frac{3d+z}{n}} \tag{1.13}$$

with probability at least $1-e^{-z}$, where $d = d_1 + d_2$.

Proposition 1.3.2. Assume Conditions (A2) and (A3) hold. For any $s, l > 0$ and $z > 0$, let τ and n satisfy

$$\tau \geq 4\nu_0\sqrt{(2\sigma_0^2 + 96\nu_0^2 s^2)/c_l} \quad \text{and} \quad n \geq C_1(\tau/s)^2(l^2 d + z), \tag{1.14}$$

where $d = d_1 + d_2$ and $C_1 > 0$ is a constant depending only on ν_0 and c_l. Then, the local RSC event $\mathscr{E}(s, l, \kappa)$ with $\kappa = c_l/4$ occurs with probability at least $1 - e^{-z}$.

Combining these high probability bounds with Theorem 1.3.1 leads to the convergence rate of $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$, as stated in the following theorem.

Theorem 1.3.2. Assume Conditions (A1)–(A3) hold. For any $z > 0$, the robust (approximately) low-rank matrix estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ defined in (1.7) with $\mathscr{C} = \mathbb{R}^{d_1 \times d_2}$, $\tau \asymp \sigma_0\sqrt{n/(d+z)}$ and $\lambda \asymp \sigma_0\sqrt{(d+z)/n}$ satisfies

$$\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \lesssim \sigma_0^{1-q/2}\sqrt{\rho}\left(\frac{d+z}{n}\right)^{1/2-q/4} \quad \text{and} \quad \|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_* \lesssim \sigma_0^{1-q}\rho\left(\frac{d+z}{n}\right)^{(1-q)/2}$$

with probability at least $1 - 2e^{-z}$ as long as $n \gtrsim \max\left\{(\rho/\sigma_0^q)^{2/(2-q)}, 1\right\}(d+z)$, where $d = d_1 + d_2$.

Remark 1.3.2. In the exact low-rank case, i.e. $q = 0$ and $\rho = r = \mathrm{rank}(\boldsymbol{\Theta}^*)$, the results in Theorem 1.3.2 imply that with high probability (over both the random sensing matrices $\boldsymbol{X}_i$ and noise variables ε_i), the robust estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ satisfies with high probability that

$$\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \lesssim \sigma_0\sqrt{\frac{rd}{n}} \text{ as long as } n \gtrsim rd.$$

Within a constant independent of (n, r, d_1, d_2) and noise scale σ_0, this upper bound matches the information-theoretic lower bound established by Candès and Plan (2011) when ε_i are i.i.d. $\mathscr{N}(0, \sigma_0^2)$. The robustness manifests in two aspects. First, the noise distribution is only required to have bounded variance as opposed to sub-Gaussian tails. Secondly, we assume the random vector $\mathrm{vec}(\boldsymbol{X}_i)$ is sub-exponential, whereas $\mathrm{vec}(\boldsymbol{X}_i)$ is often assumed to have i.i.d. Gaussian/sub-Gaussian entries in the literature.

Remark 1.3.3. For matrix compressed sensing, based on a shrinkage principle Fan, Wang and Zhu (2021) also proposed a robust low-rank estimator, which achieves near-optimal rate under heavy-tailed noise distributions. Its recovery guarantees (see Theorem 3 therein), however, depend on more stringent assumptions as needed in Theorem 1.3.2. In addition to Conditions (A2) and (A3), Fan, Wang and Zhu (2021) assumed further that (i) $\mathrm{vec}(\boldsymbol{X}_i)$ is sub-Gaussian, and (ii) $\mathbb{E}|y_i|^{2k} \leq M_k$ for some $k > 1$. Under these conditions and in the exact low-rank case (for brevity), their truncate/shrinkage estimator, denoted by $\widetilde{\boldsymbol{\Theta}}$, satisfies with high probability the bound

$$\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \leq M_k^{1/(2k)} \sqrt{\frac{rd}{n}} \text{ as long as } n \gtrsim rd.$$

The above convergence rate is sub-optimal in terms of its dependence on the noise scale σ_0. As the noise scale decreases, $M_k^{1/(2k)}$ remains to be bounded away from zero because

$$M_k^{1/k} > \mathbb{E} y_i^2 = \mathbb{E}\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle^2 + \mathbb{E}(\varepsilon_i^2).$$

Remark 1.3.4. The sample size requirement in Theorem 1.3.2 becomes more stringent as σ_0 goes to 0 when $q \neq 0$. This is an artifact of the technical argument used in the proof of the theorem. A similar sample size requirement, characterized by its inverse proportionality to the moment of the response, can be found in Theorem 3 of Fan, Wang and Zhu (2021). To modify

the sample size requirement, we can choose

$$\tau \asymp \max(\sigma_0, 1)\sqrt{n/(d+z)} \text{ and } \lambda \asymp \max(\sigma_0, 1)\sqrt{(d+z)/n}$$

for a given $z > 0$. This results in a revised sample size requirement of $n \gtrsim \max(\rho^{2/(2-q)}, 1)(d+z)$, accompanied by an error bound

$$\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \lesssim \max(\sigma_0^{1-q/2}, 1)\sqrt{\rho}\left(\frac{d+z}{n}\right)^{1/2-q/4}$$

with probability at least $1 - 2e^{-z}$. This error bound is similar to the deviation bound in Corollary 5 of Negahban and Wainwright (2011), but it should be noted that they are not proportional to the noise level. Additionally, Theorem 1.3.4 in Section 1.3.3 also relies on a similar technical argument, necessitating a larger sample size as σ_0 approaches zero.

1.3.2 Matrix completion

This subsection investigates matrix completion under the following assumptions.

(B1) $\boldsymbol{\Theta}^* \in \mathscr{B}_q(\rho)$ and $\|\boldsymbol{\Theta}^*\|_\infty \le \alpha_0$ for some $\alpha_0 > 0$. We thus set $\mathscr{C} = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Theta}\|_\infty \le \alpha_0\}$ in (1.7) so that $\boldsymbol{\Theta}^* \in \mathscr{C}$.

(B2) $\boldsymbol{X}_i \in \mathbb{R}^{d_1 \times d_2}$ is uniformly sampled from $\{\boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\mathrm{T}}(d_2)\}_{1 \le j \le d_1, 1 \le k \le d_2}$, where $\{\boldsymbol{e}_j(d)\}_{j=1}^d$ are the canonical basis vectors in $\mathbb{R}^d$. Specifically, $\mathbb{P}\{\boldsymbol{X}_i = \boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\mathrm{T}}(d_2)\} = (d_1 d_2)^{-1}$.

(B3) $\mathbb{E}(\varepsilon_i|\boldsymbol{X}_i) = 0$ and $\mathbb{E}(\varepsilon_i^2|\boldsymbol{X}_i) \le \sigma_0^2$ (almost surely) for some constant $\sigma_0 > 0$.

Remark 1.3.5. In addition to the assumption that $\boldsymbol{\Theta}^*$ is of (approximately) low-rank, we require in Condition (B1) that $\|\boldsymbol{\Theta}^*\|_\infty \le \alpha_0$ for some $\alpha_0 > 0$. Past works on noisy matrix completion also imposed the same or similar conditions. For instance, Klopp (2014) and Minsker (2018) assumed that $\|\boldsymbol{\Theta}^*\|_\infty$ is bounded; Negahban and Wainwright (2012) and Fan, Wang and Zhu (2021) required the spikiness ratio $\|\boldsymbol{\Theta}^*\|_\infty / \|\boldsymbol{\Theta}^*\|_{\mathrm{F}}$ to be bounded; Candès and Plan (2009) and Candès

and Recht (2009) relied on matrix incoherence conditions. Without such extra conditions, the number of measurements should satisfy $n \asymp d_1 d_2$ in order to recover $\mathbf{\Theta}^*$ in the worst case; see Candès and Recht (2009) and Negahban and Wainwright (2012) for details.

Similarly to the matrix sensing case, the key steps to establish the convergence rate of $\widehat{\mathbf{\Theta}}_{\tau,\lambda}$ are (i) an upper bound of $\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2$ as shown in Proposition 1.3.3 below, and (ii) a lower bound for

$$\langle \nabla \widehat{L}_\tau(\mathbf{\Theta}) - \nabla \widehat{L}_\tau(\mathbf{\Theta}^*), \mathbf{\Theta} - \mathbf{\Theta}^* \rangle$$

uniformly over $\mathbf{\Theta}$ in a neighborhood of $\mathbf{\Theta}^*$. The sparsity of $\boldsymbol{X}_i$ in this case (see Condition (B2)) introduces more subtleties into the analysis of the latter, as we will see from Proposition 1.3.4.

Proposition 1.3.3. Assume Conditions (B2) and (B3) hold. For any $\sigma \geq \sigma_0$ and $z > 0$, the loss function $\widehat{L}_\tau(\cdot)$ with $\tau = \sigma\sqrt{n/\{d_2(z+\log d)\}}$ satisfies with probability at least $1 - e^{-z}$ that

$$\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2 \leq (3\sigma_0 + 2\sigma/3)\sqrt{\frac{z+\log d}{d_2 n}}, \tag{1.15}$$

where $d = d_1 + d_2$.

Proposition 1.3.4. Assume Conditions (B2) and (B3) hold. For any $s, l > 0$ and $z > 0$, let τ and n satisfy

$$\tau^2 \geq 16 \max[ns^2/\{l^2 d_1^2 d_2(z+\log d)\}, \sigma_0^2] \quad \text{and} \quad n \geq d_2 \log d,$$

where $d = d_1 + d_2$. Define the constrain set

$$\mathbb{A}(s,l) = \left\{ \mathbf{\Delta} \in \mathbb{B}(s) \cap \mathbb{C}(l) : \frac{\|\mathbf{\Delta}\|_\infty^2}{\|\mathbf{\Delta}\|_{\mathrm{F}}^2/(d_1 d_2)} \leq \frac{1}{8}\sqrt{\frac{n}{z+\log d}} \right\}, \tag{1.16}$$

where $\mathbb{B}(s)$ and $\mathbb{C}(l)$ are given in (1.10). Then, for all $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$ with $\mathbf{\Delta} := \mathbf{\Theta} - \mathbf{\Theta}^* \in \mathbb{A}(s,l)$,

we have with probability at least $1-e^{-z}$ that

$$\langle \nabla \widehat{L}_\tau(\boldsymbol{\Theta}) - \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*), \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \rangle \geq \frac{1}{4d_1d_2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 - C_0 l^2 \frac{d_1(z+\log d)}{n}\|\boldsymbol{\Delta}\|_\infty^2,$$

where $C_0 > 1$ is an absolute constant.

With the above preparations, we now state the statistical guarantees for matrix completion under heavy-tailed noise.

Theorem 1.3.3. Assume Conditions (B1)–(B3) hold. For any $z > 0$, set

$$\tau \asymp \sigma \sqrt{\frac{n}{d_2(z+\log d)}} \text{ and } \lambda \asymp \sigma \sqrt{\frac{z+\log d}{d_2 n}},$$

where $\sigma = \max\{\sigma_0, \alpha_0\}$ and $d = d_1 + d_2$. Then, the robust (approximately) low-rank matrix estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ defined in (1.7) with $\mathscr{C} = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Theta}\|_\infty \leq \alpha_0\}$ satisfies

$$\frac{1}{d_1 d_2}\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 \lesssim \max\left[\sigma^{2-q}\rho\left\{\frac{d_1(z+\log d)}{n}\right\}^{1-q/2}, \alpha_0^2\sqrt{\frac{z+\log d}{n}}\right] \tag{1.17}$$

and

$$\begin{aligned}
&\frac{1}{\sqrt{d_1 d_2}}\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_* \\
&\qquad \lesssim \max\left[\sigma^{1-q}\rho\left\{\frac{d_1(z+\log d)}{n}\right\}^{\frac{1-q}{2}}, \alpha_0^{\frac{2(1-q)}{2-q}} \frac{\rho^{\frac{1}{2-q}}}{(d_1 d_2)^{\frac{q}{2(2-q)}}}\left(\frac{z+\log d}{n}\right)^{\frac{1-q}{2(2-q)}}\right]
\end{aligned}$$

with probability at least $1-2e^{-z}$ whenever $n \gtrsim d_2(z+\log d)$.

Remark 1.3.6. In the context of matrix completion, one is interested in recovering a large low-rank data matrix from a highly incomplete subset of its entries. A natural assumption is $n \leq d_1 d_2$, which in turn implies $\sqrt{\log(d)/n} \leq d_1 \log(d)/n$, where $d = d_1 + d_2$ and $d_1 \geq d_2$. Therefore, taking $z = \log n$, the maximum in (1.17) is often given by its first term. In the exact

low-rank case, the general results in Theorem 1.3.3 imply that the proposed robust estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ with $\tau \asymp \sigma_0\sqrt{n/(d_2 \log d)}$ and $\lambda \asymp \sigma_0\sqrt{\log(d)/(d_2 n)}$ satisfies the bound

$$\frac{1}{d_1 d_2}\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 \lesssim \max\{\alpha_0^2, \sigma_0^2\}\frac{r d_1 \log d}{n} \tag{1.18}$$

with high probability as long as $n \gtrsim d_2 \log d$. Under our notations, Theorem 6 in Koltchinskii, Lounici and Tsybakov (2011) shows that when $\varepsilon_i \sim \mathcal{N}(0, \sigma_0^2)$ is independent of $\boldsymbol{X}_i$, there exist absolute constants $\beta \in (0,1)$ and $c > 0$ such that

$$\inf_{\widehat{\boldsymbol{\Theta}}} \sup_{\mathrm{rank}(\boldsymbol{\Theta}^*)\le r, \|\boldsymbol{\Theta}^*\|_\infty \le \alpha_0} \mathbb{P}\left\{\frac{1}{d_1 d_2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 > c\min(\sigma_0^2, \alpha_0^2)\frac{r d_1}{n}\right\} \ge \beta,$$

where $\inf_{\widehat{\boldsymbol{\Theta}}}$ is the infimum over all estimators $\widehat{\boldsymbol{\Theta}} \in \mathbb{R}^{d_1 \times d_2}$. Therefore, the rate derived in Theorem 1.3.3 is minimax optimal up to a logarithmic factor and a trailing term.

Remark 1.3.7 (Comparison to existing work on robust (noisy) matrix completion)**.** In the context of matrix completion with heavy-tailed noise, several robust estimators have been proposed and studied. Minsker (2018) proposed a two-step method that computes a truncation-type matrix estimator, denoted by $\widetilde{\boldsymbol{\Theta}}$, in step one and then solves the nuclear norm penalized optimization $\|\boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2/(d_1 d_2) + \lambda\|\boldsymbol{\Theta}\|_*$. In the exact low-rank case, this two-step estimator satisfies a high probability bound, which is similar to (1.17) with $q = 0$, when ε_i is independent of $\boldsymbol{X}_i$ and has bounded variance. The independence assumption can be removed by slightly modifying the proof in Minsker (2018). For matrix sensing and multitask regression, it is unclear whether such a two-step procedure will also lead to robust estimates that satisfy sharp error bounds proportional only to the noise scale. Concurrently, Fan, Wang and Zhu (2021) considered a similar two-step estimator, but their theoretical result requires a slightly stronger moment condition, i.e. $\mathbb{E}\{\mathbb{E}(\varepsilon_i^2|\boldsymbol{X}_i)^k\} \le M_k$ for some $k > 1$. Our proposal is more relevant to Elsener and van de Geer (2018), who also used the Huber loss for matrix completion in the presence of heavy-tailed errors. Their results, however, depend on stronger assumptions on the error distribution. In addition to

Conditions (B1) and (B2), they further assumed that (i) the distribution of ε_i is symmetric around 0, and (ii) there exists a constant $C_1 > 0$ such that the cumulative distribution function $F(\cdot)$ of ε_i satisfies

$$F(x+\tau) - F(x-\tau) \geq 1/C_1^2 \quad \text{for all } |x| \leq 2\alpha_0 \text{ and } \tau \leq 2\alpha_0.$$

Under theses conditions and in the exact low-rank case, Elsener and van de Geer (2018) proved that the nuclear norm penalized Huber regression estimator, denoted by $\check{\boldsymbol{\Theta}}$, satisfies

$$\frac{1}{d_1 d_2}\|\check{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 = O_{\mathbb{P}}\left\{ \max(\alpha_0^2, \tau^2) C_1^4 \frac{r d_1 \log(d_1+d_2)}{n} \right\}$$

under the sample size requirement $n \gtrsim d_2 \log(d_2)\log(d_1+d_2)$.

1.3.3 Multitask regression

In this section, we establish the statistical properties of the robust low-rank multitask (reduced-rank) regression estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ (1.8). With slight abuse of notation, we write

$$\widehat{L}_\tau(\boldsymbol{\Theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell_\tau\big(\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2\big), \;\; \boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}, \tag{1.19}$$

where $\tau > 0$ is the robustification parameter.

(C1) $\boldsymbol{\Theta}^* \in \mathscr{B}_q(\rho)$ for some $0 \leq q \leq 1$ and $\rho > 0$.

(C2) $\boldsymbol{x}_i \in \mathbb{R}^{d_1}$ are i.i.d. zero-mean sub-Gaussian vectors, that is, there exists a (dimension-free) constant $\nu_0 \geq 1$ such that

$$\mathbb{E}e^{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_i} \leq e^{\nu_0^2\|\boldsymbol{u}\|_2^2/2}, \;\; \text{valid for any } \boldsymbol{u} \in \mathbb{R}^{d_1}.$$

Moreover, there exists a constant $c_l > 0$ such that $\lambda_{\min}\big(\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\big) \geq c_l$.

(C3) The noise vectors $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{d_2}$ are such that $\mathbb{E}(\boldsymbol{\varepsilon}_i|\boldsymbol{X}_i) = \boldsymbol{0}$ and $\lambda_{\max}\big(\mathbb{E}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}}|\boldsymbol{x}_i)\big) \leq \sigma_0^2$ (almost surely).

Proposition 1.3.5. Assume Conditions (C2) and (C3) hold. For any $\sigma \geq \sigma_0$ and $z > 0$, choose $\tau = \sigma\sqrt{n/(z+\log d)}$ with $d = d_1 + d_2$. Then, it holds with probability at least $1 - e^{-z}$ that

$$\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2 \leq C\nu_0\sigma\sqrt{\frac{d(z+\log d)}{n}},$$

where $C > 0$ is a universal constant.

Proposition 1.3.6. Assume Conditions (C2) and (C3) hold, and let $s, \tau > 0$ and $z > 0$ satisfy

$$\tau \geq \max\left\{4\sigma_0\sqrt{d_2}, 2\nu_0 s\sqrt{2d_1 + 3z + 3\log n}\right\}. \tag{1.20}$$

Then, with probability at least $1 - 2e^{-z}$,

$$\langle \nabla \widehat{L}_\tau(\mathbf{\Theta}) - \nabla \widehat{L}_\tau(\mathbf{\Theta}^*), \mathbf{\Theta} - \mathbf{\Theta}^* \rangle \geq \frac{c_l}{2}\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathrm{F}}^2 \text{ for all } \mathbf{\Theta} \in \mathbf{\Theta}^* + \mathbb{B}(s),$$

provided that $n \gtrsim \nu_0^4 c_l^{-2}(d_1 + z)$.

Theorem 1.3.4. Assume Conditions (C1)-(C3) hold. For any $z > 0$, the robust (approximately) low-rank matrix estimator $\widehat{\mathbf{\Theta}}_{\tau,\lambda}$ defined in (1.8) with $\tau \asymp \sigma_0\sqrt{n/(z+\log d)}$ and $\lambda \asymp \sigma_0\sqrt{d(z+\log d)/n}$ ($d = d_1 + d_2$) satisfies the bounds

$$\|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_{\mathrm{F}} \lesssim \sigma_0^{1-q/2}\sqrt{\rho}\left\{\frac{d(z+\log d)}{n}\right\}^{\frac{1}{2}-\frac{q}{4}}$$

and

$$\|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_* \lesssim \sigma_0^{1-q}\rho\left\{\frac{d(z+\log d)}{n}\right\}^{\frac{1-q}{2}}$$

with probability at least $1 - 3e^{-z}$ as long as

$$n \gtrsim \max\{(\rho/\sigma_0^q)^{2/(4-q)}, 1\} \cdot (d + z + \log n)(z + \log d).$$

Remark 1.3.8. Again, in the exact low-rank case where $\rho = r = \mathrm{rank}(\mathbf{\Theta}^*)$ and $q = 0$, Theorem 1.3.4 shows that for an arbitrary accuracy $\varepsilon > 0$, we have $\|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_{\mathrm{F}} \le \varepsilon$ with an overwhelming probability provided that the number of measurements satisfies

$$n \gtrsim \sigma_0^2 \frac{rd\log d}{\varepsilon^2}. \tag{1.21}$$

This result improves Theorem 5 in Fan, Wang and Zhu (2021) in several aspects. Under the multitask regression model (1.3), they assumed that

$$\lambda_{\max}\big(\mathbb{E}(\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}})\big) \le R < \infty, \qquad \max_{1\le i\le n, 1\le k\le d_2} \mathbb{E}\big\{\mathbb{E}(\varepsilon_{ik}^2|\boldsymbol{x}_i)^k\big\} \le M_k < \infty \text{ for some } k > 1,$$

and for each i, $\varepsilon_{i1}, \ldots, \varepsilon_{id_2}$ are pairwise (conditionally) independent given $\boldsymbol{x}_i$. In contrast, Condition (C3) only assumes bounded variances and allows arbitrary dependency between ε_{ik}'s. For an arbitrary accuracy $\varepsilon > 0$, the truncated/shrinkage matrix estimator $\widetilde{\mathbf{\Theta}}$ proposed by Fan, Wang and Zhu (2021) satisfies $\|\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}} \le \varepsilon$ with high probability provided

$$n \gtrsim (R + M_k^{1/k}) \frac{rd\log d}{\varepsilon^2}. \tag{1.22}$$

Here the term $R + M_k^{1/k}$ can be much larger than σ_0^2 in (1.21). More importantly, as the noise scale σ_0 decays, R stays away from zero because

$$R \ge \lambda_{\max}\big(\mathbb{E}\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}\big) = \lambda_{\max}\big((\mathbf{\Theta}^*)^{\mathrm{T}}\mathbf{\Sigma}\mathbf{\Theta}^* + \mathbb{E}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}}\big) \ge \lambda_{\max}\big((\mathbf{\Theta}^*)^{\mathrm{T}}\mathbf{\Sigma}\mathbf{\Theta}^*\big).$$

1.4 Numerical studies

1.4.1 Finite-sample performance

In this section, we perform simulation studies to assess the finite-sample performance of the nuclear norm penalized adaptive Huber trace regression method (Nuclear-AH) in all three

Table 1.1. Mean relative Frobenius error $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}/\|\mathbf{\Theta}^*\|_{\mathrm{F}}$ (with standard deviations in parentheses), averaged over 500 replications, under the matrix sensing, matrix completion and multitask regression settings.

Matrix Sensing	Normal	t	Pareto
Nuclear-LS	0.227 (0.010)	0.173 (0.052)	0.169 (0.093)
Nuclear-AH	0.227 (0.010)	0.132 (0.008)	0.107 (0.007)
Matrix Completion	**Normal**	**t**	**Pareto**
Nuclear-LS	0.424 (0.021)	0.280 (0.047)	0.315 (0.041)
Nuclear-AH	0.445 (0.022)	0.223 (0.023)	0.252 (0.022)
Multitask Regression	**Normal**	**t**	**Pareto**
Nuclear-LS	0.228 (0.005)	0.213 (0.112)	0.237 (0.181)
Nuclear-AH	0.228 (0.005)	0.148 (0.003)	0.120 (0.003)

problems. As a benchmark, we implement the nuclear norm penalized least squares (Nuclear-LS) estimator also via the LAMM algorithm.

In addition to the regularization parameter λ, the use of an adaptive Huber loss also involves a robustification parameter τ that changes with data scales. We set $\tau = c_\tau \cdot a_{n,d}$ and $\lambda = c_\lambda \cdot b_{n,d}$, where c_τ and c_λ are positive constants that are independent of (n,d) but depend on the noise scale, and $a_{n,d}$ and $b_{n,d}$ are determined by the theoretical results in Section 1.3, as follows:

(a) For matrix sensing, we choose $a_{n,d} = \sqrt{n/d}$ and $b_{n,d} = \sqrt{d/n}$.

(b) For matrix completion, we choose $a_{n,d} = \sqrt{n/(d\log d)}$ and $b_{n,d} = \sqrt{\log(d)/(dn)}$.

(c) For multitask learning, we choose $a_{n,d} = \sqrt{n/\log d}$ and $b_{n,d} = \sqrt{d\log(d)/n}$.

Then we follow the following steps to tune c_τ and c_λ.

(i) First, choose the constant c_λ in the Nuclear-LS method via five-fold CV with the absolute median loss as the criterion. In particular, we use the "one-standard-error" rule, which yields the most parsimonious model within one standard error of the minimum CV error.

(ii) Next, let $\{r_i\}_{i=1}^n$ be the Nuclear-LS residuals with c_λ selected via CV as in Step (i). As

a rule-of-thumb, we set c_τ as the median absolute deviation of $\{r_i\}$, i.e. median$\{|r_i - \text{median}(r_i)|\}/0.6745$.

(iii) With c_τ determined after Step (ii), we choose the constant c_λ in the Nuclear-AH method again via five-fold CV under the one-standard-error rule.

Under the matrix sensing and matrix completion setups, the data $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^n$ are generated from $y_i = \langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle + \varepsilon_i$, where ε_i follows one of the following three distributions: (i) $\mathcal{N}(0, 0.5^2)$—centered normal distribution with standard deviation 0.5 (lighted-tailed and symmetric), (ii) $t_{2.1}/8$—scaled t-distribution with 2.1 degrees of freedom (heavy-tailed and symmetric), and (iii) $\text{Par}(2,1)/8$—scaled Pareto distribution with scale parameter 1 and shape parameter 2 (heavy-tailed and asymmetric). For matrix sensing, we set $(d_1, d_2, n) = (50, 50, 1500)$, $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ is such that $\text{rank}(\boldsymbol{\Theta}^*) = 5$ and all nonzero singular values of $\boldsymbol{\Theta}^*$ are 1, and the design matrix $\boldsymbol{X}_i$ consists of i.i.d. standard normal entries. For matrix completion, we set $(d_1, d_2, n) = (50, 50, 2000)$, $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ is such that $\|\boldsymbol{\Theta}^*\|_{\text{F}} = \sqrt{d_1 d_2}$ and $\text{rank}(\boldsymbol{\Theta}^*) = 5$, and $\boldsymbol{X}_i$ is uniformly sampled from $\{\boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\text{T}}(d_2)\}_{1 \le j \le d_1, 1 \le k \le d_2}$. To implement LAMM, we use the initial estimates $\boldsymbol{\Theta}^{(0)} = \boldsymbol{0}$ and $\boldsymbol{\Theta}^{(0)} = (d_1 d_2/n)\sum_{i=1}^n y_i \boldsymbol{X}_i$, respectively, under the two setups. In the case of multitask regression, the data vectors $\{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}_{i=1}^n$ are generated from $\boldsymbol{y}_i = (\boldsymbol{\Theta}^*)^{\text{T}}\boldsymbol{x}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ is the same as in the matrix sensing setting, $\boldsymbol{x}_i \in \mathbb{R}^{d_1}$ are i.i.d. standard normal and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id_2})^{\text{T}}$ consists of i.i.d. entries following one of the above three errors distributions. In this case we set $d_1 = d_2 = 80$ and $n = 2000$.

Simulation results on the relative Frobenius error $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\text{F}}/\|\boldsymbol{\Theta}^*\|_{\text{F}}$, averaged over 500 repetitions, are presented in Table 1.1. To better demonstrate the robustness property of Nuclear-AH, Figure 1.1 shows the boxplots of (relative) Frobenius errors for the cross-validated Nuclear-LS and Nuclear-AH estimators under three error distributions. We see that Nuclear-LS and Nuclear-AH have almost identical performance when errors have symmetric and light tails, while the latter achieves considerably better performance under all three settings in the presence of heavy-tailed and/or asymmetric errors.

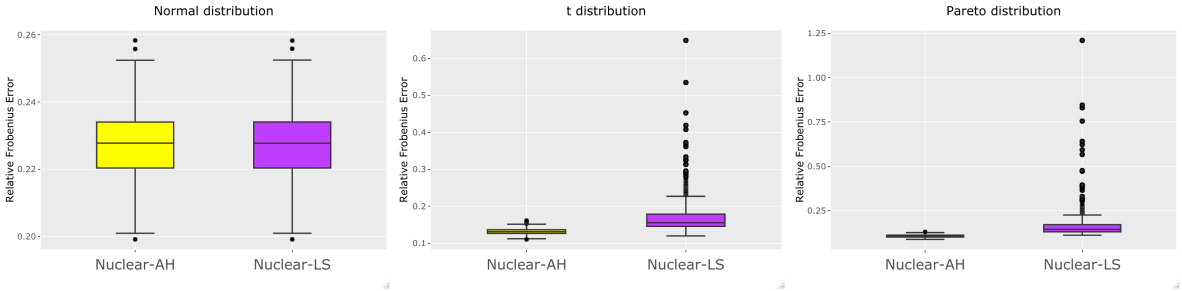
1.4.2 Convergence rate versus noise scale

In this section, we numerically examine the dependence of $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}$ on the noise scale under the matrix sensing and multitask regression settings. Our theoretical results, Theorem 1.3.2 and Theorem 1.3.4, indicate that in the exact low-rank case, $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}$ should be proportional to the noise scale σ, where $\sigma^2 = \mathbb{E}(\varepsilon_i^2)$ or $\sigma^2 = \lambda_{\max}(\mathbb{E}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^{\mathrm{T}}))$. To verify this, given a sequence of σ values ranging from 10^{-3} to 1, we generate ε_i from either $\sigma \cdot \mathscr{N}(0,1)$ or $\sigma \cdot t_{2.1}/16$. The specifications of $\boldsymbol{\Theta}^*$ and $\boldsymbol{X}_i$ or $\boldsymbol{x}_i$ are the same as in Section 1.4.1.

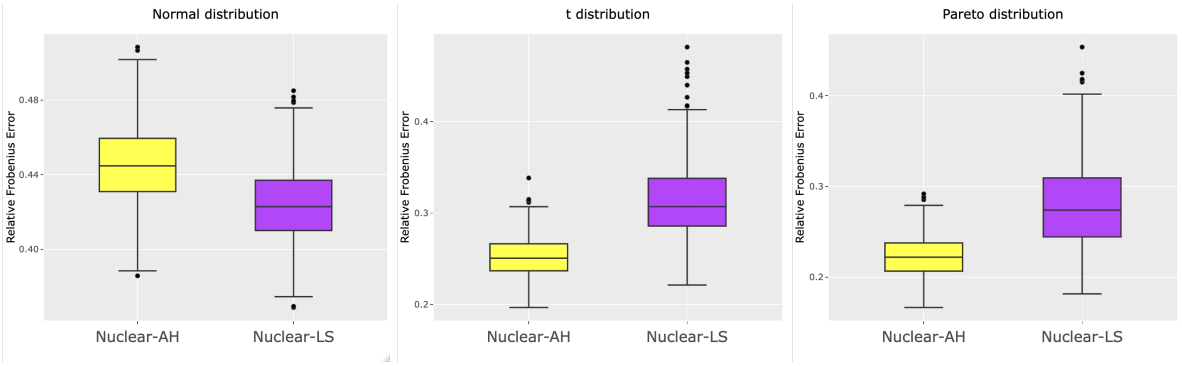
Under the matrix sensing setting, we set $(d_1, d_2, n) = (50, 50, 2000)$ and choose $\tau = 2\sigma\sqrt{n/d}$ and $\lambda = \sigma\sqrt{d/n}$ with $d = d_1 + d_2$. For multitask regression, we set $(d_1, d_2, n) = (100, 100, 3000)$ and choose $\tau = \sigma\sqrt{n/\log d}$ and $\lambda = 0.5\sigma\sqrt{d\log(d)/n}$. Figure 1.2 shows the plots of the Frobenius error versus noise scale, based on 200 replications, under these two settings and two error distributions. Consistent with the predictions of Theorems 1.3.2 and 1.3.4, we observe a nearly perfect linear growth in all four plots.

1.5 Acknowledgements

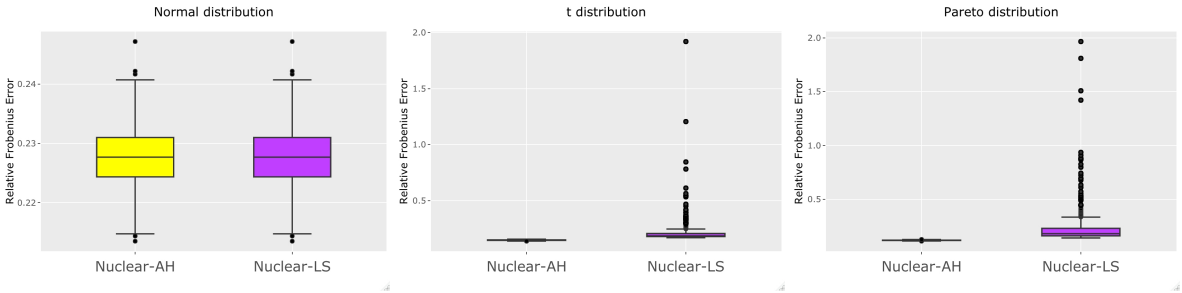
This chapter, in full, is a reprint of the material in the paper "Low-rank matrix recovery under heavy-tailed errors", Yu, Myeonghun, Sun, Qiang and Zhou, Wen-Xin. The paper has been published on *Bernoulli*, **30**, 2326–2345. The dissertation author was the primary investigator and author of this paper.



(a) Matrix sensing setting with $(d_1, d_2, n) = (50, 50, 1500)$

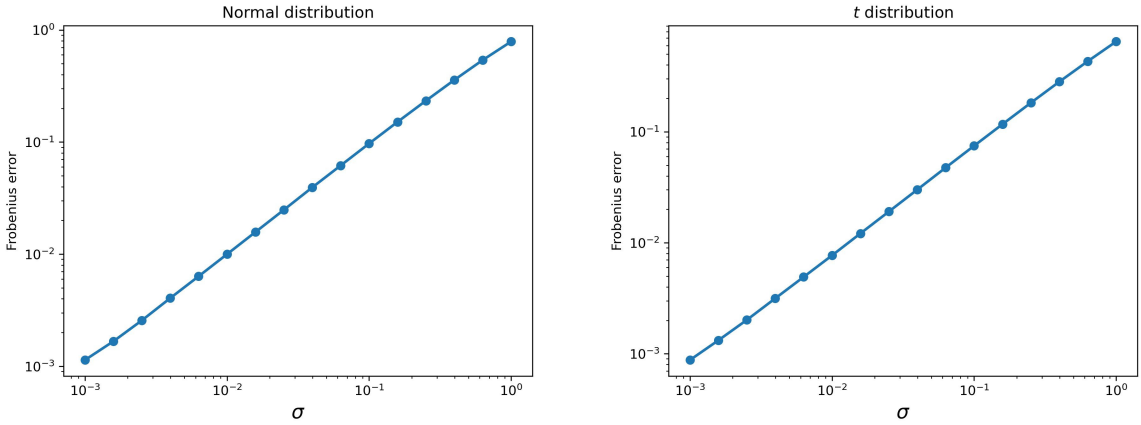


(b) Matrix completion setting with $(d_1, d_2, n) = (50, 50, 2000)$

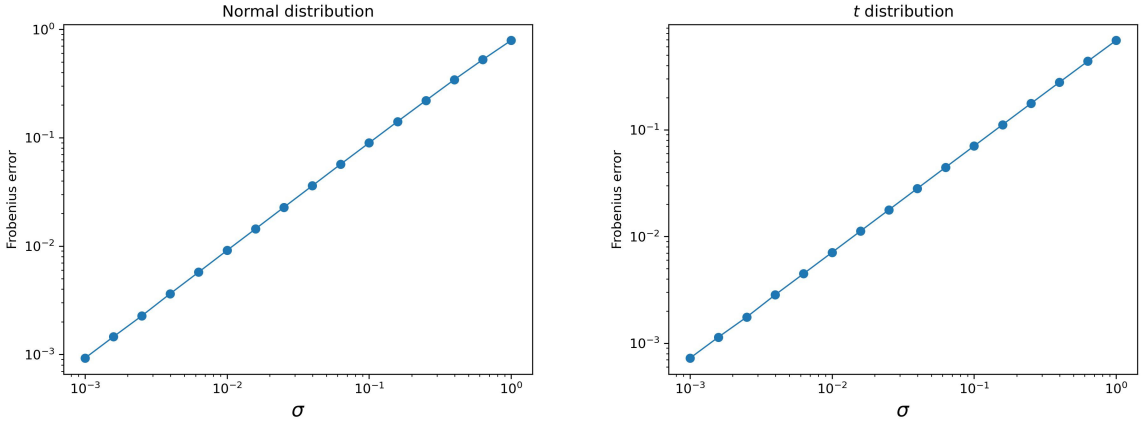


(c) Multitask regression setting with $(d_1, d_2, n) = (80, 80, 2000)$

Figure 1.1. Boxplots of relative Frobenius errors $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} / \|\boldsymbol{\Theta}^*\|_{\mathrm{F}}$ (based on 500 repetitions) for the Nuclear-AH and Nuclear-LS estimators under the matrix sensing, matrix completion and multitask regression settings.



(a) Matrix sensing setting with $(d_1, d_2, n) = (50, 50, 2000)$



(b) Multitask regression setting with $(d_1, d_2, n) = (50, 50, 2000)$

Figure 1.2. Plots of Frobenius error $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}$ versus noise scale based on 200 simulations under the matrix sensing and multitask regression settings.

Chapter 2

Gaussian differentially private robust mean estimation and inference

2.1 Introduction

We consider the problem of estimating the mean of a random vector $\boldsymbol{x} \in \mathbb{R}^d$ based on independent and identically distributed (i.i.d.) samples $\{\boldsymbol{x}_i\}_{i=1}^n$. When the data are generated from heavy-tailed distributions and/or contaminated with outliers, this problem, known as robust mean estimation, has received a lot of attention recently in both statistical and machine learning communities; see, for example, Catoni (2012); Bubeck, Cesa-Bianchi and Lugosi (2013); Minsker (2015); Devroye *et al.* (2016); Chen, Gao and Ren (2018); Lugosi and Mendelson (2019a); Hopkins (2022); Hopkins, Li and Zhang (2020); Lugosi and Mendelson (2021); Depersin and Lecué (2022a,b); Mathieu (2022) for an unavoidably incomplete overview. For a more thorough review of robust mean estimation and beyond, we refer to the survey articles Diakonikolas and Kane (2019) and Lugosi and Mendelson (2019b).

It is well-known that the sample/empirical mean estimator has desired tail behaviors when the distribution of $\boldsymbol{x}$ is light-tailed, but its performance deteriorates quickly and becomes sub-optimal for heavy-tailed distributions. For example, for a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, the following deviation bound of the sample mean is optimal (Catoni, 2012): for any $z \geq 0$, $\|\bar{\boldsymbol{x}}_n - \boldsymbol{\mu}\|_2 \leq \sqrt{\mathrm{tr}(\Sigma)/n} + \sqrt{2\|\Sigma\|_2 \cdot z/n}$ with probability at least $1 - e^{-z}$, where $\bar{\boldsymbol{x}}_n = (1/n)\sum_{i=1}^n \boldsymbol{x}_i$. The worst-case analysis in (Catoni, 2012) shows that the deviations

of the sample estimate significantly increase when the sample distribution is far from being Gaussian. Over the past decade, significant effort has been dedicated to developing robust mean estimators, both univariate and multivariate, that offer optimal Gaussian-type deviation bounds, as demonstrated above, commonly referred to as sub-Gaussian deviation bounds. Although certain estimators, such as the median-of-means tournaments (Lugosi and Mendelson, 2019a) and the trimmed mean estimator (Lugosi and Mendelson, 2021), are capable of achieving the sharp concentration bound under the bounded second moment condition, most of them are not computationally feasible. Some recent works such as Hopkins (2022) and Depersin and Lecué (2022a) have proposed polynomial-time mean estimation algorithms that achieve sub-Gaussian rates. However, implementing these algorithms in practice remains a significant challenge. In contrast, Huber's M-estimator and its variants considered by Mathieu (2022) are computationally more efficient as they are directly defined as minima of convex optimization problems. It is worth noting that the M-estimation approach comes with a minor caveat. Specifically, Proposition 2 of Mathieu (2022) demonstrates that Huber's M-estimator can attain the sub-Gaussian deviation bound within a limited range of z when the distribution of $\boldsymbol{x}$ has finite q-th moment with $q > 2$. However, when $\boldsymbol{x}$ only exhibits finite variance, the estimator attains the sub-optimal deviation bound; see also Remark 2.2.1.

While most of the aforementioned results solely focus on statistical properties without taking into account the potentially sensitive information contained in the data, there has been an increasing demand for data privacy guarantees in statistical methods during the last decade. Differential privacy (DP), arguably the first widely accepted rigorous definition of data privacy, was introduced in Dwork *et al.* (2006b) and has since gained widespread acceptance and success. Informally, a mechanism is said to be differentially private if its distribution over outputs is insensitive to the change of only one datum. Gaussian differential privacy (GDP) Dong, Roth and Su (2022) is an attractive variant of DP, especially for statisticians, due to its neat hypothesis testing interpretation. The study of mean estimation with differential privacy is mostly limited in the computer science literature. For example, Bun and Steinke (2019); Kamath *et al.* (2019);

Karwa and Vadhan (2018) considered optimal private mean estimation in terms of sample complexity under different differential privacy frameworks. Another work in statistics literature Cai, Wang and Zhang (2021) proved minimax optimal mean estimation under squared error loss under DP. However, these results all depend on the assumption that the underlying distribution is sub-Gaussian.

Recently, the problem of private robust mean estimation under heavy-tailed distributions has gained increasing interest in the literature. For instance, based on pairwise comparisons, Kamath, Singhal and Ullman (2020) introduced an algorithm for private mean estimation under concentrate, pure, and (ε, δ)-DP when a distribution has a bounded q-th moment for $q \geq 2$. Additionally, Liu et al. (2021) proposed a private iterative filtering-based algorithm designed to estimate the mean vector of heavy-tailed distributions under (ε, δ)-DP, even when the data is corrupted by arbitrary outliers. Hopkins, Kamath and Majid (2022) utilized the sum-of-squares method to design private algorithms that are robust to heavy-tailed distribution and arbitrary outliers under pure DP. However, most of the proposed methods, although achieved by polynomial-time algorithms, are still not as computationally tractable as those based on convex optimization.

Despite the growing interest in developing robust non-private and private mean estimators with sub-Gaussian deviation bounds, existing results have mainly focused on providing concentration bounds. Robust inference with heavy-tailed data, however, has often been neglected. Due to the high complexity of existing robust mean algorithms, it is challenging to track the limiting distributions of the resulting estimators. Constructing differentially private confidence sets presents an even greater challenge since it involves accounting for the additional noise needed to guarantee privacy.

The main goal of this paper is to develop an easy-to-implement GDP robust mean estimator and construct privacy-preserving confidence intervals for heavy-tailed data. To achieve robustness, we adopt a Huber (robust) loss function with a diverging robustification parameter τ (Catoni, 2012; Mathieu, 2022). On the other hand, data privacy is typically guaranteed by

randomly perturbing the output of non-private algorithms (Dwork *et al.*, 2006b; McSherry and Talwar, 2007). In particular, to privately release a non-private Huber-type robust estimator, inspired by Song et al. (2013); Bassily, Smith and Thakurta (2014), we take a noisy optimization approach by adding Gaussian noises in each iteration of the gradient descent method. This noisy gradient decent procedure guarantees that the desired privacy level can still be met along a sequence of outputs by carefully choosing the scale of the added noise. To make valid inferences, one needs to leverage the distributional properties of the resulting private robust mean estimator. Existing concentration/deviation bounds such as those in Mathieu (2022) do not allow us to achieve this goal, even for the non-private Huber-type mean estimator. To this end, we first provide a refined non-asymptotic analysis and establish Bahadur representation of the non-private Huber-type mean estimator, which paves the road for the more challenging inference problem of its private counterpart. In constructing the private confidence intervals, we show that the scale of the privacy-inducing noise critically depends on the robustification parameter τ, which also balances the bias and robustness of the non-private Huber-type estimator. The cost of privacy is further revealed by our different choices of τ and the resulting deviation bounds together with Gaussian approximation bounds for private and non-private robust mean estimators.

Our contributions are mainly three-fold: (a) *A comprehensive analysis of a Huber-type robust mean estimator.* While a concentration study already appeared in the literature for robust M-estimators of locations, our first contribution is to go beyond deviation analysis and establish Bahadur representation and (uniform) Gaussian approximation, which are key ingredients to construct both non-private and private robust confidence intervals. Notably, our analysis of the Berry-Esseen bound reveals that the choice of robustification parameter τ that leads to the smallest concentration bound results in a sub-optimal Berry-Esseen bound; see Remark 2.2.2 for details. It is also worth mentioning that even for the concentration bounds with bounded second moment assumption, our result still slightly improves that in Proposition 2 of Mathieu (2022) due to using a different analysis. (b) *Noisy gradient descent of Huber mean estimator.* Our second contribution is to privatize the Huber-type robust estimator via a noisy gradient descent

algorithm. We provide a complete finite-sample convergence analysis, demonstrating that private iterates converge linearly to a ball centered at the non-private Huber estimator with a radius comparable to the noise added in each step. Different from most existing methods, one novelty is that the privacy-inducing noise level critically depends on the robustification parameter τ, which in turn controls the bias and robustness. In contrast to the non-private counterpart, the trade-off between bias, robustness and privacy leads to a choice of τ explicitly depending on the privacy level. Consequently, we show the cost of privacy in a deviation bound for our private robust mean estimator and demonstrate its optimality in terms of the dependence on privacy and moment conditions for some scenarios. In particular, the cost of privacy of our proposed estimator with an appropriate choice of τ achieves the minimax optimal bound under the finite second moment, and the estimator has the same cost as in Kamath, Singhal and Ullman (2020), which is the smallest one in the literature under higher-moment assumptions. (c) *Private robust confidence intervals.* The last but not least contribution is to construct both non-private and private robust confidence intervals for linear projections of the mean under a bounded fourth moment condition. We allow increasing dimension d due to the new Gaussian approximation results. The novel construction of private robust confidence intervals is based on a noisy Studentized statistic. In particular, to guarantee the privacy of the confidence interval, besides the private Huber-type mean estimator, we further employ a robust and private estimator of the covariance.

Other related literature. In the statistics literature, a series of works are devoted to developing differentially private approaches for statistical estimation with a focus on optimal rates of convergence, including Wasserman and Zhou (2010); Barber and Duchi (2014); Duchi, Jordan and Wainwright (2022); Cai, Wang and Zhang (2020, 2021); Rohde and Steinberger (2020); Wang, Kifer and Lee (2019); Avella-Medina (2021); Avella-Medina, Bradshaw and Loh (2023). For example, under the local differential privacy, a slightly stronger notion of DP, Wasserman and Zhou (2010) revealed that existing private mechanisms lead to slower rates than the minimax rates, and (Duchi, Jordan and Wainwright, 2022; Rohde and Steinberger, 2020) further derived new minimax rates and corresponding private algorithms for several models.

(Cai, Wang and Zhang, 2020, 2021) considered minimax optimality of mean estimation and generalized linear regression with given differential privacy (DP) constraint under both the low-dimensional and sparse high-dimensional settings. The studies in hypothesis testing and confidence intervals with differential privacy are still limited in the statistical community. The most relevant work to the current paper is Avella-Medina, Bradshaw and Loh (2023), which considered optimization-based approaches for Gaussian differentially private M-estimators. In particular, parametric inference problems are tackled by constructing private variance estimators. While their general noisy gradient descent method can be applied for our robust mean estimation, the inference analysis and results do not allow increasing dimensional settings. In contrast, our newly established Gaussian approximation results together with a careful global convergence analysis of the noisy optimization reveal the critical role of the robustification parameter, which makes the inference under growing dimensions possible.

The rest of the paper is structured as follows. We first revisit the non-private robust mean estimation problem under heavy-tailed distributions in Section 2.2. New concentration bounds and normal approximation results are established for the proposed Huber estimator to conduct robust inference, including constructing confidence intervals and sets in this section. Section 2.3 introduces the basic background of Gaussian differential privacy and presents our private Huber mean estimator via a noisy gradient descent algorithm with finite-sample convergence analysis. New approaches for constructing private robust confidence intervals are further presented in Section 2.3. Section 2.4 presents the numerical studies that evaluate the performance of the proposed robust mean estimators, both non-private and private. Additionally, a data-driven approach is proposed to choose the robustification parameter. Some proofs of theorems in Section 2.2 are given in Appendix 2.5. The extension of our construction of private robust estimators to other notions of differential privacy, a detailed description of the numerical algorithm for computing private robust estimators, and remaining proofs for theoretical results are relegated to the Supplementary Material Yu, Ren and Zhou (2023).

NOTATION. The following notations will be used throughout this paper. For every integer $d \geq 1$, we use $\mathbb{R}^d$ to denote the d-dimensional Euclidean space. For any vector $\boldsymbol{u} = (u_1, \ldots, u_d) \in \mathbb{R}^d$, we use $\|\boldsymbol{u}\|_p (1 \leq p \leq \infty)$ to denote its ℓ_p-norm in $\mathbb{R}^p$: $\|\boldsymbol{u}\|_p = (\sum_{j=1}^d |u_j|^p)^{1/p}$ and $\|\boldsymbol{u}\|_\infty = \max_{1 \leq j \leq d} |u_j|$. The unit $(d-1)$-sphere $\mathbb{S}^{d-1}$ is defined as $\mathbb{S}^{d-1} = \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_2 = 1\}$. We write $a \lesssim b$ if there exists an absolute constant $C > 0$ such that $a \leq Cb$, and $a \gtrsim b$ if $b \lesssim a$. Moreover, we write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$.

2.2 Robust mean estimation and inference via Huber loss

In this section, we consider robust (multivariate) mean estimation using Huber loss minimization. A more general version of this approach was proposed by Mathieu (2022), in which concentration bounds are established. The idea of using a robust loss function with a diverging robustification parameter (as a function of sample size) dates back to Catoni (2012), and has also been employed in regression settings (Fan, Li and Wang, 2017; Zhou et al., 2018). In Section 2.2.1, we first provide a concentration bound for the Huber mean estimator, denoted by $\widehat{\boldsymbol{\mu}}_\tau$ parameterized by $\tau > 0$, based on a different technical argument compared to that employed in Mathieu (2022). Next, we provide a non-asymptotic Bahadur representation result, indicating that $\sqrt{n}(\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu})$ can be approximated by a linear statistic with higher-order remainders. Based on this result, in Section 2.2.2 we establish several normal approximation results (through Berry-Esseen-type bounds) for the proposed robust estimator, which pave the way for constructing robust confidence intervals under heavy-tailed distributions.

Throughout, let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be independent observations from a random vector $\boldsymbol{x} \in \mathbb{R}^d$ with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^{\mathrm{T}}$ and covariance matrix $\Sigma = (\sigma_{kl})_{1 \leq k,l \leq d}$, both assumed to be unknown.

2.2.1 A concentration study of Huber mean estimator

Given $\tau > 0$, define the loss function $\rho_\tau(u) = \tau^2 \rho(u/\tau)$ for some continuously differentiable convex function $\rho : \mathbb{R} \to [0, \infty)$. Assume that $\psi(u) = \rho'(u)$ is Lipschitz continuous, concave, and differentiable almost everywhere on $\mathbb{R}_+$. Mathieu (2022) provided a concentration

study of the following M-estimator:

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_\tau \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\{ \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\big(\|\boldsymbol{x}_i - \boldsymbol{\theta}\|_2\big) \right\}. \tag{2.1}$$

Let $\psi_\tau(u) = \rho_\tau'(u) = \tau\psi(u/\tau)$ be the score function. By the convexity of $\rho_\tau(\cdot)$ and hence of $\widehat{\mathscr{L}_\tau}(\cdot)$, the M-estimator $\widehat{\boldsymbol{\mu}}$ can be equivalently defined as the solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\theta}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\theta}\|_2}(\boldsymbol{x}_i - \boldsymbol{\theta}) = \mathbf{0}.$$

In particular, Mathieu (2022) considered three robust mean estimators that are determined by their corresponding score functions, which are

(i) (Huber's score) $\psi(u) = u\mathbb{1}(|u| \le 1) + \operatorname{sign}(u)\mathbb{1}(|u| > 1)$;

(ii) (Catoni's score) $\psi(u) = \log(1 + u + u^2/2)\mathbb{1}(u \ge 0) - \log(1 - u + u^2/2)\mathbb{1}(u < 0)$;

(iii) (Polynomial score) For $p \ge 1$, $\psi(u) = \frac{u}{1+u^{1-1/p}}\mathbb{1}(u \ge 0) - \frac{u}{1+(-u)^{1-1/p}}\mathbb{1}(u < 0)$.

As demonstrated in Mathieu (2022), these three robust estimators exhibit similar theoretical and numerical performance. Therefore, we restrict attention to Huber's estimator (Huber, 1964). The Huber loss is defined as

$$\rho(u) = \min(u^2/2, |u| - 1/2),$$

with its score function listed in (i) above. A variety of smoothed Huber loss functions have been discussed in the robust statistics literature (Hampel, Hennig and Ronchetti, 2011). See, for example, Examples 1 and 2 in Avella-Medina, Bradshaw and Loh (2023).

Theorem 2.2.1 below provides a concentration bound for the (multivariate) Huber mean estimator $\widehat{\boldsymbol{\mu}}_\tau$ with a sufficiently large τ, explicitly dependent on the robustification bias. Through-

out the rest, we write

$$\bar{\lambda} = \|\Sigma\|_2 := \max_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \|\Sigma\boldsymbol{u}\|_2, \ \ \underline{\lambda} = \min_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \boldsymbol{u}^{\mathrm{T}}\Sigma\boldsymbol{u} \ \ \text{and} \ \ \mathrm{r}(\Sigma) = \mathrm{tr}(\Sigma)/\|\Sigma\|_2$$

as the largest eigenvalue, smallest eigenvalue, and effective rank of the covariance matrix Σ, respectively.

Theorem 2.2.1. Assume that the random vector $\boldsymbol{x} \in \mathbb{R}^d$ has mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. For any $z > 0$, the Huber mean estimator $\widehat{\boldsymbol{\mu}}_\tau$ given in (2.1) with $\tau \gtrsim \sqrt{\mathrm{tr}(\Sigma)}$ satisfies the bound

$$\|\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\|_2 \lesssim \bar{\lambda}^{1/2}\sqrt{\frac{\mathrm{r}(\Sigma)+z}{n}} + \frac{\tau z}{n} + b_\tau \tag{2.2}$$

with probability at least $1-2e^{-z}$ as long as $n \gtrsim \mathrm{r}(\Sigma)+z$, where

$$b_\tau := \left\| \mathbb{E}\left\{ \frac{\psi_\tau(\|\boldsymbol{x}-\boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2}(\boldsymbol{x}-\boldsymbol{\mu}) \right\} \right\|_2 \le \frac{\sqrt{\bar{\lambda}\,\mathrm{tr}(\Sigma)}}{\tau}. \tag{2.3}$$

We refer to b_τ as the robustification bias. When $\tau = \infty$, it is easy to see that $b_\infty = 0$; in general, $b_\tau > 0$ for any fixed $\tau > 0$ unless the distribution of $\boldsymbol{x}$ is symmetric around $\boldsymbol{\mu}$.

To determine the optimal robustification parameter τ that minimizes the upper bound (2.2) under higher moment assumptions, we next derive a bound for b_τ. Before doing so, we first introduce some additional notations. Assuming that $m_q := \mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^q$ is finite for some $q \ge 2$, we define

$$\nu_q = \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \frac{\mathbb{E}|\langle \boldsymbol{x}-\boldsymbol{\mu}, \boldsymbol{u}\rangle|^q}{(\mathbb{E}\langle \boldsymbol{x}-\boldsymbol{\mu}, \boldsymbol{u}\rangle^2)^{q/2}} \ \ \text{and} \ \ \kappa_q = \max_{1\le k\le d} \frac{\mathbb{E}|x_k-\mu_k|^q}{\{\mathbb{E}(x_k-\mu_k)^2\}^{q/2}}. \tag{2.4}$$

In particular, ν_4 and κ_4 denote, respectively, the supremum of the kurtosises of all linear combinations of $\boldsymbol{x}$ and the maximum of the kurtosises of all coordinates of $\boldsymbol{x}$. These quantities

characterize the degree of skewness of the random vector $\boldsymbol{x}$. It is easy to see that $\nu_4 \geq \kappa_4 > 1$ if $\boldsymbol{x}$ is non-degenerate, and $\nu_4 = \kappa_4 = 3$ when $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Also, note that when $m_q < \infty$ for $q \geq 2$, we have $m_q^{1/q} \geq \mathrm{tr}(\Sigma)^{1/2}$ by Jensen's inequality, and Hölder's inequality yields

$$\begin{aligned} m_q &= \mathbb{E}\left\{ \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^{q-2} \sum_{k=1}^{d} (x_k - \mu_k)^2 \right\} \\ &\leq \sum_{k=1}^{d} (\mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^q)^{1-2/q} (\mathbb{E}|x_k - \mu_k|^q)^{2/q} \leq m_q^{1-2/q} \cdot \kappa_q^{2/q} \mathrm{tr}(\Sigma), \end{aligned} \tag{2.5}$$

so that $\mathrm{tr}(\Sigma)^{1/2} \leq m_q^{1/q} \leq \kappa_q^{1/q} \mathrm{tr}(\Sigma)^{1/2}$. With the notation, we now present the bound of the bias b_τ.

Lemma 2.2.1. Assume that there exists some $q \geq 2$ such that $m_q = \mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^q$ is finite. Then, the bias term b_τ satisfies

$$b_\tau \leq \min\left\{ \nu_q^{1/q} \frac{\bar{\lambda}^{1/2} m_q^{1-1/q}}{\tau^{q-1}}, \frac{m_q}{\tau^{q-1}} \right\}.$$

Remark 2.2.1. By combining Lemma 2.2.1 and Theorem 2.2.1, we can choose τ that minimizes $b_\tau + \tau z/n$. For instance, when the variance exists ($q = 2$), the optimal choice for τ is $\tau \asymp \bar{\lambda}^{1/4} \mathrm{tr}(\Sigma)^{1/4} (n/z)^{1/2}$, which leads to the bound

$$\|\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\|_2 \lesssim \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \bar{\lambda}^{1/2} \mathrm{r}(\Sigma)^{1/4} \sqrt{\frac{z}{n}} \tag{2.6}$$

with probability at least $1 - 2e^{-z}$ as long as $n \gtrsim \max\{\mathrm{r}(\Sigma), \mathrm{r}(\Sigma)^{1/2} z\}$. For heavy-tailed data without adversarial corruption, the above bound slightly improves that in Proposition 2 of Mathieu (2022) with $q = 2$ and $\varepsilon_n = 0$. In detail, Proposition 2 of Mathieu (2022) establishes that the Huber estimator $\widehat{\boldsymbol{\mu}}_\tau$ with $\tau \asymp \mathrm{tr}(\Sigma)^{1/2} (n/z)^{1/2}$ satisfies

$$\|\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\|_2 \lesssim \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \bar{\lambda}^{1/2} \sqrt{\frac{z}{n}} + \bar{\lambda}^{1/2} \mathrm{r}(\Sigma)^{1/2} \sqrt{\frac{z}{n}}$$

with probability at least $1-4e^{-z}-e^{-n/32}$ as long as $n \gtrsim z$ in our notations. Consequently, our derived bound improves upon the multiplicative factor of $\mathrm{r}(\Sigma)^{1/2}$ in the bound of Mathieu (2022), refining it to $\mathrm{r}(\Sigma)^{1/4}$.

Yet, the deviation bound (2.6) is still sub-optimal in terms of its dependence on $\bar{\lambda}$, $\mathrm{tr}(\Sigma)$ and z. Specifically, it includes an extra multiplicative factor of $\mathrm{r}(\Sigma)^{1/4}$, compared to the optimal Gaussian concentration bound. However, the main advantage of Huber loss minimization is threefold: (i) the estimator is defined as the solution to a convex optimization problem, for which the objective function is also locally strongly convex; (ii) the asymptotic distribution is easily tractable, which significantly facilitates statistical inference; (iii) via noisy gradient descent, we can construct differentially private robust mean estimator and the correspondent confident intervals/sets as discussed in Section 2.3.

Moreover, the Huber estimator $\widehat{\boldsymbol{\mu}}$ attains the optimal concentration bound as long as z is small under higher-moment assumptions. Specifically, when $m_q < \infty$ for $q > 2$, we can choose $\tau \asymp m_q^{1/q}(n/z)^{1/q}$ to obtain a tighter concentration bound given by

$$\|\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\|_2 \lesssim \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + m_q^{1/q}\left(\frac{z}{n}\right)^{1-1/q} \tag{2.7}$$

with probability at least $1-2e^{-z}$. Applying the inequality $m_q^{1/q} \le \kappa_q^{1/q}\mathrm{tr}(\Sigma)^{1/2}$, we can see that $\widehat{\boldsymbol{\mu}}_\tau$ satisfies the optimal Gaussian concentration bound provided that $z = O(n^{(q-2)/(2q-2)} + n \cdot \mathrm{r}(\Sigma)^{-q/(q-2)})$. In this regime, the Huber estimator $\widehat{\boldsymbol{\mu}}_\tau$ attains the optimal deviation bound.

Theorem 2.2.1 is restricted to establishing concentration/deviation bounds and thus falls short in addressing the distributional characteristics of $\widehat{\boldsymbol{\mu}}_\tau$. However, the latter is the cornerstone for statistical inference. To fill this gap, we further establish a non-asymptotic Bahadur representation result for $\widehat{\boldsymbol{\mu}}_\tau$, which is the key to deriving Gaussian approximation results with explicit error bounds.

Theorem 2.2.2. Assume that there exists some $q \ge 2$ such that $m_q = \mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^q < \infty$. Given $t > 0$,

let the sample size satisfy $n \gtrsim \mathrm{r}(\Sigma) + z$. Then, the Huber mean estimator $\widehat{\boldsymbol{\mu}}_\tau$ with $\tau \gtrsim \sqrt{\mathrm{tr}(\Sigma)}$ satisfies

$$\left\| \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu} - \frac{1}{n}\sum_{i=1}^{n} \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}) \right\|_2 \lesssim \left\{ \bar{\lambda}^{1/2} \sqrt{\frac{\mathrm{r}(\Sigma) + z}{n}} + \frac{\tau z}{n} + b_\tau \right\} \left(\frac{m_q}{\tau^q} + \sqrt{\frac{z}{n}} \right) \tag{2.8}$$

with probability at least $1 - 3e^{-z}$, where $\psi_\tau(u) = \tau\psi(u/\tau)$ and b_τ is defined in (2.3).

Theorem 2.2.2 shows that with high probability, $\sqrt{n}(\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu})$ is first-order equivalent to the linear term

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}),$$

which determines the asymptotic distribution of $\widehat{\boldsymbol{\mu}}_\tau$. Based on the Bahadur representation (2.8), in Section 2.2.2, we establish several Gaussian approximation results for $\widehat{\mu}_\tau$ under the bounded third or fourth moment condition. In particular, the boundedness of the fourth moment is crucial for robust covariance estimation (Minsker, 2018; Mendelson and Zhivotovskiy, 2020).

2.2.2 Gaussian approximations

In this section, we present two Gaussian approximation results for the Huber mean estimator $\widehat{\boldsymbol{\mu}}_\tau$ under the bounded third or fourth moment condition. The dimension d is allowed to grow with the sample size n and enters the Gaussian approximation error bounds through the moment parameter $m_q = \mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^q$ for $q \geq 3$.

Theorem 2.2.3 below provides a Berry-Esseen bound for all (deterministic) linear combinations of $\widehat{\boldsymbol{\mu}}_\tau$, from which the asymptotic normality immediately follows.

Theorem 2.2.3. Assume that $m_q = \mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^q < \infty$ for some $q \geq 3$, and let the sample size satisfy $n \gtrsim \mathrm{r}(\Sigma) + \log n$. Then, the Huber mean estimator $\widehat{\boldsymbol{\mu}}_\tau$ with $\tau \asymp m_q^{1/q}(n/\log n)^\gamma$ for some

$\gamma \in [1/(q-1), 1/2]$ satisfies

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d, x\in\mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}\langle \boldsymbol{u}/\|\boldsymbol{u}\|_\Sigma, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle \le x\right) - \Phi(x)\right| \lesssim \frac{m_q^{1/q}}{\underline{\lambda}^{1/2}} \frac{\log n}{\sqrt{n}} + \nu_q^{2/q} \left(\frac{\log n}{n}\right)^{(q-2)/(q-1)} + \frac{\nu_3}{\sqrt{n}}, \tag{2.9}$$

where $\|\boldsymbol{u}\|_\Sigma^2 := \boldsymbol{u}^{\mathrm{T}}\Sigma\boldsymbol{u}$ and $\Phi(x)$ is the cumulative distribution function of $\mathcal{N}(0,1)$.

Remark 2.2.2. In Theorem 2.2.3, we require $\boldsymbol{x}$ to have at least the finite third moment so that the upper bound in (2.9) depends on n through $n^{-1/2}$. Instead, if $m_{2+\iota} = \mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^{2+\iota} < \infty$ for some $0 < \iota < 1$, the dependence on n can at best be $n^{-\iota/2}$; see Heyde (1967) for details. To achieve the optimal $n^{-1/2}$-rate, Theorem 2.2.3 shows that the choice of τ becomes more flexible as higher-order moments are bounded. It is worth noting that the choice $\tau \asymp m_q^{1/q}(n/\log n)^\gamma$ with $\gamma \in [1/(q-1), 1/2]$ for Gaussian approximation does not lead to the smallest concentration bound, as shown in (2.7) with a choice $\tau \asymp m_q^{1/q}(n/\log n)^{1/q}$. Yet, for this choice of τ, we will obtain an $n^{-1/2+1/q}$-rate for the Berry-Esseen bound.

When the q-th moment ($q \ge 3$) is finite, the two parameters ν_q and κ_q defined in (2.4) are essentially dimension-free. Using the inequality $m_q^{1/q} \le \kappa_q^{1/q}\mathrm{tr}(\Sigma)^{1/2}$ from (2.5), we can substitute this bound into (2.9) to obtain a further bound for the first term on the right-hand side:

$$\kappa_q^{1/q}(\bar{\lambda}/\underline{\lambda})^{1/2}(\log n)\sqrt{\frac{\mathrm{r}(\Sigma)}{n}}.$$

From an asymptotic view, with two dimension-free parameters ν_q and κ_q defined in (2.4) and a bounded condition number of Σ, this shows that any linear combination of the coordinates of $\sqrt{n}(\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu})$ converges in distribution to the correspondent linear combination of $\mathcal{N}(\mathbf{0}, \Sigma)$ as $n \to \infty$ under the growth condition $\mathrm{r}(\Sigma)\log^2(n) = o(n)$ as $n \to \infty$. Since $\mathrm{r}(\Sigma) \le d$, a sufficient condition on the dimension is $d\log^2(n) = o(n)$.

To construct confidence intervals/sets based on the above result, we also need to robustly estimate the variance $\|\boldsymbol{u}\|_\Sigma^2 = \boldsymbol{u}^{\mathrm{T}}\Sigma\boldsymbol{u}$, or the covariance matrix Σ. To this end, we consider a

U-type robust covariance estimator proposed and studied by Fan *et al.* (2019) and Ke et al. (2019). Given a robustification parameter $\xi > 0$, the U-type covariance estimator $\widehat{\Sigma}_\xi$ is defined as

$$\widehat{\Sigma}_\xi = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \psi_\xi \left(\frac{1}{2} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \right) \frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}, \tag{2.10}$$

where $\psi_\xi(t) = \xi \psi(t/\xi)$. By choosing $\delta = e^{-z}$ in Theorem 3.2 of Ke et al. (2019) with a suitably chosen ξ, the following proposition provides an exponential-type deviation bound for $\widehat{\Sigma}_\xi$ under a bounded fourth moment condition.

Proposition 2.2.1 (Theorem 3.2 in Ke et al. (2019))**.** Assume $\boldsymbol{x} \in \mathbb{R}^d$ has bounded fourth moment, and write

$$v_0^2 := \frac{1}{4} \left\| \mathbb{E}\{(\boldsymbol{x}_1 - \boldsymbol{x}_2)(\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\mathrm{T}}\}^2 \right\|_2. \tag{2.11}$$

Let $n_0 = \lfloor n/2 \rfloor$ be the largest integer not exceeding $n/2$. For any $z > 0$, the U-type covariance estimator $\widehat{\Sigma}_\xi$ defined in (2.10) with $\xi = v_0 \sqrt{n_0 / \{\log(2d) + z\}}$ satisfies

$$\left\| \widehat{\Sigma}_\xi - \Sigma \right\|_2 \le 2 v_0 \sqrt{\frac{\log(2d) + z}{n_0}}$$

with probability at least $1 - e^{-z}$.

Remark 2.2.3. To compute $\widehat{\Sigma}_\xi$, the major barrier is due to the U-statistics structure of (2.10), in which the sum consists of $O(n^2)$ terms. Clémençon, Bellet and Colin (2016) proposed a resampling technique named incomplete U-statistics, which reduces the computation complexity to $O(n)$. Alternatively, we can use the following truncated plug-in covariance estimator

$$\widetilde{\Sigma}_\xi = \frac{1}{n} \sum_{i=1}^{n} \frac{\psi_\xi(\|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}\|_2^2)}{\|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}\|_2^2} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^{\mathrm{T}}, \tag{2.12}$$

where $\xi > 0$ is a robustification parameter and $\widehat{\boldsymbol{\mu}}$ is a prespecified robust mean estimator. Given ξ and $\widehat{\boldsymbol{\mu}}$, the computational complexity of $\widetilde{\Sigma}_\xi$ is $O(nd^2)$. Assume $\boldsymbol{x}$ has a bounded fourth moment and let

$$\sigma_0^2 = \|\mathbb{E}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\}^2\|_2.$$

For any $z > 0$, following the proof of Lemma 2.1 in Wei and Minsker (2017), it can be similarly shown that conditioned on the event $\{\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \le C_1\sqrt{\mathrm{tr}(\Sigma)z/n}\}$ for some $C_1 > 0$, the truncated plug-in estimator $\widetilde{\Sigma}_\xi$ with $\xi = \sigma_0\sqrt{n/(z+\log d)}$ satisfies

$$\|\widetilde{\Sigma}_\xi - \Sigma\|_2 \lesssim \sigma_0\sqrt{\frac{z+\log d}{n}} \tag{2.13}$$

with probability at least $1-4e^{-z}$ as long as $n \ge C_2(\sigma_0/\bar{\lambda})^2(z+\log d)$, where $C_2 > 0$ is a constant depending only on C_1. Since $\sigma_0^2 \le \nu_4\bar{\lambda}\,\mathrm{tr}(\Sigma)$ (see Lemma 4.1 in Minsker and Wei (2020)), a sufficient sample size requirement for (2.13) is $n \gtrsim \nu_4\,\mathrm{r}(\Sigma)(z+\log d)$. On the other hand, it follows from Theorem 2.2.1 and Lemma 2.2.1 that the Huber mean estimator $\widehat{\boldsymbol{\mu}}_\tau$ with $\tau \asymp (m_4 n/z)^{1/4}$ satisfies

$$\begin{aligned}\|\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\|_2 &\lesssim \sqrt{\frac{\mathrm{tr}(\Sigma)+\bar{\lambda}z}{n}} + m_4^{1/4}\left(\frac{z}{n}\right)^{3/4} \\ &\lesssim \sqrt{\frac{\mathrm{tr}(\Sigma)+\bar{\lambda}z}{n}} + \kappa_4^{1/4}\mathrm{tr}(\Sigma)^{1/2}\left(\frac{z}{n}\right)^{3/4} \lesssim \sqrt{\frac{\mathrm{tr}(\Sigma)z}{n}}\end{aligned}$$

with probability at least $1-2e^{-z}$ when the sample size satisfies $n \gtrsim \nu_4\,\mathrm{r}(\Sigma)(z+\log d)$ and $z > 1$. In other words, the Huber mean estimator satisfies the required bound (with high probability) for the plug-in estimate in (2.12).

The following result complements Theorem 2.2.3 by providing a Berry-Esseen-type bound for the studentized robust statistic $\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle/(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_\xi\boldsymbol{u})^{1/2}$ uniformly over $\boldsymbol{u} \in \mathbb{R}^d$.

Theorem 2.2.4. Assume $m_4 = \mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^4 < \infty$ and let the sample size satisfy $n \gtrsim \mathrm{r}(\Sigma) + \log n$. For any $\gamma \in [1/3, 1/2]$, the Huber estimator $\widehat{\boldsymbol{\mu}}_\tau$ with $\tau \asymp m_4^{1/4}(n/\log n)^\gamma$ satisfies

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d, x\in\mathbb{R}} \left|\mathbb{P}\big\{\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle/(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_\xi \boldsymbol{u})^{1/2} \le x\big\} - \Phi(x)\right| \lesssim \nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)\log n}{n}}, \quad (2.14)$$

where $\widehat{\Sigma}_\xi$ is the U-type covariance estimator defined in (2.10) with $\xi \asymp v_0\sqrt{n/\log(nd)}$. In particular, $v_0^2 \le 2\nu_4\bar{\lambda}\,\mathrm{tr}(\Sigma)$.

From Theorem 2.2.4 we see that a sufficient condition for the asymptotic normality of the Studentized statistic $\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle/(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_\xi \boldsymbol{u})^{1/2}$ is $d\log^2(n) = o(n)$, the same as discussed following Theorem 2.2.3. Consequently, for any (deterministic) vector $\boldsymbol{u} \in \mathbb{R}^d$ of interest and $\alpha \in (0,1)$, we can construct robust (approximate) $100(1-\alpha)\%$ confidence interval for $\langle \boldsymbol{u}, \boldsymbol{\mu}\rangle$ as

$$\left[\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau\rangle - z_{\alpha/2}\frac{(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_\xi \boldsymbol{u})^{1/2}}{\sqrt{n}}, \langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau\rangle + z_{\alpha/2}\frac{(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_\xi \boldsymbol{u})^{1/2}}{\sqrt{n}}\right], \quad (2.15)$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ denotes the $(1-\alpha/2)$-th quantile of $\mathcal{N}(0,1)$.

We end this subsection with a uniform Gaussian approximation result, which provides theoretical guarantees for multiple testing procedures based on Studentized robust statistics.

Theorem 2.2.5. Assume $m_4 = \mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^4 < \infty$ and let the sample size satisfy $n \gtrsim \mathrm{r}(\Sigma) + \log n$. Let $\boldsymbol{G} = (G_1, \ldots, G_d)^{\mathrm{T}}$ be a d-dimensional zero-mean Gaussian random vector with covariance matrix $\mathrm{cov}(\boldsymbol{G}) = \mathrm{corr}(\Sigma) := (\sigma_{kl}/\sqrt{\sigma_{kk}\sigma_{ll}})_{1\le k,l\le d}$. For any $\gamma \in [1/3, 1/2]$, the Huber estimator $\widehat{\boldsymbol{\mu}}_\tau$ with $\tau \asymp m_4^{1/4}(n/\log n)^\gamma$ satisfies

$$\sup_{x\ge 0}\left|\mathbb{P}\left\{\max_{1\le k\le d}\left|\frac{\sqrt{n}(\widehat{\mu}_k - \mu_k)}{\sqrt{\widehat{\sigma}_{kk}}}\right| \le x\right\} - \mathbb{P}(\|\boldsymbol{G}\|_\infty \le x)\right| \lesssim \nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log^2(d)\log(n)\sqrt{\frac{d}{n}}, \quad (2.16)$$

where $\widehat{\sigma}_{kk}$ is the k-th diagonal element of $\widehat{\Sigma}_\xi$ defined in (2.10) and $\xi \asymp v_0\sqrt{n/\log(nd)}$.

Based on Theorem 2.2.5, we construct the confidence set

$$\times_{k=1}^{d}\left[\widehat{\mu}_k-\omega_\alpha\sqrt{\frac{\widehat{\sigma}_{kk}}{n}},\widehat{\mu}_k+\omega_\alpha\sqrt{\frac{\widehat{\sigma}_{kk}}{n}}\right] \tag{2.17}$$

for $\boldsymbol{\mu}\in\mathbb{R}^d$, which has level $1-\alpha$ asymptotically under the growth condition $d\log^4(d)\log^2(n)=o(n)$, where ω_α is the $(1-\alpha)$-quantile of $\|\boldsymbol{G}\|_\infty$. This confidence set is less conservative than the conventional multiple testing methods, such as the Bonferroni method and the Šidák method, which ignore the dependence structure among the d coordinates.

Another challenge is to compute ω_α due to the unknown covariance matrix $\mathrm{cov}(\boldsymbol{G})=\mathrm{corr}(\Sigma)$, or equivalently Σ. To this end, we apply a plug-in method by replacing Σ with its robust estimate $\widehat{\Sigma}_\xi$, and then compute the quantile of $\|\widehat{\boldsymbol{G}}\|_\infty$ with $\widehat{\boldsymbol{G}}\sim\mathscr{N}(\mathbf{0},\mathrm{corr}(\widehat{\Sigma}_\xi))$ via Monte Carlo simulations. Its validity (consistency) is guaranteed by Proposition 2.2.2 below as long as the right-hand side of the inequality is $o(1)$.

Proposition 2.2.2. Assume $m_4=\mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^4<\infty$, and let

$$\boldsymbol{G}=(G_1,\ldots,G_d)^{\mathrm{T}}\sim\mathscr{N}(\mathbf{0},\mathrm{corr}(\Sigma))\quad\text{and}\quad\widehat{\boldsymbol{G}}=(\widehat{G}_1,\ldots,\widehat{G}_d)^{\mathrm{T}}\sim\mathscr{N}(\mathbf{0},\mathrm{corr}(\widehat{\Sigma})),$$

where $\widehat{\Sigma}=\widehat{\Sigma}_\xi$ is the U-type covariance estimator defined in (2.10) with $\xi\asymp v_0\sqrt{n/\log(nd)}$. Then, with probability at least $1-2n^{-1}$, we have

$$\begin{aligned}\sup_{t\geq0}\left|\mathbb{P}\left(\max_{1\leq k\leq d}|\widehat{G}_k|\leq t\middle|x_1,\ldots,x_n\right)-\mathbb{P}\left(\max_{1\leq k\leq d}|G_k|\leq t\right)\right|&\\ \lesssim v_4^{1/2}(\bar{\lambda}/\underline{\lambda})^2\log(d)\log(n)\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}}.&\end{aligned} \tag{2.18}$$

Remark 2.2.4. In this section, the inference results of the Huber estimator $\widehat{\boldsymbol{\mu}}_\tau$ are limited to constructing a confidence interval for the one-dimensional projection of $\langle\boldsymbol{u},\boldsymbol{\mu}\rangle$, where $\boldsymbol{u}$ is a fixed direction in $\mathbb{R}^d$, or for obtaining confidence intervals simultaneously for each coordinate of $\boldsymbol{\mu}$. It is interesting to explore the possibility of extending these results to establish a multivariate

confidence region for the mean vector $\boldsymbol{\mu}$.

Following the idea from Spokoiny and Zhilova (2015); Chen and Zhou (2020), we propose a likelihood-based confidence set using the multiplier bootstrap method. To elaborate, let $u_1, \ldots, u_n$ be independent and identically distributed random variables that are independent of the observed data $\mathscr{D}_n := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and satisfy $\mathbb{E}(u_i) = 0, \mathrm{var}(u_i) = 1$ and $\mathbb{E}\exp(u_i^2/A^2) < \infty$ for some constant $A > 0$. Introducing the random weights $w_i = 1 + u_i$, we define the bootstrap loss and bootstrap Huber estimator as

$$\widehat{\mathscr{L}_\tau^{\mathrm{b}}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} w_i \rho_\tau(\|\boldsymbol{x}_i - \boldsymbol{\theta}\|_2) \quad \text{for} \quad \boldsymbol{\theta} \in \mathbb{R}^d,$$

and $\widehat{\boldsymbol{\mu}}_\tau^{\mathrm{b}} \in \mathrm{argmin}_{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\mu}}_\tau\|_2 \le R} \widehat{\mathscr{L}_\tau^{\mathrm{b}}}(\boldsymbol{\theta})$, respectively, where $R > 0$ is a prespecified radius parameter. Let $\mathbb{P}^*$ denote the conditional probability over the random multipliers given $\mathscr{D}_n$. Then, we denote $z_\alpha^{\mathrm{b}} = z_\alpha^{\mathrm{b}}(\mathscr{D}_n)$ to be the upper α-quantile ($0 < \alpha < 1$) of $\widehat{\mathscr{L}_\tau^{\mathrm{b}}}(\widehat{\boldsymbol{\mu}}_\tau) - \widehat{\mathscr{L}_\tau^{\mathrm{b}}}(\widehat{\boldsymbol{\mu}}_\tau^{\mathrm{b}})$, that is,

$$z_\alpha^{\mathrm{b}} = \inf\big\{z \ge 0 : \mathbb{P}^*\{\widehat{\mathscr{L}_\tau^{\mathrm{b}}}(\widehat{\boldsymbol{\mu}}_\tau) - \widehat{\mathscr{L}_\tau^{\mathrm{b}}}(\widehat{\boldsymbol{\mu}}_\tau^{\mathrm{b}}) > z\} \le \alpha\big\}.$$

Based on this, a confidence region for $\boldsymbol{\mu}$ at the given confidence level $1 - \alpha$ is given by

$$\{\boldsymbol{\theta} \in \mathbb{R}^d : \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}_\tau) \le z_\alpha^{\mathrm{b}}\}.$$

Practically, the conditional quantiles of $\widehat{\mathscr{L}_\tau^{\mathrm{b}}}(\widehat{\boldsymbol{\mu}}_\tau) - \widehat{\mathscr{L}_\tau^{\mathrm{b}}}(\widehat{\boldsymbol{\mu}}_\tau^{\mathrm{b}})$ can be computed with arbitrary precision by using Monte Carlo simulations.

Since a significant amount of additional work, including the derivation of the concentration property of the Wilks' expansion for the excess risk $\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}_\tau)$ and theoretical analysis of the bootstrap estimators, is still needed, we leave a rigorous theoretical investigation and validation of this approach to future work.

2.3 Differentially private robust mean estimation and inference

In this section, we propose a Gaussian differentially private robust mean estimator via the use of Huber loss and noisy gradient descent. The key observation is that the derivative of the Huber loss $\rho_\tau(\cdot)$, denoted by $\psi_\tau(\cdot)$, is bounded in magnitude by τ. Therefore, we can utilize the Gaussian mechanism (surveyed later in Section 2.3.1) to gain privacy. Note that $\widehat{\boldsymbol{\mu}}_\tau$ is defined as the minimum of a convex loss function, solvable by gradient descent and its many variants, we thus apply a noisy gradient descent method (Bassily, Smith and Thakurta, 2014) to construct a private version of $\widehat{\boldsymbol{\mu}}_\tau$ that is also statistically robust. We provide a deviation study of this private robust mean estimator and establish a Bahadur representation result based on which the validity of Gaussian approximation is also provided. This enables us to construct private confidence intervals for any linear combination of the mean vector.

2.3.1 Background on Gaussian differential privacy

The notion of differential privacy (DP) was first proposed to formalize the ad-hoc data privacy idea that a DP mechanism (randomized algorithm) M should make the distributions of $M(\boldsymbol{X})$ and $M(\boldsymbol{X}')$ similar for any pair of datasets $\boldsymbol{X}$ and $\boldsymbol{X}'$ that differ by only one entry or datum. Intuitively, an attacker is not able to detect whether any datum $\boldsymbol{x}$ belongs to the dataset $\boldsymbol{X}$ when a DP algorithm is applied to $\boldsymbol{X}$.

Definition 2.3.1 (Dwork *et al.* (2006a,b))**.** A dataset $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) \in \mathscr{X}^n$ consist of n data from some space $\mathscr{X}$. We say two datasets $\boldsymbol{X}$ and $\boldsymbol{X}'$ are neighbors if they differ by one entry. A randomized algorithm $M : \mathscr{X}^n \to \mathscr{Y}$ is said to be (ε, δ)-differentially private ((ε, δ)-DP) for $\varepsilon, \delta > 0$ if for any neighboring datasets $\boldsymbol{X}$ and $\boldsymbol{X}'$, and any measurable set $E \subseteq \mathscr{Y}$,

$$\mathbb{P}\{M(\boldsymbol{X}) \in E\} \leq e^{\varepsilon}\mathbb{P}\{M(\boldsymbol{X}') \in E\} + \delta,$$

where the probabilities are computed only over the randomness of the mechanism M.

From a statistical viewpoint, it is more natural to understand differential privacy in a hypothesis testing problem that takes the form

$$H_0 : \text{the underlying dataset is } \boldsymbol{X} \quad \text{vs} \quad H_1 : \text{the underlying dataset is } \boldsymbol{X}'. \tag{2.19}$$

As revealed by Wasserman and Zhou (2010), for any $0 < \alpha < 1$, the power of α-level test based on the output of an (ε, δ)-DP mechanism is upper bounded by $e^{\varepsilon}\alpha + \delta$. Therefore, it is impossible to construct a powerful test based on the output of an (ε, δ)-DP mechanism for small ε and δ.

Built upon the hypothesis testing interpretation, Dong, Roth and Su (2022) further proposed and advocated a notion of Gaussian differential privacy (GDP). GDP has an attractive interpretation to statisticians: the testing problem (2.19), e.g., identifying whether an individual is in a dataset, is at least as difficult as distinguishing between $\mathcal{N}(0,1)$ and $\mathcal{N}(\varepsilon,1)$ based on a single draw for some $\varepsilon > 0$. In other words, the privacy requirement in the notion of GDP can be precisely characterized by a single parameter ε. The formal definition is as follows.

Definition 2.3.2 (Dong, Roth and Su (2022))**.** Let M be a randomized algorithm. We say M is ε-Gaussian differentially private (GDP) if any α-level test ϕ for (2.19) has a power function

$$\beta(\alpha) \leq 1 - \Phi(\Phi^{-1}(1-\alpha) - \varepsilon)$$

for all $\alpha \in [0,1]$, where $\Phi(\cdot)$ is the standard normal distribution function.

The definition might not be as transparent as the intuition described in the univariate Gaussian distribution testing problem. Here, the function $\Phi(\Phi^{-1}(1-\alpha) - \varepsilon)$ describes the supreme of the type II errors of all α-level tests for distinguishing $\mathcal{N}(0,1)$ and $\mathcal{N}(\varepsilon,1)$ based on a single draw, which is achieved by the likelihood ratio test. For formal proof, we refer to Appendix A in Dong, Roth and Su (2022) for more details.

Despite the remarkable success of (ε, δ)-DP, GDP has a number of appealing properties compared to (ε, δ)-DP, as highlighted in Dong, Roth and Su (2022). Notably, among these distinct attributes, GDP has been proven to provide a tight privacy guarantee under composition, a feature that is absent in the (ε, δ)-DP mechanism (Murtagh and Vadhan, 2016). Furthermore, the GDP mechanism preserves a transparent hypothesis testing interpretation, while other relaxations of the (ε, δ)-DP mechanism, including concentrated differential privacy (CDP) (Dwork and Rothblum, 2016; Bun and Steinke, 2016) and Rényi differential privacy (Mironov, 2017), no longer have hypothesis testing interpretations.

We summarize several properties of GDP in the remainder of this subsection which are central in developing our private robust mean estimator. A variety of basic algorithms such as the gradient descent method used in Section 2.3.2 can be made private by simply adding a properly scaled Gaussian noise in the output. To this end, for any (non-private) statistics $\boldsymbol{h}(\boldsymbol{X}) \in \mathbb{R}^d$ of the dataset $\boldsymbol{X}$, define the sensitivity of $\boldsymbol{h}$ as

$$\text{sens}(\boldsymbol{h}) = \sup_{\boldsymbol{X}, \boldsymbol{X}'} \|\boldsymbol{h}(\boldsymbol{X}) - \boldsymbol{h}(\boldsymbol{X}')\|_2, \tag{2.20}$$

where the supremum is taken over all pairs of datasets $\boldsymbol{X}$ and $\boldsymbol{X}'$ that differ by one entry or datum. The following lemma provides the key device to construct Gaussian differentially private estimators. It is worth mentioning that only the univariate case ($d = 1$) was stated in Theorem 1 of Dong, Roth and Su (2022) but the extension to general $d \geq 1$ is straightforward.

Lemma 2.3.1. (Theorem 1 in Dong, Roth and Su (2022)) Define the Gaussian mechanism that operates on a statistic $\boldsymbol{h} \in \mathbb{R}^d$ as

$$M(\boldsymbol{X}) = \boldsymbol{h}(\boldsymbol{X}) + \frac{\text{sens}(\boldsymbol{h})}{\varepsilon} \boldsymbol{g},$$

where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_d)$. Then, the Gaussian mechanism M is ε-GDP.

Many algorithms, including our gradient descent approach in this paper, involve a

sequence of differentially private steps where the computation of each step relies on both the same dataset and outputs from previous steps. The joint mechanism is called "k-fold composition". Intuitively, the privacy would be gradually decayed along a sequence of outputs as the same dataset is used several times. One critical question is how privacy degrades given that each step alone is private. While the computation of precise privacy guarantees for compositions of (ε, δ)-DP mechanisms can be computationally challenging Murtagh and Vadhan (2016), the overall privacy guarantee for a composition of GDP mechanisms can be accurately reduced to the privacy guarantee of a single GDP mechanism. Indeed, this is one of the major reasons that GDP is advocated.

Lemma 2.3.2. (Corollary 2 in Dong, Roth and Su (2022)) Let $M_1 : \mathscr{X}^n \to \mathscr{Y}_1$ be the first mechanism and $M_t : \mathscr{X}^n \times \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_{t-1} \to \mathscr{Y}_t$ be the t-th mechanism for $t = 2, \ldots, k$. We define the k-fold composed mechanism $M : \mathscr{X}^n \to \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_k$ as $M(\boldsymbol{X}) = (y_1, y_2, \ldots, y_k)$ where $y_1 = M_1(\boldsymbol{X})$ and $y_t = M_t(\boldsymbol{X}, y_1, \ldots, y_{t-1})$ for $t = 2, \ldots, k$. If M_1 is ε_1-GDP and $M_t(\cdot, y_1, \ldots, y_{t-1})$ is ε_t-GDP for any $y_1 \in \mathscr{Y}_1, \ldots, y_{t-1} \in \mathscr{Y}_{t-1}$, then the k-fold composed mechanism M is $\sqrt{\varepsilon_1^2 + \ldots + \varepsilon_k^2}$-GDP.

Of note, the k-fold composition is different from the traditional composition of functions which is termed "post-processing" in the literature of privacy. In fact, privacy will not deteriorate if a GDP mechanism/algorithm is simply post-processed independently of the original dataset, as summarized in the lemma below.

Lemma 2.3.3. (Proposition 4 in Dong, Roth and Su (2022)) Let $M : \mathscr{X}^n \to \mathscr{Y}$ be ε-GDP. Denote a post-processing (randomized) algorithm $Proc : \mathscr{Y} \to \mathscr{Z}$ that maps the input $M(\boldsymbol{X})$ to some space $\mathscr{Z}$. Then the post-processing $Proc \circ M : \mathscr{X}^n \to \mathscr{Z}$ is also ε-GDP.

2.3.2 Private robust mean estimation: Finite sample theory

In this section, under the Gaussian differential privacy mechanism, we propose a differentially private Huber mean estimator via noisy gradient descent and provide a finite-sample

convergence analysis. Recall the non-private Huber estimator $\widehat{\boldsymbol{\mu}}_\tau$ defined in (2.1), which can be computed by gradient descent

$$\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} + \frac{\eta_0}{n}\sum_{i=1}^{n}\frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}), \;\; t = 0, 1, \ldots,$$

where $\eta_0 > 0$ is the step size (learning rate) and $\boldsymbol{\mu}^{(0)}$ is the initial value. To achieve a certain level of privacy, we consider the following noisy version of gradient descent (Bassily, Smith and Thakurta, 2014). For a predetermined number of iterations T, it computes

$$\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} + \frac{\eta_0}{n}\sum_{i=1}^{n}\frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}) + 2T^{1/2}\tau\frac{\eta_0}{\varepsilon n}\boldsymbol{g}_t \tag{2.21}$$

for $t = 0, 1, \ldots, T-1$, where $\eta_0 > 0$ is the step size, $\{\boldsymbol{g}_t\}_{t=0}^{T-1}$ is a sequence of independent standard d-variate normal random vectors, and $\varepsilon > 0$ is the privacy parameter. The final private estimator is denoted by $\boldsymbol{\mu}^{(T)}$. Here the scale of the Gaussian noise is carefully chosen based on the properties of GDP, i.e., Lemmas 2.3.1-2.3.2.

Proposition 2.3.1. Given an initial estimate $\boldsymbol{\mu}^{(0)} \in \mathbb{R}^d$ and dataset $\boldsymbol{X}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, consider the noisy gradient descent iterates $\{\boldsymbol{\mu}^{(t)}\}_{t=0}^{T}$ defined in (2.21). Then the final output $\boldsymbol{\mu}^{(T)}$ is ε-GDP.

Proof. Consider two datasets $\boldsymbol{X}_n$ and $\boldsymbol{X}'_n$ that differ by one datum, say $\boldsymbol{x}_1 \in \boldsymbol{X}_n$ versus $\boldsymbol{x}'_1 \in \boldsymbol{X}'_n$. Let the (vanilla) gradient update be

$$\boldsymbol{h}(\boldsymbol{X}_n, \boldsymbol{\mu}^{(t)}) = \boldsymbol{\mu}^{(t)} + \frac{\eta_0}{n}\sum_{i=1}^{n}\frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}),$$

and define $\boldsymbol{h}(\boldsymbol{X}_n', \boldsymbol{\mu}^{(t)})$ similarly. At the first iteration, note that

$$\begin{aligned}
&\|\boldsymbol{h}(\boldsymbol{X}_n, \boldsymbol{\mu}^{(0)}) - \boldsymbol{h}(\boldsymbol{X}_n', \boldsymbol{\mu}^{(0)})\|_2 \\
&= \frac{\eta_0}{n}\left\| \frac{\psi_\tau(\|\boldsymbol{x}_1 - \boldsymbol{\mu}^{(0)}\|_2)}{\|\boldsymbol{x}_1 - \boldsymbol{\mu}^{(0)}\|_2}(\boldsymbol{x}_1 - \boldsymbol{\mu}^{(0)}) - \frac{\psi_\tau(\|\boldsymbol{x}_1' - \boldsymbol{\mu}^{(0)}\|_2)}{\|\boldsymbol{x}_1' - \boldsymbol{\mu}^{(0)}\|_2}(\boldsymbol{x}_1' - \boldsymbol{\mu}^{(0)}) \right\|_2 \leq \frac{2\tau\eta_0}{n}.
\end{aligned}$$

Therefore, the sensitivity of $\boldsymbol{h}$ is upper bounded by $2\tau\eta_0/n$. By Lemma 2.3.1, adding a Gaussian noise $2T^{1/2}\tau\eta_0(\varepsilon n)^{-1}\boldsymbol{g}_0$ to the gradient update makes this step $(T^{-1/2}\varepsilon)$-GDP. Consequently, $\boldsymbol{\mu}^{(1)}$ is $(T^{-1/2}\varepsilon)$-GDP since the initial estimate $\boldsymbol{\mu}^{(0)}$ is deterministic. The second iterate $\boldsymbol{\mu}^{(2)} = \boldsymbol{\mu}^{(2)}(\boldsymbol{X}_n)$ takes $\boldsymbol{\mu}^{(1)}$ as input in addition to the dataset. It thus follows from Lemma 2.3.2 that the two-fold composed (joint) mechanism $(\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)})$ is $\sqrt{\varepsilon^2/T + \varepsilon^2/T}$-GDP. Using the same argument repeatedly, we conclude that the T-fold composed mechanism $(\boldsymbol{\mu}^{(1)}, \ldots, \boldsymbol{\mu}^{(T)})$ is ε-GDP, and so is $\boldsymbol{\mu}^{(T)}$. □

To establish the statistical properties of the ε-GDP robust estimate $\boldsymbol{\mu}^{(T)}$, we first derive a concentration bound conditioning on some "good" event with a set of parameters. Next, we show that this event occurs with high probability when the parameters are properly chosen. To begin with, given parameters $r_0 > 0$ and $\chi \in (0,1)$, define the event

$$\mathscr{E}_1 = \mathscr{E}_1(r_0, \chi) = \big\{\widehat{\boldsymbol{\mu}} \in \Theta(r_0/2)\big\} \cap \big\{\nabla^2 \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) \succeq (1-\chi)\mathbf{I}_d,\ \forall \boldsymbol{\theta} \in \Theta(r_0)\big\}, \tag{2.22}$$

where

$$\Theta(r) := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \boldsymbol{\mu}\|_2 \leq r\} \ \text{ for every } r > 0, \tag{2.23}$$

and $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_\tau$ is the non-private robust estimator defined in (2.1). We are now ready to present an oracle-type concentration bound of the private estimator $\boldsymbol{\mu}^{(T)}$ around $\widehat{\boldsymbol{\mu}}$ conditioning on $\mathscr{E}_1$.

Theorem 2.3.1. Consider the private estimate $\boldsymbol{\mu}^{(T)}$ obtained from noisy gradient descent (2.21) with step size $\eta_0 \in (0,1]$ and the initial estimate $\boldsymbol{\mu}^{(0)} \in \Theta(r_0)$ for some $r_0 > 0$. Let $\chi \in (0,1), z >$

0 and $T \geq 1$. Define the optimization error r_{opt} and the privacy error r_{p} as

$$r_{\text{opt}}^2(T) = (1-\rho)^T r_0^2 \quad \text{and} \quad r_{\text{p}}^2(T) = \eta_0 T\{\eta_0 + (1-\chi)^{-1}\}\left(\frac{d}{\rho} + z\right)\left(\frac{\tau}{\varepsilon n}\right)^2,$$

where $\rho = (1-\chi)^2\eta_0^2$. Assume that the sample size satisfies

$$n \gtrsim T^{1/2}\tau\frac{\sqrt{d} + \sqrt{\log T + z}}{(1-\chi)\varepsilon r_0}. \tag{2.24}$$

Then, conditioning on the event $\mathscr{E}_1 = \mathscr{E}_1(r_0, \chi)$, $\boldsymbol{\mu}^{(T)}$ satisfies

$$\|\boldsymbol{\mu}^{(T)} - \widehat{\boldsymbol{\mu}}\|_2^2 \lesssim r_{\text{opt}}^2(T) + r_{\text{p}}^2(T)$$

with probability (over $\{\boldsymbol{g}_t\}_{t=0}^{T-1}$) at least $1 - 2e^{-z}$.

Theorem 2.3.1 provides a concentration bound with two terms: optimization error $r_{\text{opt}}(T)$ and privacy error $r_{\text{p}}(T)$. As the number of iterations T increases and the step size η_0 approaches to 1, the optimization error decreases, whereas the privacy error increases. In addition to these two errors, we also need to account for the statistical error of $\widehat{\boldsymbol{\mu}}$ in (2.2) to obtain a deviation bound for $\boldsymbol{\mu}^{(T)}$ around the true mean $\boldsymbol{\mu}$. Hence, we need to select an appropriate number of iterations T to balance $r_{\text{opt}}(T)$ and $r_{\text{p}}(T)$, while also choosing τ to balance bias, robustness and privacy error.

Before selecting appropriate parameters in Theorem 2.3.1 to consider the trade-off between different sources of error and make the event $\mathscr{E}_1$ occur with high probability, we provide a few remarks regarding the assumption on the initial iterate $\boldsymbol{\mu}^{(0)}$. In Theorem 2.3.1, the minimum sample size required and the event $\mathscr{E}_1$ depend on r_0, the ℓ_2 distance between the initial value $\boldsymbol{\mu}^{(0)}$ and the true mean $\boldsymbol{\mu}$. The following proposition shows that if $\|\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu}\|_2 > r_0$, implying $R_0 := \|\boldsymbol{\mu}^{(0)} - \widehat{\boldsymbol{\mu}}\|_2 > r_0/2$ conditioning on the event $\mathscr{E}_1(r_0, \chi)$, then it takes as many as $T_0 = O((R_0/r_0)^2)$ noisy gradient descent iterations to ensure that the above initial value condition

is met, that is, $\|\boldsymbol{\mu}^{(T_0)} - \boldsymbol{\mu}\|_2 \leq r_0$.

Proposition 2.3.2. Assume the step size $\eta_0 \in (0,1]$ and let $R_0 = \|\boldsymbol{\mu}^{(0)} - \widehat{\boldsymbol{\mu}}\|_2$. For any $z > 0$ and $\Delta > 0$, let $T_0 \geq R_0^2/(\eta_0 \Delta)$ and the sample size satisfy

$$n \gtrsim \frac{T^{1/2} B_{T_0}}{\varepsilon} \max\left\{ \frac{\tau(R_0 + T_0\tau)}{\Delta}, T_0 \frac{\tau\eta_0}{R_0}, T_0 \left(\frac{\tau\eta_0}{R_0}\right)^2 \right\},$$

where $B_{T_0} = B_{T_0}(z) = \sqrt{d} + \sqrt{2(\log T_0 + z)}$ and T is the predetermined number of iterations in the definition of noisy gradient descent (2.21). Then, $\boldsymbol{\mu}^{(T_0)}$ satisfies $\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(T_0)}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) \leq \Delta$ with probability (over $\{\boldsymbol{g}_t\}_{t=0}^{T_0-1}$) at least $1 - e^{-z}$. In particular, conditioning on $\mathscr{E}_1(r_0, \chi)$ and taking $\Delta = (1-\chi) r_0^2/8$, we have

$$\|\boldsymbol{\mu}^{(T_0)} - \boldsymbol{\mu}\|_2 \leq r_0 \tag{2.25}$$

with probability (over $\{\boldsymbol{g}_t\}_{t=0}^{T_0-1}$) at least $1 - e^{-z}$.

Next, the following proposition shows that, with suitably chosen (r_0, χ), the event $\mathscr{E}_1(r_0, \chi)$ occurs with high probability.

Proposition 2.3.3. Assume the same conditions as in Theorem 2.2.1. Moreover, for a given $z > 0$, let (r_0, χ, τ) and n satisfy

$$r_0 = \frac{\tau}{2} \text{ and } \chi = \chi(n, z) := \frac{4\mathrm{tr}(\Sigma)}{\tau^2} + \sqrt{\frac{z}{2n}}.$$

Then, the event $\mathscr{E}_1(r_0, \chi)$ with $0 < \chi < 1$ occurs with probability $1 - 3e^{-z}$ as long as $\tau \gtrsim \sqrt{\mathrm{tr}(\Sigma)}$ and $n \gtrsim \mathrm{r}(\Sigma) + z$.

Combining Proposition 2.3.3 with Theorem 2.3.1 yields the following result.

Corollary 2.3.1. Let $\varepsilon > 0$ be a predetermined privacy parameter. For any $z > 1$, let the sample

size satisfy

$$n \gtrsim \max\left\{ \mathrm{r}(\Sigma) + z, T^{1/2}\frac{\sqrt{d}+\sqrt{\log T + z}}{\varepsilon} \right\} \tag{2.26}$$

with $\tau \gtrsim \sqrt{\mathrm{tr}(\Sigma)}$. Starting at $\boldsymbol{\mu}^{(0)} \in \Theta(\tau/2)$, the ε-GDP robust estimator $\boldsymbol{\mu}^{(T)}$ defined through noisy gradient descent (2.21) with $\eta_0 = 1$ and $T \asymp \log(n/z)$ satisfies the bounds

$$\|\boldsymbol{\mu}^{(T)} - \widehat{\boldsymbol{\mu}}\|_2 \lesssim \tau\frac{z}{n} + (d+z)^{1/2}(\log n)^{1/2}\frac{\tau}{\varepsilon n} \tag{2.27}$$

and

$$\|\boldsymbol{\mu}^{(T)} - \boldsymbol{\mu}\|_2 \lesssim \bar{\lambda}^{1/2}\sqrt{\frac{\mathrm{r}(\Sigma)+z}{n}} + \tau\frac{z}{n} + b_\tau + (d+z)^{1/2}(\log n)^{1/2}\frac{\tau}{\varepsilon n} \tag{2.28}$$

with probability at least $1 - 5e^{-z}$, where b_τ is the bias term defined in (2.3).

Remark 2.3.1. Taking $r_0 = \tau/2$ in Proposition 2.3.2, we observe that even when the initial iterate $\boldsymbol{\mu}^{(0)}$ fails to meet the assumption of Corollary 2.3.1, that is, when $\|\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu}\|_2 > \tau/2$ which implies

$$\frac{\tau}{4} < R_0 := \|\boldsymbol{\mu}^{(0)} - \widehat{\boldsymbol{\mu}}\|_2 \le \|\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu}\|_2 + \frac{\tau}{2}$$

conditioning on the event $\mathscr{E}_1(r_0, \chi)$, we only need $T_0 \asymp R_0^2/\tau^2$ iterations to satisfy the initial condition. Then, provided that $T_0 < T$, we can consider $\boldsymbol{\mu}^{(T_0)}$ as an initial estimate instead of $\boldsymbol{\mu}^{(0)}$ in Theorem 2.3.1, resulting in

$$\|\boldsymbol{\mu}^{(T)} - \widehat{\boldsymbol{\mu}}\|_2 \lesssim r_{\mathrm{opt}}(T - T_0) + r_{\mathrm{p}}(T - T_0)$$

with high probability, where $r_{\mathrm{opt}}(\cdot)$ and $r_{\mathrm{p}}(\cdot)$ are defined in Theorem 2.3.1. Note that we require $T \asymp \log n$ in Corollary 2.3.1, and we choose τ to diverge as $n \to \infty$ to control the bias b_τ,

implying $T_0 \asymp R_0^2/\tau^2 = O(1)$. Consequently, we have $T - T_0 \asymp T$, and the deviation bound of Corollary 2.3.1 remains valid even when the initial condition is not satisfied. Furthermore, we also note that since $T_0 \asymp R_0^2/\tau^2$, the sample size requirement of Proposition 2.3.2 reduces to

$$n \gtrsim \frac{T^{1/2} B_{T_0}}{\varepsilon} \max\left(\frac{R_0}{\tau}, T_0\right) \asymp \frac{T^{1/2} B_{T_0}}{\varepsilon} T_0.$$

Given that we have $T_0 = O(1)$, the sample size requirement of Corollary 2.3.1 implies that the above inequality holds as long as n is sufficiently large. Therefore, Corollary 2.3.1 and Proposition 2.3.2 together ensure that the accuracy of the initial estimator does not significantly impact the algorithm's convergence.

Remark 2.3.2. From Corollary 2.3.1 we see that the parameter τ not only controls the bias-robustness tradeoff, but also determines the global sensitivity. The latter is the key to the privacy-preserving Gaussian mechanism (Dong, Roth and Su, 2022), as summarized in Lemma 2.3.1. Assume that $\boldsymbol{x}$ has bounded q-th moment $m_q = \mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^q$ ($q \geq 2$), satisfying $\mathrm{tr}(\Sigma)^{1/2} \leq m_q^{1/q} \leq \kappa_q^{1/q} \mathrm{tr}(\Sigma)^{1/2}$ according to (2.5). Taking $z = \log n$ and

$$\tau \asymp \nu_q^{1/q} \bar{\lambda}^{1/(2q)} m_q^{(q-1)/q^2} \left\{ \frac{\varepsilon n}{\sqrt{(d + \log n) \log n}} \right\}^{1/q},$$

employing Lemma 2.2.1 yields

$$\begin{aligned} &\|\boldsymbol{\mu}^{(T)} - \boldsymbol{\mu}\|_2 \\ &\quad \lesssim \bar{\lambda}^{1/2} \sqrt{\frac{\mathrm{r}(\Sigma) + \log n}{n}} + \nu_q^{1/q} \bar{\lambda}^{1/(2q)} \mathrm{tr}(\Sigma)^{(q-1)/(2q)} \left\{ \frac{(\log n)^{1/2} (d + \log n)^{1/2}}{\varepsilon n} \right\}^{1-1/q} \end{aligned}$$

with probability exceeding $1 - 5n^{-1}$. Comparing this result with the bound (2.7) for non-private robust estimator $\widehat{\boldsymbol{\mu}}$, with a dimension-free parameter ν_q and bounded $\bar{\lambda}$, we have a larger second

term

$$\nu_q^{1/q}\bar{\lambda}^{1/(2q)}\mathrm{tr}(\Sigma)^{(q-1)/(2q)}\left\{\frac{(\log n)^{1/2}(d+\log n)^{1/2}}{\varepsilon n}\right\}^{1-1/q} \lesssim \left(\frac{d\log n}{\varepsilon n}\right)^{1-1/q},$$

which quantifies the "cost of privacy" of our ε-GDP robust mean estimator $\boldsymbol{\mu}^{(T)}$ compared to its non-private counterpart $\widehat{\boldsymbol{\mu}}$.

Recently, Cai, Wang and Zhang (2021) showed that the minimax ℓ_2 risk of sub-Gaussian mean estimation with (ε,δ)-differential privacy is at least $O(\sqrt{\frac{d}{n}}+\frac{d\log^{1/2}(1/\delta)}{\varepsilon n})$, explicitly demonstrating its dependence on ε and δ. By Corollary 1 in Dong, Roth and Su (2022), an algorithm is ε-GDP if and only if $(\varepsilon,\delta(\varepsilon))$-DP, where $\delta(\varepsilon)=\Phi(-1+\varepsilon/2)-e^{\varepsilon}\Phi(-1-\varepsilon/2)$. Consequently, the cost of privacy of sub-Gaussian mean estimation with ε-GDP is thus at least $O(\frac{d}{\varepsilon n})$, up to logarithmic factors. In fact, $\sup_{q\geq 1}\kappa_q^{1/q}$ is upper bounded by a constant if $\boldsymbol{x}$ is sub-Gaussian with a finite Orlicz ψ_2-norm Vershynin (2018). In this case, it can be shown from Corollary 2.3.1 that with $\tau\asymp\sqrt{d+\log n}$, the resulting ε-GDP Huber estimator attains the minimax-optimal ℓ_2 convergence rate, up to logarithmic factors.

For mean estimation under bounded q-th moment, the ℓ_2 error of the proposed robust ε-GDP estimator with the optimal τ is of order $O(\sqrt{d/n}+(\frac{d}{\varepsilon n})^{1-1/q})$ with high probability, ignoring the $\log(n)$-factor. The slower term $(\frac{d}{\varepsilon n})^{1-1/q}$ characterizes the impact of heavy-tailedness and privacy. For $q=2$, we find that this matches the lower bound on the ℓ_2-risk (Kamath, Mouzakis and Singhal, 2022). The latter proposed an algorithm for achieving (ε,δ)-DP with polynomial-time complexity, albeit with a more intricate implementation. The lower bound for $q>2$ remains unknown. Furthermore, for $q>2$, the privacy cost of the ℓ_2-risk of our estimator aligns with that of the (ε,δ)-differentially private estimator proposed in Kamath, Singhal and Ullman (2020). Finally, it is worth noting that the tail probability bound for the private robust estimator we obtained decays exponentially with z, while the proof of Theorem 39 in Kamath, Singhal and Ullman (2020) employs Markov's inequality, resulting in a bound with a polynomial decay.

Combining the deviation bound (2.27) with Theorem 2.2.2, we obtain a non-asymptotic Bahadur representation for the ε-GDP Huber estimator $\boldsymbol{\mu}^{(T)}$ as stated below.

Corollary 2.3.2. For any $z > 0$, assume that all the conditions in Corollary 2.3.1 hold. Then, the ε-GDP Huber estimator $\boldsymbol{\mu}^{(T)}$ satisfies

$$\begin{aligned}
&\left\| \boldsymbol{\mu}^{(T)} - \boldsymbol{\mu} - \frac{1}{n}\sum_{i=1}^{n} \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}) \right\|_2 \\
&\lesssim \left\{ \bar{\lambda}^{1/2} \sqrt{\frac{\mathrm{r}(\Sigma) + z}{n}} + \frac{\tau z}{n} + b_\tau \right\} \left(\frac{m_q}{\tau^q} + \sqrt{\frac{z}{n}} \right) + (d+z)^{1/2} (\log n)^{1/2} \frac{\tau}{\varepsilon n} \qquad (2.29)
\end{aligned}$$

with probability at least $1 - 8e^{-z}$.

Corollary 2.3.2 shows that with high probability, $\sqrt{n}(\boldsymbol{\mu}^{(T)} - \boldsymbol{\mu})$ is first-order equivalent to the linear term

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}),$$

which determines the asymptotic distribution of $\boldsymbol{\mu}^{(T)}$ when τ is chosen in a suitable way. Based on the Bahadur representation (2.29), in Section 2.3.3 we obtain a Gaussian approximation result for $\boldsymbol{\mu}^{(T)}$ under a bounded third or fourth moment condition.

2.3.3 Construction of private confidence intervals

In this section, we present a Gaussian approximation result for the ε-GDP Huber estimator $\boldsymbol{\mu}^{(T)}$ under the bounded q-th moment condition with $q \geq 3$, based on which differentially private confidence intervals can be constructed. Without loss of generality, we assume $\varepsilon \leq 1$.

Theorem 2.3.2. Assume $m_q = \mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^q < \infty$ for some $q \geq 3$. Let the sample size satisfy (2.26) and $n \gtrsim \sqrt{(d + \log n)\log n}/\varepsilon$ with $z = \log n$ and $\tau \asymp m_q^{1/q}\{\varepsilon n / \sqrt{(d+\log n)\log n}\}^{1/q}$. For $\boldsymbol{\mu}^{(0)} \in$

$\Theta(\tau/2)$, the ε-GDP Huber estimator $\boldsymbol{\mu}^{(T)}$ with $\eta_0 = 1$ and $T \asymp \log(n/\log n)$ satisfies

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d, x\in\mathbb{R}} \left|\mathbb{P}(\sqrt{n}\langle \boldsymbol{u}/\|\boldsymbol{u}\|_\Sigma, \boldsymbol{\mu}^{(T)} - \boldsymbol{\mu}\rangle \le x) - \Phi(x)\right| \lesssim \frac{m_q^{1/q}}{\underline{\lambda}^{1/2}} \left\{\frac{\sqrt{(d+\log n)\log n}}{\varepsilon}\right\}^{1-1/q} \left(\frac{1}{n}\right)^{1/2-1/q} + \nu_q^{2/q} \left\{\frac{\sqrt{(d+\log n)\log n}}{\varepsilon n}\right\}^{1-2/q}, \tag{2.30}$$

where ν_q is defined in (2.4).

Remark 2.3.3. Since $m_q^{1/q} \le \kappa_q^{1/q} \mathrm{tr}(\Sigma)^{1/2}$, the first term on the right-hand side of (2.30) is further bounded, up to constants, by

$$\mathrm{r}(\Sigma)^{1/2} \left\{\frac{\sqrt{(d+\log n)\log n}}{\varepsilon}\right\}^{1-1/q} \left(\frac{1}{n}\right)^{1/2-1/q},$$

which is the leading term under mild conditions. This term quantifies the impact of the proposed privacy-preserving random noise mechanism and the heavy-tailedness of $\boldsymbol{x}$. When $\boldsymbol{x}$ follows a sub-Gaussian distribution with a finite Orlicz ψ_2-norm, the above rate can be improved to $\varepsilon^{-1}\sqrt{\mathrm{r}(\Sigma)(d+\log n)\log(n)/n}$ (as if $q = \infty$). Comparing this result with Theorem 2.2.3 for non-private robust estimator $\widehat{\boldsymbol{\mu}}$, the different choice of $\tau \asymp m_q^{1/q}\{\varepsilon n/\sqrt{(d+\log n)\log n}\}^{1/q}$ is due to the tradeoff among bias, robustness and global sensitivity. Consequently, we have a slower rate for the Berry-Esseen bound. Similar to the discussion following Theorem 2.2.3, from an asymptotic view with a fixed value of ε, any linear combination of the coordinates of $\sqrt{n}(\boldsymbol{\mu}^{(T)} - \boldsymbol{\mu})$ converges in distribution to a normal distribution under a sufficient growth condition $d^{(2q-1)/(q-2)}(\log n)^{(q-1)/(q-2)} = o(n)$.

To construct confidence intervals/sets in the differential privacy setting, the plug-in method described in Section 2.2.2 cannot be directly applied. In the following, we introduce a differentially private counterpart of the robust covariance estimator $\widehat{\Sigma}_\xi$ given in (2.10).

Proposition 2.3.4. Let $\mathbf{E} \in \mathbb{R}^{d\times d}$ be a symmetric random matrix whose upper-triangular and diagonal entries are i.i.d. $\mathcal{N}(0,1)$. For any robustification parameter $\xi > 0$, the perturbed robust estimate $\widehat{\Sigma}_\xi + \frac{4\xi}{\varepsilon n}\mathbf{E}$ is ε-GDP.

Proof. Let $D = \frac{d(d+1)}{2}$, and denote by $\boldsymbol{h}(\boldsymbol{X}_n)$ the D-dimensional vector that consists of the upper-triangular and diagonal entries of the covariance estimator $\widehat{\Sigma}_\xi = \widehat{\Sigma}_\xi(\boldsymbol{X}_n) \in \mathbb{R}^{d\times d}$. Consider two datasets $\boldsymbol{X}_n$ and $\boldsymbol{X}'_n$ that differ by one datum, say $\boldsymbol{x}_1 \in \boldsymbol{X}_n$ versus $\boldsymbol{x}'_1 \in \boldsymbol{X}'_n$. We have

$$
\begin{aligned}
&\|\boldsymbol{h}(\boldsymbol{X}_n) - \boldsymbol{h}(\boldsymbol{X}'_n)\|_2 \le \|\widehat{\Sigma}_\xi(\boldsymbol{X}_n) - \widehat{\Sigma}_\xi(\boldsymbol{X}'_n)\|_{\mathrm{F}} \\
&\le \left\| \frac{2}{n(n-1)} \sum_{2\le i\le n} \left\{ \psi_\xi\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_i\|_2^2}{2} \right) \frac{(\boldsymbol{x}_1 - \boldsymbol{x}_i)(\boldsymbol{x}_1 - \boldsymbol{x}_i)^{\mathrm{T}}}{\|\boldsymbol{x}_1 - \boldsymbol{x}_i\|_2^2} \right.\right. \\
&\qquad\qquad \left.\left. - \psi_\xi\left(\frac{\|\boldsymbol{x}'_1 - \boldsymbol{x}_i\|_2^2}{2} \right) \frac{(\boldsymbol{x}'_1 - \boldsymbol{x}_i)(\boldsymbol{x}'_1 - \boldsymbol{x}_i)^{\mathrm{T}}}{\|\boldsymbol{x}'_1 - \boldsymbol{x}_i\|_2^2} \right\} \right\|_{\mathrm{F}} \\
&\le \frac{4\xi}{n}.
\end{aligned}
$$

By Lemma 2.3.1, $\boldsymbol{h}(\boldsymbol{X}_n) + \frac{4\xi}{\varepsilon n}\boldsymbol{g}$ with $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$ is ε-GDP. Then it follows from Lemma 2.3.3 that $\widehat{\Sigma}_\xi + \frac{4\xi}{\varepsilon n}\mathbf{E}$ is also ε-GDP. □

Remark 2.3.4. Based on Remark 2.2.3, we further consider a differentially private counterpart of the truncated covariance estimator $\widetilde{\Sigma}_\xi$ given in (2.12), which has a much smaller computational complexity than $\widehat{\Sigma}_\xi$. Let $\mathbf{E} \in \mathbb{R}^{d\times d}$ be the same random matrix as above. Following a similar argument as in Propositions 2.3.1 and 2.3.4, we see that given a robustification parameter $\xi > 0$ and an ε-GDP mean estimator $\widehat{\boldsymbol{\mu}}$, the perturbed plug-in covariance estimator $\widetilde{\Sigma}_\xi + \frac{2\xi}{\varepsilon n}\mathbf{E}$ is $\sqrt{2}\varepsilon$-GDP.

Note that the perturbed matrix $\widehat{\Sigma}_\xi + \frac{4\xi}{\varepsilon n}\mathbf{E}$ may not be positive semi-definite, and therefore is not always a valid covariance estimator. To avoid this issue, we project $\widehat{\Sigma}_\xi + \frac{4\xi}{\varepsilon n}\mathbf{E}$ onto a cone

of positive definite matrices $\{\mathbf{H} : \mathbf{H} \succeq \zeta \mathbf{I}\}$ and obtain

$$\widehat{\Sigma}_{\xi,\varepsilon} = \operatorname*{argmin}_{\mathbf{H} \succeq \zeta \mathbf{I}} \left\| \mathbf{H} - \left(\widehat{\Sigma}_{\xi} + \frac{4\xi}{\varepsilon n} \mathbf{E} \right) \right\|_2, \tag{2.31}$$

where $\zeta > 0$ is sufficiently small. By Lemma 2.3.3, $\widehat{\Sigma}_{\xi,\varepsilon}$ is also ε-GDP because it is the outcome of a deterministic post-processing step. The following proposition provides a non-asymptotic concentration bound of the private covariance estimator $\widehat{\Sigma}_{\xi,\varepsilon}$.

Proposition 2.3.5. Assume $\boldsymbol{x}$ has the finite fourth moment so that v_0^2 given in (2.11) is well-defined. Let $n_0 = \lfloor n/2 \rfloor$ be the largest integer not exceeding $n/2$. Then, the private covariance estimator $\widehat{\Sigma}_{\xi,\varepsilon}$ defined in (2.31) with $\xi = v_0 \sqrt{n_0 / \log(2nd)}$ satisfies

$$\|\widehat{\Sigma}_{\xi,\varepsilon} - \Sigma\|_2 \lesssim v_0 \sqrt{\frac{\log(nd)}{n}} + \frac{v_0}{\varepsilon} \sqrt{\frac{d}{n}}$$

with probability at least $1 - 2n^{-1}$.

Similarly to Theorem 2.2.4, we establish below a Berry-Esseen-type bound for the studentized private statistic $\sqrt{n}\langle \boldsymbol{u}, \boldsymbol{\mu}^{(T)} - \boldsymbol{\mu} \rangle / (\boldsymbol{u}^{\mathrm{T}} \widehat{\Sigma}_{\xi,\varepsilon} \boldsymbol{u})^{1/2}$ for any $\boldsymbol{u} \in \mathbb{R}^d$.

Corollary 2.3.3. Under the same conditions as in Theorem 2.3.2 with $q \geq 4$, we have

$$\begin{aligned} &\sup_{\boldsymbol{u} \in \mathbb{R}^d, x \in \mathbb{R}} \left| \mathbb{P}\{ \sqrt{n} \langle \boldsymbol{u}, \boldsymbol{\mu}^{(T)} - \boldsymbol{\mu} \rangle / (\boldsymbol{u}^{\mathrm{T}} \widehat{\Sigma}_{\xi,\varepsilon} \boldsymbol{u})^{1/2} \leq x \} - \Phi(x) \right| \\ &\lesssim \frac{m_q^{1/q}}{\underline{\lambda}^{1/2}} \left\{ \frac{\sqrt{d + \log n) \log n}}{\varepsilon} \right\}^{1-1/q} \left(\frac{1}{n} \right)^{1/2-1/q} + \nu_4^{1/2} \frac{\bar{\lambda}}{\underline{\lambda}} \sqrt{\mathrm{r}(\Sigma) \log n} \left(\sqrt{\frac{\log n}{n}} + \frac{1}{\varepsilon} \sqrt{\frac{d}{n}} \right), \end{aligned} \tag{2.32}$$

where $\xi = v_0 \sqrt{n / \log(2nd)}$ and $\widehat{\Sigma}_{\xi,\varepsilon}$ is the differentially private covariance estimator defined in (2.31).

Recall from Theorem 2.2.4 that $v_0^2 \leq 2\nu_4 \bar{\lambda} \operatorname{tr}(\Sigma)$. Based on Theorem 2.3.2 and Proposition 2.3.5, the proof of (2.32) is almost identical to that of Theorem 2.2.4, and thus is omitted.

Ignoring the moment parameters and the condition number $\bar{\lambda}/\underline{\lambda}$ of Σ, the leading term on the right-hand side of (2.32) is

$$\mathrm{r}(\Sigma)^{1/2}\left\{\frac{\sqrt{(d+\log n)\log n}}{\varepsilon}\right\}^{1-1/q}\left(\frac{1}{n}\right)^{1/2-1/q},$$

which essentially matches the upper bound in (2.30). In other words, the covariance estimation error is dominated by the Gaussian approximation error under privacy.

Based on the Gaussian approximation result in Corollary 2.3.3, for any $\alpha \in (0,1)$ and deterministic vector $\boldsymbol{u} \in \mathbb{R}^d$, we construct the following $(\sqrt{2}\varepsilon)$-GDP (approximate) $100(1-\alpha)\%$ confidence interval of $\langle \boldsymbol{u}, \boldsymbol{\mu} \rangle$:

$$\left[\langle \boldsymbol{u}, \boldsymbol{\mu}^{(T)} \rangle - z_{\alpha/2}\frac{(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_{\xi,\varepsilon}\boldsymbol{u})^{1/2}}{\sqrt{n}}, \langle \boldsymbol{u}, \boldsymbol{\mu}^{(T)} \rangle + z_{\alpha/2}\frac{(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_{\xi,\varepsilon}\boldsymbol{u})^{1/2}}{\sqrt{n}}\right], \tag{2.33}$$

where $z_{\alpha/2}$ denotes the $(1-\alpha/2)$-th quantile of $\mathcal{N}(0,1)$.

2.4 Numerical studies

In this section, we perform simulation studies to evaluate the numerical performance of the Huber mean estimator and its differentially private counterpart. Regarding the choice of robustification parameter τ, cross-validation provides a viable option but can be computationally expensive and blind to problem structure. Recall from Theorem 2.2.3 that when the fourth moment is finite, the Huber estimator with $\tau \asymp m_4^{1/4}(n/\log n)^{\gamma}$ for any $\gamma \in [1/3, 1/2]$ satisfies the Berry-Esseen bound (2.9) that is of order $m_4^{1/4}(\underline{\lambda} n)^{-1/2}\log n + v_4^{1/2}\{\log(n)/n\}^{3/4}$. Motivated by this, we propose a heuristic data-driven approach to choose τ as described below.

Let $\boldsymbol{\mu}^{(0)} = (1/n)\sum_{i=1}^n \boldsymbol{x}_i$ be an initial estimate. At iteration $t = 1, 2, \ldots$, we take

$$\tau^{(t)} = 0.2 \times \widehat{s}^{(t)} \times \left(\frac{n}{\log n}\right)^{\gamma} \quad \text{with} \quad \widehat{s}^{(t)} = \mathrm{Med}\big(\{\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t-1)}\|_2\}_{i=1}^n\big),$$

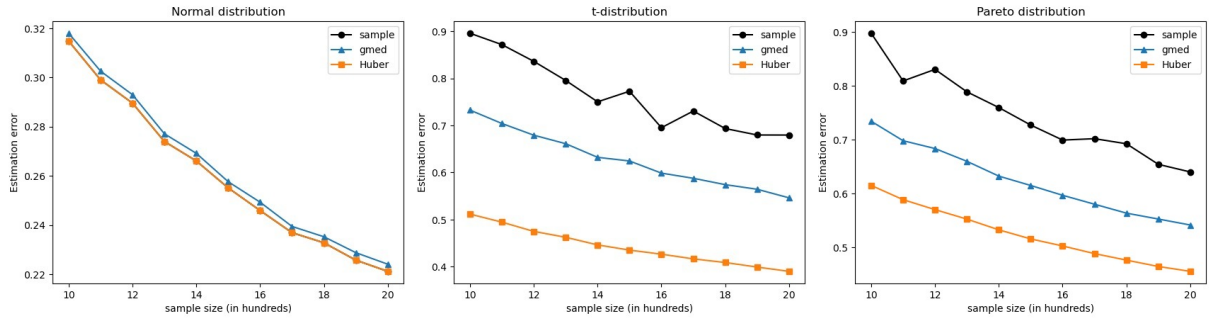


Figure 2.1. Plots of estimation error (under ℓ_2-norm) versus sample size based on 500 repetitions when $d = 100$.

and compute the gradient descent iterate

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \frac{\eta_0}{n} \sum_{i=1}^{n} \frac{\psi_{\tau^{(t)}}(\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t-1)}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t-1)}\|_2} (\boldsymbol{x}_i - \boldsymbol{\mu}^{(t-1)}),$$

where $\eta_0 > 0$ is the step size and $\gamma \in [1/3, 1/2]$. Here, we compute the median of $\{\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t-1)}\|_2\}_{i=1}^n$, which is equivalent to taking the fourth root of the median of $\{\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t-1)}\|_2^4\}_{i=1}^n$, for a robust estimation of $m_4^{1/4} = (\mathbb{E}\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^4)^{1/4}$. Repeat the above two steps until convergence, or until the maximum number of iterations is reached. Since the loss function is locally strongly convex with high probability, we can either use a fixed step size, say $\eta_0 = 1$, or apply the Barzilai-Borwein method (Barzilai and Borwein, 1988) to compute the step size automatically without requiring any parameters. We choose $\gamma = 1/2$ in the following simulation studies. The algorithm for computing the GDP Huber estimator and its confidence interval is provided in the Supplementary Material Yu, Ren and Zhou (2023).

2.4.1 Robust mean estimation and inference

For estimation purposes, we compare the Huber mean estimator, computed by the above algorithm with automatically tuned τ, with the sample mean estimator and the geometric median estimator (gmed) (Minsker, 2015) under the following three distributions, the multivariate normal (lighted-tailed and symmetric), multivariate t (heavy-tailed and symmetric) and Pareto (heavy-tailed and asymmetric).

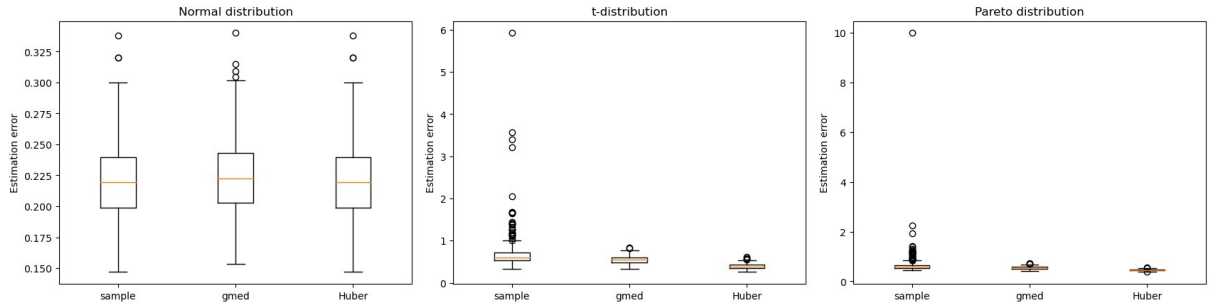


Figure 2.2. Boxplots of estimation error (under ℓ_2-norm) based on 500 repetitions when $(n,d) = (2000, 100)$.

(i) $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^{\mathrm{T}}$ with μ_j's independently drawn from Rademacher distribution, and $\Sigma = (0.8^{|k-l|})_{1 \le k,l \le d}$.

(ii) $\boldsymbol{x}$ follows a multivariate t distribution with 2.1 degrees of freedom. The mean vector $\boldsymbol{\mu}$ is generated the same way as in (i), and the covariance matrix is set to be $\Sigma = 21 * (0.8^{|k-l|})_{1 \le k,l \le d}$.

(iii) $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ has independent coordinates, and each x_j follows a Pareto distribution with shape parameter $\alpha = 2.5$ and scale parameter 1.

We refer to Mathieu (2022) for more comparisons on the estimation errors. For statistical inference, we only compare the proposed robust confidence construction with that of the sample mean. How to construct confidence intervals/sets for other well-known robust mean estimators, such as the geometric median and the geometric median of means, remains an open question.

We fix $d = 100$ and let the sample size n increase from 1000 to 2000. Figure 2.1 depicts the ℓ_2-error versus sample size for the three methods, averaged over 500 repetitions. The Huber estimator is almost identical to the sample mean with normally distributed data, and considerably outperforms the latter for t and Pareto distributed data. The robustness of Huber can be further demonstrated by the boxplot comparison (when $(n,d) = (2000, 100)$) in Figure 2.2. These numerical results provide evidence that the Huber approach gains robustness against heavy-tailedness without compromising efficiency.

Next, we compare the proposed robust confidence intervals (CIs) based on the Huber estimator with the standard CIs constructed from the sample mean and the sample covariance matrix. We fix $(n,d) = (3000, 100)$, and randomly generate a unit vector $\boldsymbol{u}$. The robust 95% CI for $\langle \boldsymbol{u}, \boldsymbol{\mu} \rangle$ takes the form of (2.15) but with $\widehat{\Sigma}_\xi$ replaced by $\widetilde{\Sigma}_\xi$ in (2.12) to reduce computational cost. After obtaining the Huber mean estimator $\widehat{\boldsymbol{\mu}}$, we use $\xi = \widehat{s}\sqrt{n/\log(nd)}$ with $\widehat{s} = \mathrm{Med}(\{\|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}\|_2\}_{i=1}^n)$ to construct the robust covariance estimate. The empirical coverage probabilities and average interval width (with its standard deviation in the parenthesis), averaged over 500 Monte Carlo simulations, are reported in Table **??**. Both methods achieve the nominal coverage under the three distributions, but the robust CIs are consistently narrower and much less variable in the case of heavy-tailed distributions.

In addition, we also conduct a comparative analysis of the performance of the proposed robust multiple CIs against the Bonferroni method and the Šidák method. For $\alpha \in \{0.1, 0.05\}$, we construct robust multiple $100(1-\alpha)\%$ CIs for $\boldsymbol{\mu}$, which take the form of (2.17). For the Bonferroni and Šidák methods, we replace $\omega_{1-\alpha}$ by $z_{1-\alpha/(2d)}$ and $z_{1-\{1-(1-\alpha)^{1/d}\}/2}$, respectively. The empirical coverage probabilities under the multivariate normal and multivariate t-distribution, averaged over 1000 Monte Carlo simulations, are presented in Table 2.2. The multiple CIs based on the uniform Gaussian approximation consistently achieve the nominal coverage. In contrast, the other two methods demonstrate a conservative behavior, indicated by their coverage probabilities surpassing the nominal coverage. Hence, this empirical result supports the assertion that our proposed multiple CIs are less conservative than the Bonferroni and Šidák methods.

2.4.2 Privacy-preserving robust mean estimation and inference

In this subsection, we first examine the numerical performance of the proposed private robust algorithm for mean estimation when $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ consists of i.i.d. $t_{2.1}$-distributed coordinates. The marginal means $\mu_j = \mathbb{E}(x_j)$'s are generated independently from the Rademacher distribution so that $|\mu_j| = 1$ for all $j = 1, \ldots, d$. We fix the initial estimate $\boldsymbol{\mu}^{(0)} = \mathbf{0} \in \mathbb{R}^d$ and step size $\eta_0 = 1$, and set the number of iterations as $T = \lfloor \log n \rfloor$. We implement the private

Table 2.1. Empirical coverage probabilities and average interval widths (with standard deviation in parenthesis) of two normal-based 95% CIs for $\langle \boldsymbol{u}, \boldsymbol{\mu} \rangle$ using the sample mean and the Huber estimator, respectively. The results are based on 500 Monte Carlo simulations when $(n, d) = (3000, 100)$.

	Normal		t		Pareto	
	Coverage	width (sd)	Coverage	width (sd)	Coverage	width (sd)
Sample mean	0.954	0.067 (0.001)	0.944	0.166 (0.076)	0.948	0.101 (0.020)
Huber	0.954	0.067 (0.001)	0.938	0.101 (0.003)	0.954	0.090 (0.002)

Table 2.2. Empirical coverage probabilities of three multiple $100(1-\alpha)\%$ CIs for $\boldsymbol{\mu}$ using the Huber estimator with $\alpha \in \{0.1, 0.05\}$. The results are based on 1000 Monte Carlo simulations when $(n, d) = (3000, 100)$.

	Normal		t	
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$
Proposed CIs	0.905	0.951	0.885	0.945
Bonferroni method	0.933	0.959	0.923	0.957
Šidák method	0.931	0.959	0.918	0.957

	Normal		$t_{2.5}$	
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$
Coverage	0.898	0.960	0.896	0.934

Huber estimator under the following two scenarios.

(i) Fix $d = 64$, let n increase from 10000 to 50000, and set $\varepsilon \in \{0.3, 0.5, 0.9, \infty\}$, the privacy parameter. Here "$\varepsilon = \infty$" corresponds to the non-private Huber estimator.

(ii) Fix $\varepsilon = 0.5$, set $d \in \{32, 64, 128\}$, and let n increase from 10000 to 50000.

The logarithmic ℓ_2-errors ($\log(\|\widehat{\boldsymbol{\mu}}^{(T)} - \boldsymbol{\mu}\|_2)$) versus sample size, averaged over 100 repetitions, are depicted in Figure 2.3. As n increases, the correspondent logarithmic ℓ_2-errors with various privacy parameters differ by a constant. This is consistent with the theoretical rate of convergence stated in Theorem 2.3.1.

Next, we proceed to assess the performance of the proposed robust GDP CIs based on the private robust estimator. We fix the parameters $(n, d) = (50000, 32), \varepsilon = 0.5$, and randomly

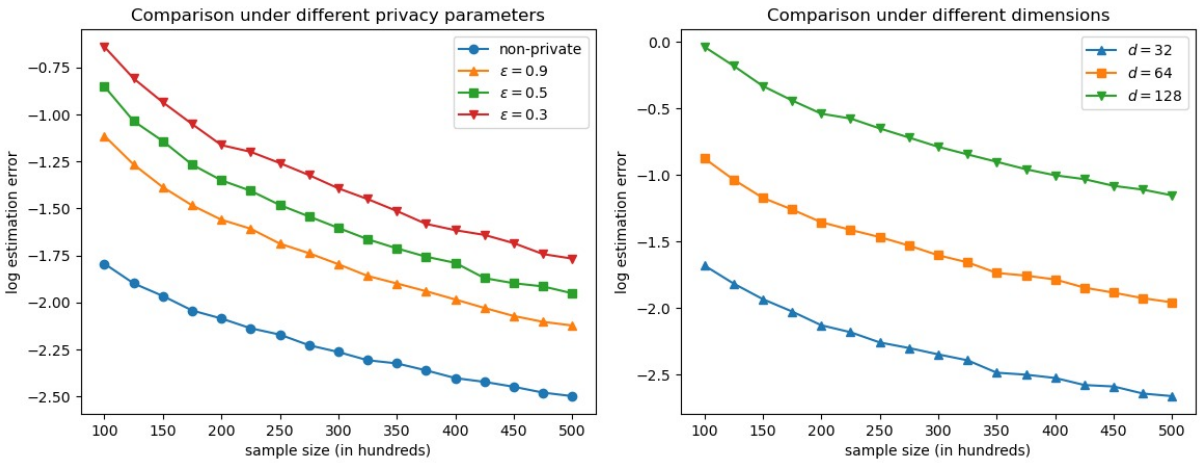


Figure 2.3. Plots of logarithmic ℓ_2-error versus sample size, averaged over 100 repetitions, for the private Huber mean estimator under the $t_{2.1}$ sampling distribution.

generate a unit vector $\boldsymbol{u} \in \mathbb{S}^{d-1}$. For $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^{\mathrm{T}}$ with μ_j's independently drawn from the Rademacher distribution, we generate i.i.d. coordinates x_j's from (i) $\mathcal{N}(0,1)$ and (ii) the t distribution with 2.5 degrees of freedom. We construct the $(\sqrt{2}\varepsilon)$-GDP robust 95% CI for $\langle \boldsymbol{u}, \boldsymbol{\mu} \rangle$ following the formulation outlined in (2.33). However, we replace $\widehat{\Sigma}_{\xi,\varepsilon}$ with the perturbed plug-in covariance estimator outlined in Remark 2.3.4 to reduce computational cost. The empirical coverage probabilities, averaged over 500 Monte Carlo simulations, are presented in Table **??**. The result demonstrates that private confidence intervals achieve nominal coverage as long as the sample size is sufficiently large to compensate for the efficiency loss due to privacy protection.

To highlight the robustness property of the proposed method, we further compare the ε-GDP Huber estimator with the (ε, δ)-DP truncated mean estimator with $\delta = \Phi(-1+\varepsilon/2) - e^{\varepsilon}\Phi(-1-\varepsilon/2)$ (see Algorithm 3.1 in Cai, Wang and Zhang (2021)) under normal and Pareto distributions. For simplicity, we generate independent coordinates x_j's from $\mathcal{N}(0,1)$ and the Pareto distribution with shape parameter $\alpha = 2.1$ and scale parameter 1. We fix $d = 50$, $\varepsilon = 0.5$ (so that $\delta \approx 0.05$), and let the sample size n increase from 10000 to 50000. As before, we set $T = \lfloor \log n \rfloor$ and $\eta_0 = 1$ in the noisy gradient descent algorithm. Note that Algorithm 3.1 in Cai, Wang and Zhang (2021) involves a truncation tuning parameter R. For normal distributions, we use the theoretically optimal choice $R = 4\sqrt{\log n}$ as suggested in Cai, Wang and Zhang (2021);

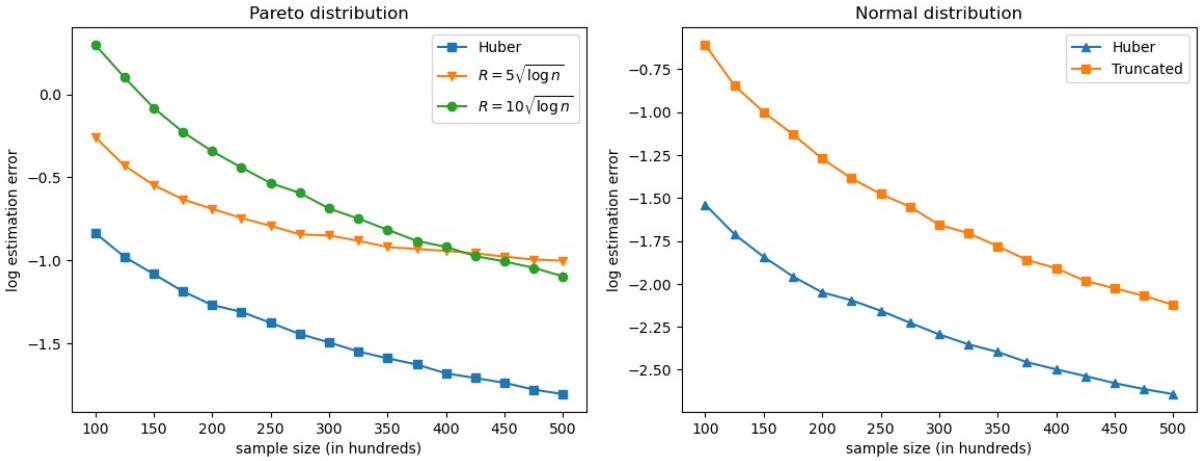


Figure 2.4. Plots of logarithmic ℓ_2-error versus sample size, averaged over 100 repetitions, for the ε-GDP Huber estimator and (ε, δ)-DP truncated mean estimator (Cai, Wang and Zhang, 2021) when $d = 50$.

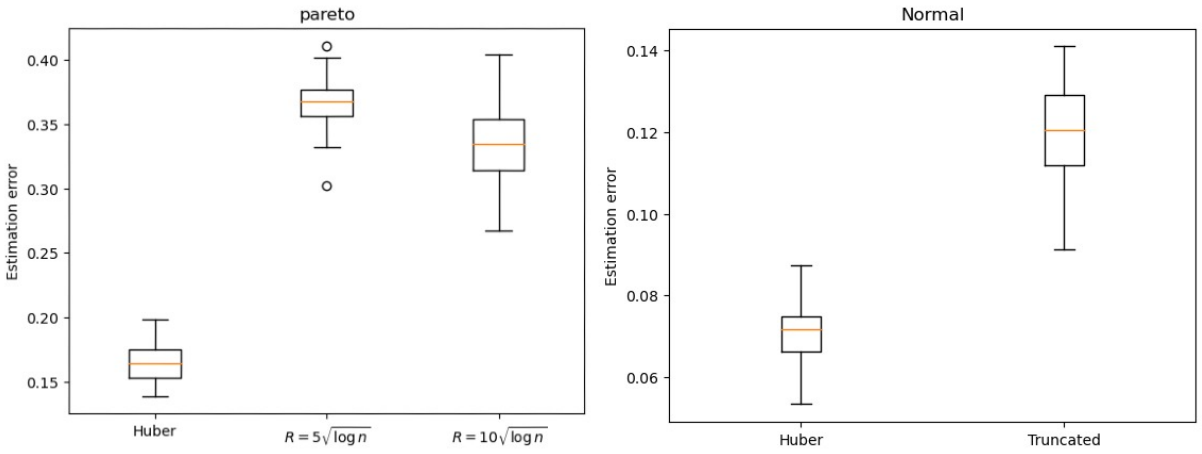


Figure 2.5. Boxplots of logarithmic ℓ_2 error based on 100 repetitions for the ε-GDP Huber estimator and (ε, δ)-DP truncated mean estimator (Cai, Wang and Zhang, 2021) when $(n, d) = (50000, 50)$.

for the heavy-tailed Pareto distribution, there is no theoretical guidance for choosing R. We thus take $R \in \{5\sqrt{\log n}, 10\sqrt{\log n}\}$ in this case.

Figures 2.4 and 2.5 show that the two methods perform similarly in the normal case. Interestingly, the private Huber estimator does exhibit a visible improvement. In the heavy-tailed case (Pareto distribution), the private Huber method considerably outperforms the noisy truncated sample mean, at least under the prespecified truncation levels. Together, the numerical results in Sections 2.4.1 and 2.4.2 provide strong evidence that the Huber mean estimator, either non-private or private, achieves a high degree of robustness against heavy-tailedness while maintaining high efficiency under light-tailed (e.g., sub-Gaussian) distributions.

2.5 Proofs of main results

2.5.1 Proof of Theorem 2.2.1

For simplicity, we write $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_\tau$. For some $r > 0$ to be determined, define $\widetilde{\boldsymbol{\mu}} = (1-u)\boldsymbol{\mu} + u\widehat{\boldsymbol{\mu}}$, where $u = \sup\{t \in [0,1] : t(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \in \Theta(r)\}$. By this definition, $u = 1$ if $\widehat{\boldsymbol{\theta}} \in \Theta(r)$, and $u \in (0,1)$ otherwise. For the latter, $\widetilde{\boldsymbol{\mu}} \in \partial\Theta(r)$.

Since $\widehat{\boldsymbol{\mu}}$ minimizes the convex objection function $\widehat{\mathscr{L}_\tau}(\cdot)$, the first-order condition holds, that is, $\nabla\widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) = \mathbf{0}$. Further, applying Lemma C.1 in the supplementary material of Sun, Zhou and Fan (2020) implies

$$\langle \nabla\widehat{\mathscr{L}_\tau}(\widetilde{\boldsymbol{\mu}}) - \nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}), \widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\rangle \le u\langle \nabla\widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) - \nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}), \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\rangle \le \|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu})\|_2 \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2.$$

For the left-hand side, since $\widetilde{\boldsymbol{\mu}} \in \Theta(r)$, it follows from the mean value theorem that

$$\langle \nabla\widehat{\mathscr{L}_\tau}(\widetilde{\boldsymbol{\mu}}) - \nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}), \widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\rangle \ge \inf_{\boldsymbol{\theta}\in\Theta(r)} \lambda_{\min}\big(\nabla^2\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta})\big) \cdot \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2,$$

where $\lambda_{\min}(\nabla^2\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}))$ is the smallest eigenvalue of $\nabla^2\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta})$. For any $z > 0$ and $r < \tau$, Lemma

D.1 in Yu, Ren and Zhou (2023) implies that, with probability at least $1-e^{-z}$,

$$1-\mathbb{P}\big(\|\boldsymbol{x}-\boldsymbol{\mu}\|_2>\gamma\big)-\sqrt{\frac{z}{2n}}\le \boldsymbol{u}^{\mathrm{T}}\nabla^2\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta})\boldsymbol{u}\le 1 \tag{2.34}$$

holds uniformly over $\boldsymbol{\theta}\in\Theta(r)$ and $\boldsymbol{u}\in\mathbb{S}^{d-1}$, where $\gamma=\tau-r$ and $\Theta(r)$ is defined in (2.23). Furthermore, by Lemma D.2 in Yu, Ren and Zhou (2023), we have

$$\|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu})\|_2\le 2\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}+\sqrt{\frac{2\|\Sigma\|_2 z}{n}}+\frac{4\tau z}{3n}+b_\tau \tag{2.35}$$

with probability at least $1-e^{-z}$. Therefore, denoting $\mathscr{G}_z$ to be the event that (2.34) and (2.35) hold, $\mathscr{G}_z$ occurs with probability at least $1-2e^{-z}$. By Markov's inequality, $\mathbb{P}(\|\boldsymbol{x}-\boldsymbol{\mu}\|_2>\gamma)\le\gamma^{-2}\,\mathrm{tr}(\Sigma)$. Then, conditioned on $\mathscr{G}_z$, the above upper and lower bounds yield

$$\left(1-\gamma^{-2}\,\mathrm{tr}(\Sigma)-\sqrt{\frac{z}{2n}}\right)\cdot\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2^2\le\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2\left\{2\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}+\sqrt{\frac{2\|\Sigma\|_2 z}{n}}+b_\tau+\frac{4\tau z}{3n}\right\}.$$

This, combined with the local constraint $\widetilde{\boldsymbol{\mu}}\in\Theta(r)$, implies

$$\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2\le 2\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}+\sqrt{\frac{2\|\Sigma\|_2 z}{n}}+b_\tau+\frac{4\tau z}{3n}+r\cdot\left\{\frac{\mathrm{tr}(\Sigma)}{\gamma^2}+\sqrt{\frac{z}{2n}}\right\}.$$

To conclude the proof, note from Lemma D.2 in Yu, Ren and Zhou (2023) that $b_\tau\le\tau^{-1}\sqrt{\|\Sigma\|_2\,\mathrm{tr}(\Sigma)}$. Taking $r=\gamma=\tau/2$, and let (n,τ) satisfy $n\gtrsim \mathrm{r}(\Sigma)+z$ and $\gamma\gtrsim\sqrt{\mathrm{tr}(\Sigma)}$, the right-hand side of the above inequality is strictly less than r, indicating that $\widetilde{\boldsymbol{\mu}}$ falls in the interior of $\Theta(r)$. Via proof by contradiction, we reach the conclusion $\widehat{\boldsymbol{\mu}}=\widetilde{\boldsymbol{\mu}}\in\Theta(r)$ (otherwise $\widetilde{\boldsymbol{\mu}}$ must be on the boundary of $\Theta(r)$), and hence the same bound applies to $\widehat{\boldsymbol{\mu}}$. □

2.5.2 Proof of Theorem 2.2.2

For $\boldsymbol{h} \in \mathbb{R}^d$, define the function $\Delta(\boldsymbol{h}) = \nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}+\boldsymbol{h}) - \nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}) - \boldsymbol{h}$. By the mean value theorem for vector-valued functions,

$$\Delta(\boldsymbol{h}) = \int_0^1 \nabla^2\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}+t\boldsymbol{h})\mathrm{d}t\cdot\boldsymbol{h} - \boldsymbol{h} = \int_0^1 \big\{\nabla^2\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}+t\boldsymbol{h}) - \mathbf{I}_d\big\}\mathrm{d}t\cdot\boldsymbol{h}.$$

Hence, for any $r>0$, we have $\sup_{\|\boldsymbol{h}\|_2\le r}\|\Delta(\boldsymbol{h})\|_2 \le \sup_{\boldsymbol{\theta}\in\Theta(r)}\|\nabla^2\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) - \mathbf{I}_d\|_2\cdot r$. This together with Lemma D.1 in Yu, Ren and Zhou (2023) implies that, with probability at least $1-e^{-z}$,

$$\sup_{\|\boldsymbol{h}\|_2\le r}\big\|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}+\boldsymbol{h}) - \nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}) - \boldsymbol{h}\big\|_2 \le r\bigg(\gamma^{-q}\mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^q + \sqrt{\frac{z}{2n}}\bigg), \tag{2.36}$$

where $\gamma = \tau - r$.

For simplicity, we write $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_\tau$. Setting $\widehat{\boldsymbol{h}} = \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}$, Theorem 2.2.1 ensures that $\|\widehat{\boldsymbol{h}}\|_2 \le r_0$ with $r_0 \asymp \sqrt{\{\mathrm{tr}(\Sigma) + \|\Sigma\|_2 z\}/n} + \tau z/n + b_\tau$ with probability at least $1-2e^{-z}$, provided $n \gtrsim \mathrm{r}(\Sigma) + z$ and $\tau \gtrsim \sqrt{\mathrm{tr}(\Sigma)}$. Note that the gradient of the empirical loss $\widehat{\mathscr{L}_\tau}(\cdot)$ is given by

$$\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^n \frac{\psi_\tau(\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2)}{\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2}(\boldsymbol{x}_i-\boldsymbol{\theta}) \tag{2.37}$$

for $\boldsymbol{\theta}\in\mathbb{R}^d$. Taking $r = r_0$, the claimed bound (2.8) follows from (2.36), (2.37) and the fact that $\nabla\widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) = \mathbf{0}$. □

2.6 Acknowledgements

This chapter, in full, is a reprint of the material in the paper "Gaussian differentially private robust mean estimation and inference", Yu, Myeonghun, Ren, Zhao and Zhou, Wen-Xin. The paper has been accepted by *Bernoulli*, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Deep Neural Network Expected Shortfall Regression with Heavy-tailed Data

3.1 Introduction

Expected shortfall (ES), also known as conditional value-at-risk or superquantile, is defined as the expected value of a random variable, given that its realization falls below some quantile of the underlying distribution. Initially introduced as a risk measure by Artzner et al. (1997), ES has gained widespread recognition and applicability across various disciplines, including finance (Acerbi and Tasche, 2002; Rockafellar and Uryasev, 2002), operations research (Rockafellar and Uryasev, 2000; Rockafellar et al., 2008), engineering (Rockafellar and Royset, 2010), and clinical studies (He et al., 2010), among others. Notably, in the recent Fundamental Review of the Trading Book (Basel Committee, 2019), the Basel Committee on Banking Supervision confirmed the replacement of the value at risk (quantile) with ES as the standard risk measure for market risk. Furthermore, in the context of insurance regulation, ES has been adopted as a risk measure in the Swiss Solvency Test.

Formally, let Y be a real-valued random variable, denoting for example the return of an asset or investment portfolio. Let F_Y be its cumulative distribution function (CDF). Denote the quantile of Y at level $\alpha \in (0,1)$ by $q_\alpha(Y) := \inf\{y \in \mathbb{R} : F_Y(y) \geq \alpha\}$. Provided that $\mathbb{E}|Y| < \infty$,

the ES of Y at level α is defined as

$$e_\alpha(Y) := \mathbb{E}\{Y|Y \leq q_\alpha(Y)\} = \frac{1}{\alpha}\mathbb{E}[Y\mathbb{1}\{Y \leq q_\alpha(Y)\}].$$

Intuitively, the α-level ES refers to the average of the lowest $(100 \cdot \alpha)\%$ portion of Y, rescaled by $1/\alpha$. If F_Y is continuous at $q_\alpha(Y)$, the α-level ES can be equivalently expressed as $e_\alpha(Y) = \frac{1}{\alpha}\int_0^\alpha q_u(Y)\mathrm{d}u$. We refer to Sections 2.2.4 of McNeil et al. (2015) for a brief introduction to the expected shortfall and its basic properties.

In the presence of covariates $X \in \mathbb{R}^d$, the objective of this study is to estimate the conditional ES of Y given X, using a sample $\{(X_i, Y_i)\}_{i=1}^n$ of size n. One challenge of this problem lies in the fact that ES is not elicitable (Gneiting, 2011), meaning that there is no loss function such that ES is the unique minimizer of the expected loss. To tackle this challenge, Fissler and Ziegel (2016) demonstrated that ES is jointly elicitable with the quantile by constructing a class of joint loss functions that are strictly consistent. Expanding on this important property, Dimitriadis and Bayer (2019) introduced a joint linear regression framework for modeling conditional quantile and ES, while Patton et al. (2019) considered a semi-parametric model in the autoregressive context. From an alternative perspective that regards the (conditional) quantile as a nuisance parameter, Barendse (2020) and Peng and Wang (2023) each proposed two-step estimators and established their asymptotic properties under the fixed-d regime. Although their definitions differ, both methods rely on an orthogonality property, as we will revisit in Section 3.3.

In practice, the relationship between the response variable Y and covariates X often displays a high degree of nonlinearity, requiring the use of nonparametric regression techniques. To estimate nonlinear conditional ES functions, Scaillet (2005) employed the Nadaraya-Watson estimator to estimate the conditional CDF of Y given X in the initial stage, followed by the estimation of conditional ES functions. Furthermore, Cai and Wang (2008) and Kato (2012) employed weighted Nadaraya-Watson estimators to estimate conditional CDFs and ES functions. As the dimensionality of the covariate space increases, the amount of data required to obtain

accurate estimations using the Nadaraya-Watson estimator grows exponentially. This is because the estimator involves weighting each data point based on its distance from the point being estimated. In higher dimensions, the "neighborhood" of a given point becomes sparser, making it challenging to find enough nearby points for accurate weighting. Consequently, the estimator suffers from reduced accuracy and efficiency when applied to moderate-dimensional data.

In the last decade, deep learning has achieved remarkable success and emerged as an indispensable tool for analyzing nonlinear relationships between various types of outcomes and explanatory variables. With the availability of vast amounts of digitized data and the development of efficient computational algorithms, deep neural networks (DNNs) have become widely used and consistently outperformed traditional methods in various machine learning tasks, as exemplified by natural language processing (Otter, Medina and Kalita, 2021) and image classification (Krizhevsky, Sutskever and Hinton, 2017). More recently, DNNs have also demonstrated their exceptional performance in forecasting climate data. By elucidating intricate nonlinear relationships with a variety of explanatory variables, DNN-based methods have shown remarkable accuracy in predicting El Niño–Southern Oscillation, precipitation and temperature (Huang, Vega-Westhoff and Sriver, 2019; Jose, Vincent and Dwarakish, 2022; Wang et al., 2023).

From a statistical viewpoint, the success of DNNs can be attributed, in part, to their ability to effectively approximate various complex functions. In particular, recent studies (Bauer and Kohler, 2019; Schmidt-Hieber, 2020; Kohler and Langer, 2021) have shown that DNN-based regression estimators can adapt to the intrinsic low-dimensional structure of the conditional mean function, enabling them to circumvent the curse of dimensionality. Specifically, when the conditional mean function can be represented as a hierarchical composition of several smooth functions, with either a high degree of smoothness or low input dimension, DNN-based estimators demonstrate an ability to adapt to the intrinsic low-dimensional structure of the regression function. Moreover, DNNs have also found successful applications in estimating the nonlinear component of semi-parametric models to mitigate the curse of dimensionality. This enables the construction of statistically efficient inferences on the linear component of the model

(Farrell et al., 2021; Zhong et al., 2022; Zhong and Wang, 2023).

In the aforementioned literature on deep neural network regression, the response or noise variable is required to be either bounded or sub-Gaussian, which can be a fairly stringent assumption in practice. In light of this issue, our aim is to develop a robust nonparametric estimator of the conditional ES function that exhibits resilience in the presence of heavy-tailed error distributions. It is worth noting that most of the extant studies on ES estimation primarily focused on deriving the asymptotic properties of their estimators. Consequently, the asymptotic results they provided can only yield polynomial-type (high probability) deviation bounds. One notable exception is He et al. (2023), where the authors proposed a robust ES estimator that exhibits insensitivity to heavy-tailed noise under joint linear conditional quantile and ES models. They showed that the robust estimator outperforms the least-squares-type estimator from a non-asymptotic perspective. Under such linear models, we note that both two-stage estimators exhibit a convergence rate of $\mathscr{O}(\sqrt{d/n})$ under expectation, where n is the number of observations.

In the context of nonparametric regression with heavy-tailed errors, there has been a growing interest in recent times. Notably, several recent works (Han and Wellner, 2018, 2019; Kuchibhotla and Patra, 2022) have addressed the impact of heavy-tailed errors on the convergence rate of LSEs that are constrained to the nonparametric function class to which the true conditional mean function belongs. Due to the lack of robustness of LSEs, alternative robust methods have been developed to address this issue, particularly when neural networks are employed. For example, Shen et al. (2021); Padilla et al. (2022) focused on nonparametric robust regression using the check loss, establishing the convergence rate of DNN quantile regression estimators. Fan et al. (2022) considered nonparametric adaptive Huber regression and demonstrated that the corresponding robust estimator achieves a faster convergence rate compared to the LSE when the noise is heavy-tailed.

In this paper, we introduce a robust two-stage method for estimating the ES regression function using DNNs. Building upon the approach of Barendse (2020), the proposed method involves estimating the quantile regression (QR) function through any machine learning method

in the first stage, followed by nonparametric adaptive Huber regression (Fan et al., 2022) with generated response variables in the second stage; see Section 3.3.2 for a rigorous construction of the method. By employing neural networks, the estimator can adapt to the unknown hierarchical composition structure of the true conditional ES function. Furthermore, the proposed method demonstrates more robust behavior compared to a two-stage LSE in the presence of heavy-tailed errors. The main contributions of this work are summarized as follows.

First, we establish a non-asymptotic deviation bound on the L_2-error of the proposed robust estimator. In detail, given a first-stage QR estimator, we establish an oracle-type upper bound for approximate empirical risk minimizers trained on a neural network of arbitrary depth and width, using a Huber loss with a reasonably large robustification parameter τ. The resulting L_2-error bound is comprised of six distinct terms. Notably, by using orthogonal score functions, which are locally insensitive to first-stage QR estimators, the deviation bound is first-order negligible with respect to the L_4-error of the QR estimator. For comparisons, oracle-type deviation bounds are also presented for a two-stage LSE for the conditional ES.

Secondly, we derive a novel approximation error bound for a composition of Hölder smooth functions using ReLU-activated DNNs. This work builds upon the results of Jiao et al. (2023) and Fan et al. (2022). Jiao et al. (2023) derived an approximation error bound for a Hölder smooth function with the smoothness index $\beta \geq 1$, and its prefactor depends on the input dimension polynomially. Fan et al. (2022) derived an approximation error bound for a composition of Hölder smooth functions, which depends on the intrinsic dimension of the composite function. However, the prefactor of their approximation error bound depends exponentially on the input dimensions of the components of the composite function. We establish an approximation error bound for composite functions that depends on the intrinsic dimension, mitigating the curse of dimensionality to some extent. Furthermore, the prefactor of the approximation bound significantly improves, exhibiting polynomial dependence on the input dimensions of the component functions instead of exponential. Applying this approximation result, we are able to establish an exponential-type deviation bound for a nonparametric QR

estimator using DNNs, where the prefactor of this bound depends on the input dimension polynomially.

To derive a specific error bound for the two-stage ES estimator based on the oracle result, we need to select appropriate values for the depth and width of neural networks, the robustification parameter τ, and choose a suitable QR estimator. By employing DNNs to estimate the QR function and balancing the terms comprising the L_2 bound, we establish convergence rates for both the robust two-stage estimator and two-stage LSE in cases where the true quantile and ES regression functions are compositions of smooth functions. These results demonstrate that both estimators effectively overcome the curse of dimensionality, as their convergence rates depend solely on the intrinsic dimension while adapting to the underlying compositional structure. In the presence of heavy-tailed errors, the robust estimator outperforms the LSE by achieving a faster convergence rate. Moreover, from a non-asymptotic viewpoint, the robust two-stage estimator exhibits exponential-type deviation bounds, whereas the LSE only exhibits polynomial-type L_2-error bounds in high probability. Finally, by applying our new approximation results, the prefactors of error bounds for both estimators exhibit a polynomial dependence on the dimensions.

NOTATION. We use $c_1, c_2, \ldots$ to denote the global constants employed in the statements and proofs of theorems, propositions, corollaries, and lemmas. We use $C_1, C_2, \ldots$ to denote the local intermediate constants within the proof. Thus, each $c_1, c_2, \ldots$ has a distinct referred numbers, while $C_1, C_2, \ldots$ may vary from one line to another. We write $a \lesssim b$ if there exists an absolute constant $C > 0$ such that $a \leq Cb$, and $a \gtrsim b$ if $b \lesssim a$. Moreover, we write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. We denote $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ and $\mathbb{N}^+ = \{1, 2, \ldots\}$ to be the sets of nonnegative integers and positive integers, respectively. For any real-valued function h defined on a domain $\mathscr{X}$, we denote the supremum norm of h over $\mathscr{X}$ as $\|h\|_\infty$. For the sample size n, we always assume $n \geq 3$ so that $\log n \geq 1$.

3.2 Model Setup and Preliminaries

3.2.1 Model

Let $\{(Y_i, X_i)\}_{i=1}^n$ be a collection of independent observations from the random variable $(Y, X) \in \mathbb{R} \times \mathscr{X}$, where $\mathscr{X} \subseteq \mathbb{R}^d$ is a compact subset. Here, Y denotes a real-valued response variable and X represents a d-dimensional vector of covariates that follows some distribution $\mathbb{P}_X$. Without loss of generality, we assume $\mathscr{X} = [0,1]^d$, the unit cube in $\mathbb{R}^d$, throughout the following. At some probability level $\alpha \in (0,1)$ of interest, we denote the conditional α-level quantile and expected shortfall of Y given the covariates X as $q_\alpha(Y|X)$ and $e_\alpha(Y|X)$, respectively. Here, the conditional ES is formally defined as $e_\alpha(Y|X) = \mathbb{E}\{Y | Y \leq q_\alpha(Y|X), X\}$. We consider the following nonparametric joint quantile and ES regression model:

$$q_\alpha(Y_i|X_i) = f_0(X_i) \text{ and } e_\alpha(Y_i|X_i) = g_0(X_i), \tag{3.1}$$

where $f_0, g_0 : [0,1]^d \to \mathbb{R}$ are two unknown functions satisfying $\mathbb{P}\{Y \leq f_0(X) | X = \boldsymbol{x}\} = \alpha$ and $g_0(\boldsymbol{x}) = \alpha^{-1}\mathbb{E}[Y \mathbb{1}\{Y \leq f_0(X)\} | X = \boldsymbol{x}]$ for $\boldsymbol{x} \in [0,1]^d$.

Our primary object is to propose a fully nonparametric estimator $\widehat{g}$ of the function g_0, and derive its rate of convergence under L_2-norm $\|\cdot\|_2$, defined as $\|h\|_2 := \|h\|_{\mathbb{P}_X,2} = \sqrt{\mathbb{E}_{X\sim\mathbb{P}_X}|h(X)|^2}$ for any $h : [0,1]^d \to \mathbb{R}$. Imposing smoothness assumption on the regression function is essential to derive meaningful insights regarding the rate of convergence. Thus, we begin by introducing the following definition of Hölder smooth classes.

Definition 3.2.1 (Hölder class of functions $\mathscr{H}^\beta(\mathscr{X}, M_0)$)**.** Let $\beta = r + s$ for a nonnegative integer $r = \lfloor \beta \rfloor$ and $0 < s \leq 1$, where $\lfloor a \rfloor$ denotes the largest integer that is strictly smaller than $a \in \mathbb{R}$. Given a subset $\mathscr{X} \subseteq \mathbb{R}^d$ and a constant $M_0 > 0$, a function $f : \mathscr{X} \to \mathbb{R}$ is called (β, M_0)-smooth on $\mathscr{X}$ if for every $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^{\mathrm{T}} \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j \leq r$, the partial derivative $\partial^{\boldsymbol{\alpha}} f = (\partial f)/(\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d})$ exists and satisfies $\max_{\|\boldsymbol{\alpha}\|_1 \leq r} \|\partial^{\boldsymbol{\alpha}} f\|_\infty \leq M_0$ and $\max_{\|\boldsymbol{\alpha}\|_1 = r} \sup_{\boldsymbol{x}_1 \neq \boldsymbol{x}_2} |\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}_1) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}_2)| / \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^s \leq M_0$, where $\|\boldsymbol{\alpha}\|_1 = \sum_{j=1}^d \alpha_j$. We then

use $\mathscr{H}^{\beta}(\mathscr{X}, M_0)$ to denote collection of all (β, M_0)-smooth functions on $\mathscr{X}$.

Throughout the paper, we assume $M_0 \geq 1$ without loss of generality. Moreover, note that the definition of Hölder class implies that if a function f belongs to $\mathscr{H}^{\beta}(\mathscr{X}, M_0)$, then f is bounded in magnitude by M_0. This can be derived by considering $\boldsymbol{\alpha} = \mathbf{0}$ in the definition.

Nonparametric estimation of a function within Hölder classes exhibits significantly slower convergence rates as the dimension d becomes large. For example, it has been well established that the minimax rate of convergence for estimating a mean regression function within $\mathscr{H}^{\beta}(\mathscr{X}, M_0)$ is of order $n^{-\beta/(2\beta+d)}$ (Stone, 1982). This phenomenon is commonly recognized as the curse of dimensionality. In order to circumvent the curse of dimensionality, we focus on functions that have a compositional structure, also known as the hierarchical interaction model (Bauer and Kohler, 2019; Kohler and Langer, 2021).

Definition 3.2.2 (Hierarchical interaction model)**.** Let $l, d \in \mathbb{N}^+$, $M_0 \geq 1$ and $\mathscr{P}$ be a subset of $[1, \infty) \times \mathbb{N}^+$ with $\sup_{(\beta,t)\in\mathscr{P}}(\beta \vee t) < \infty$. The hierarchical interaction model $\mathscr{H}(d, l, M_0, \mathscr{P})$ is defined recursively as follows.

(i) We say that a function $h : \mathbb{R}^d \to \mathbb{R}$ satisfies the model $\mathscr{H}(d, 1, M_0, \mathscr{P})$ if there exist some $(\beta, t) \in \mathscr{P}$, $h_0 \in \mathscr{H}^{\beta}(\mathbb{R}^t, M_0)$ and $\{j_1, \ldots, j_t\} \subseteq \{1, \ldots, d\}$ such that $h(\boldsymbol{x}) = h_0(x_{j_1}, \ldots, x_{j_t})$ for all $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$.

(ii) For $l > 1$, we say that a function $h : \mathbb{R}^d \to \mathbb{R}$ satisfies the hierarchical interaction model $\mathscr{H}(d, l, M_0, \mathscr{P})$ if there exist some $(\beta, t) \in \mathscr{P}$ with $h_0 \in \mathscr{H}^{\beta}(\mathbb{R}^t, M_0)$ and $u_1, \ldots, u_t \in \mathscr{H}(d, l-1, M_0, \mathscr{P})$ such that $h(\boldsymbol{x}) = h_0(u_1(\boldsymbol{x}), \ldots, u_t(\boldsymbol{x}))$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

As discussed in Kohler and Langer (2021), this general model encompasses various well-known nonparametric and semiparametric models, including additive models (Stone, 1985), single index models (Härdle et al., 1993) and the projection pursuit (Friedman and Stuetzle, 1981). Extensive research (Bauer and Kohler, 2019; Kohler and Langer, 2021; Schmidt-Hieber, 2020) has established that the minimax optimal convergence rate for the hierarchical composition

model is determined by the most challenging (least smooth) component within the composition. This challenging component is characterized by the quantity

$$\gamma^* = \frac{\beta^*}{t^*}, \text{ where } (\beta^*, t^*) = \underset{(\beta,t)\in\mathscr{P}}{\operatorname{argmin}} \frac{\beta}{t}. \tag{3.2}$$

We refer to the ratio β/t as the dimension-adjusted degree of smoothness.

3.2.2 ReLU neural networks

Our proposed nonparametric estimators for joint quantile and ES regression are constructed using truncated fully-connected deep neural networks with the rectified linear unit (ReLU) activation function, denoted as $\sigma(\cdot) = \max(\cdot, 0)$. These networks are succinctly referred to as truncated deep ReLU neural networks. To provide a brief introduction, we begin by examining the structure of a fully-connected DNN. We introduce two positive integer parameters: a depth parameter L and a width parameter N. Define a class of deep ReLU neural networks, represented as $\mathscr{F}_{\text{DNN}}(d, L, N)$, which consists of functions $f : \mathbb{R}^d \to \mathbb{R}$ that can be expressed as $f(\boldsymbol{x}) = \mathscr{L}_{L+1} \circ \sigma \circ \mathscr{L}_L \circ \sigma \circ \cdots \mathscr{L}_2 \circ \sigma \circ \mathscr{L}_1(\boldsymbol{x})$. Each $\mathscr{L}_l$ denotes an affine transformation, that is, $\mathscr{L}_l(\boldsymbol{x}) = W_l\boldsymbol{x} + b_l$, where $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ denotes the weight matrix, $b_l \in \mathbb{R}^{d_l}$ denotes the bias vector, and $(d_0, d_1, \cdots, d_L, d_{L+1}) = (d, N, \cdots, N, 1)$ is the width vector of layers. When $\boldsymbol{x}$ is a vector, the ReLU function $\sigma(\boldsymbol{x})$ is defined by applying $\sigma(\cdot)$ to each element of $\boldsymbol{x}$.

Next, for any $M > 0$, we define a truncated ReLU neural network as

$$\mathscr{F}_{\text{DNN}}(d, L, N, M) = \mathscr{T}_M \mathscr{F}_{\text{DNN}}(d, L, N) = \{\mathscr{T}_M h : h \in \mathscr{F}_{\text{DNN}}(d, L, N)\},$$

where the truncated function $\mathscr{T}_M h$ is given by $(\mathscr{T}_M h)(\boldsymbol{x}) = \text{sgn}(h(\boldsymbol{x}))(|h(\boldsymbol{x})| \wedge M)$.

The following result provides an error bound for approximating functions within a hierarchical interaction model using truncated deep ReLU neural networks. For a given index set $\mathscr{P} \subseteq [1, \infty) \times \mathbb{N}^+$, recall the definition of γ^* in (3.2).

Proposition 3.2.1 (Neural network approximation error for $\mathscr{H}(d,l,M_0,\mathscr{P})$)**.** Given a hierarchical interaction model $\mathscr{H}(d,l,M_0,\mathscr{P})$, there exist universal constants c_1–c_3 such that, for any $L_0,N_0 \geq 3$ and a measure μ on $[0,1]^d$ that is absolutely continuous with respect to the Lebesgue measure, it holds

$$\sup_{f_0 \in \mathscr{H}(d,l,M_0,\mathscr{P})} \inf_{f^* \in \mathscr{F}_{\mathrm{DNN}}(d,L,N,M_0)} \left\{ \int_{[0,1]^d} |f^*(\boldsymbol{x}) - f_0(\boldsymbol{x})|^2 \mu(\mathrm{d}\boldsymbol{x}) \right\}^{1/2} \leq c_3 (L_0 N_0)^{-2\gamma^*},$$

where $L = c_1 \lceil L_0 \log L_0 \rceil$ and $N = c_2 \lceil N_0 \log N_0 \rceil$, and $\lceil a \rceil$ denotes the smallest integer no less than $a \in \mathbb{R}$. Here, the constants c_1–c_3 depend on $t_{\max} = \max_{(\beta,t)\in\mathscr{P}} t$ polynomially.

It is worth noting that the above approximation result holds for general neural networks without imposing any structural assumptions. Proposition 3.2.1 demonstrates the validity of the approximation results across a wide range of neural networks, irrespective of sparsity or boundedness of the network weights, or a specific architectural characteristic, such as being thin and deep or wide and shallow. In comparison to the result in Fan et al. (2022), our approximation error bound features a polynomial dependence on $t_{\max}$ through the prefactor c_3. Specifically, our prefactor c_3 depends on $t_{\max}$ through the expression $t_{\max}^{\lfloor \beta_{\max} \rfloor + \beta_{\max}/2}(1 + M_0 t_{\max}^{1/2})^{l-1}$, where $\beta_{\max} = \max_{(\beta,t)\in\mathscr{P}} \beta$. In contrast, the prefactor of the approximation bound in Proposition 3.4 of Fan et al. (2022) depends on $t_{\max}$ through $(\lfloor \beta_{\max} \rfloor + 1)^{t_{\max}}(1 + M_0 t_{\max}^{1/2})^{l-1}$. Hence, our approximation error bound is more favorable when $t_{\max}$ is larger than $\beta_{\max}$, while still being comparable to the result of Fan et al. (2022) if $\beta_{\max}$ and $t_{\max}$ are of similar magnitudes. Nevertheless, it should be noted that Proposition 3.2.1 establishes an L_2 approximation error bound, while Fan et al. (2022) derived a uniform (L_∞) bound. By applying a similar line of arguments in the proof of Propositon 3.2.1 combined with Corollary 3.1 of Jiao et al. (2023), we can derive an L_∞-approximation error bound that also features a polynomial dependence on $t_{\max}$. However, this comes at the cost of necessitating an increase of the network width N. In detail, our prefactor c_2 of the network width in Proposition 3.2.1 depends on $t_{\max}$ through the expression $(\lfloor \beta_{\max} \rfloor + 1)^2 t_{\max}^{\lfloor \beta_{\max} \rfloor + l}$, whereas the prefactor of the network width required for

L_∞ bound will depend on $t_{\max}$ exponentially through $(\lfloor \beta_{\max} \rfloor + 1)^2 t_{\max}^{\lfloor \beta_{\max} \rfloor + l} 3^{t_{\max}}$. Therefore, if we employ these neural networks with enlarged network width to define an estimator, the error bound will exhibit exponential dependence on $t_{\max}$; see Theorem 3.4.1 and Theorem 3.4.4. The exact values of c_1–c_3 are specified in the proof of Proposition 3.2.1.

3.3 Nonparametric Expected Shortfall Regression

In this section, we begin with a brief review of the joint loss minimization framework introduced in Fissler and Ziegel (2016) and its limitations. Then we introduce a generic two-step nonparametric ES regression estimator that uses an orthogonal score function to reduce sensitivity to the estimation error of a quantile regression estimate in the first stage. Subsequently, we propose a robust approach for estimating the conditional ES function in the presence of heavy-tailed errors. A non-asymptotic (finite-sample) theory for the proposed estimators will be established in Section 3.4.

Following Fissler and Ziegel (2016), let us consider a class of strictly consistent joint loss functions for the pair of quantile and ES (with slight modifications)

$$\begin{aligned} L_\alpha(q,e;Y) = {} & \{\alpha - \mathbb{1}(Y \le q)\}\{G_1(Y) - G_1(q)\} \\ & - \{\underbrace{\alpha q + \alpha(Y-q)\mathbb{1}(Y \le q) - \alpha e}_{=: S_\alpha(q,e;Y)}\} G_2(e)/\alpha - \mathscr{G}_2(e), \quad e \le q, \end{aligned} \tag{3.3}$$

where G_1 is an increasing and integrable function, $\mathscr{G}_2$ is a three-times continuously differentiable function such that both $G_2 = \mathscr{G}_2'$ and G_2' are strictly positive. With the data $\{(Y_i, X_i)\}_{i=1}^n$, we obtain the nonparametric estimator for the function pair (f_0, g_0) as

$$(\widetilde{f}_n, \widetilde{g}_n) \in \operatorname*{argmin}_{f \in \mathscr{F}_n,\, g \in \mathscr{G}_n} \frac{1}{n} \sum_{i=1}^n L_\alpha(f(X_i), g(X_i); Y_i), \tag{3.4}$$

where $\mathscr{F}_n, \mathscr{G}_n$ are pre-determined classes of functions $[0,1]^d \to \mathbb{R}$. Because the objective function in (3.4) is non-differentiable and non-convex, the above estimator is not practically feasible,

particularly when dealing with highly complex function classes.

3.3.1 A two-step approach for nonparametric ES regression

In the context of the joint conditional quantile and ES model (3.1), our primary objective is to estimate the conditional ES function g_0, treating the conditional quantile function f_0 as a nuisance function parameter. While the objective function $(q,e) \to L_\alpha(q,e;Y)$ may lack desirable properties, it has been observed by Barendse (2020) and Peng and Wang (2023) that

$$\begin{aligned}\frac{\partial^2 \mathbb{E}\{L_\alpha(f(X),g(X);Y|X)\}\}}{\partial q\,\partial e}\bigg|_{f=f_0} &= -G_2'(g(X))\frac{\partial \mathbb{E}\{S_\alpha(f(X),g(X);Y)/\alpha|X\}}{\partial q}\bigg|_{f=f_0} \\ &= G_2'(g(X))\frac{F_{Y|X}(f(X))-\alpha}{\alpha}\bigg|_{f=f_0} = 0\end{aligned}$$

as long as the conditional distribution function of Y given X, denoted by $F_{Y|X}$, is continuous. Equivalently, we have

$$\partial_q \mathbb{E}\{S_\alpha(q,e;Y)|X\}\big|_{q=f_0(X)} = \alpha - F_{Y|X}(f_0(X)) = 0, \tag{3.5}$$

where

$$S_\alpha(q,e;Y) = \alpha q + (Y-q)\mathbb{1}(Y \le q) - \alpha e \ \text{ for } q,e \in \mathbb{R}. \tag{3.6}$$

This indicates that the partial derivative of the score function $(q,e) \to \mathbb{E}\{S_\alpha(q,e;Y)|X\}$, evaluated at the true conditional quantile function, is zero. Moreover, based on the definition of (conditional) Expected Shortfall (ES), this score function satisfies the moment condition $\mathbb{E}\{S_\alpha(f_0(X),g_0(X);Y)|X\} = 0$. Thanks to this orthogonality property, both of the two-step ES regression estimators proposed in Barendse (2020) and Peng and Wang (2023) exhibit local robustness to prior quantile estimation under joint linear models.

Motivated by the orthogonal property of S_α, we propose a nonparametric two-step ES

regression estimator using deep neural networks with the ReLU activation function. Note that quantile regression estimation is a self-contained problem. Therefore, it is natural to first obtain a nonparametric QR estimator $\widehat{f}_n$ of f_0, for which various methods can be applied. Next, for each conditional quantile function candidate f, we define the surrogate response variables

$$Z_i(f) := \{Y_i - f(X_i)\}\mathbb{1}\{Y_i \leq f(X_i)\} + \alpha f(X_i). \tag{3.7}$$

By plugging-in $f = \widehat{f}_n$, we propose a two-step nonparametric ES regression estimator $\widehat{g}_n$, defined as

$$\begin{aligned} &\widehat{g}_n \in \underset{g\in\mathscr{G}_n}{\operatorname{argmin}} \widehat{\mathscr{R}}(\widehat{f}_n, g), \\ &\text{where } \widehat{\mathscr{R}}(f,g) := \frac{1}{2n}\sum_{i=1}^{n} S_\alpha^2(f(X_i), g(X_i); Y_i) = \frac{1}{2n}\sum_{i=1}^{n}\{Z_i(f) - \alpha g(X_i)\}^2, \end{aligned} \tag{3.8}$$

and $\mathscr{G}_n$ is a pre-determined class of real-valued functions on $[0,1]^d$. In Section 3.4, we choose $\mathscr{G}_n$ to be a class of truncated deep ReLU neural networks and refer to $\widehat{g}_n$ as the deep least squares ES regression (DES) estimator. We then proceed to analyze the convergence rate of $\widehat{g}_n$ that explicitly depends on the sample size, network parameters, noise scale, as well as the prior QR estimation error. This choice of the function class enables the estimator to adapt to the hierarchical compositional structure of g_0.

The formulation (3.8) provides a general two-stage approach for estimating ES regression functions using a plugged-in QR estimator. Nonparametric quantile regression methods have seen significant development and expansion, including local polynomial regression methods (Chaudhuri, 1991), tree-based methods (Meinshausen, 2006), kernel ridge regression (regression in reproducing kernel Hilbert spaces) (Li et al., 2007), QR-series method (Belloni et al., 2019), and neural network regression (Padilla et al., 2022; Shen et al., 2021). To align with the theme of this work, we focus on quantile regression using deep neural networks (DQR) with the ReLU activation function. Specifically, we define the DQR estimator within the class $\mathscr{F}_n$ of truncated

ReLU neural networks as

$$\widehat{f_n} \in \underset{f \in \mathscr{F}_n}{\operatorname{argmin}} \left\{ \widehat{\mathscr{Q}}_\alpha(f) := \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(Y_i - f(X_i)) \right\}, \tag{3.9}$$

where $\rho_\alpha(u) = \{\alpha - \mathbb{1}(u < 0)\}u$ is the check function (Koenker and Bassett, 1978). The convergence rate of $\widehat{f_n}$ (in high probability) will be established in Section 3.4.2 when the true conditional quantile function f_0 also belongs to a hierarchical interaction model. Our results complement those obtained in Padilla et al. (2022), where the authors focused on sparsely connected networks with all the weight parameters and biases bounded by 1. Furthermore, our convergence rate is faster than that in Shen et al. (2021); see Remark 3.4.2 for a detailed comparison.

3.3.2 Robust nonparametric ES regression under heavy-tailed errors

The two-step estimator $\widehat{g}_n$, defined in (3.8), can be regarded as a nonparametric least squares estimator (LSE) with response variables $Z_i(\widehat{f_n})$ generated nonparametrically. The underlying model can be expressed as $\mathbb{E}\{Z_i(f_0)|X_i = x\} = \alpha g_0(x)$, or equivalently, $Z_i(f_0) = \alpha g_0(X_i) + \omega_i$, where $\omega_i = \varepsilon_{i,-} - \mathbb{E}(\varepsilon_{i,-}|X_i)$, with $\varepsilon_{i,-}$ denoting the negative part of the quantile regression error $\varepsilon_i := Y_i - f_0(X_i)$ defined as $\varepsilon_{i,-} = \min(\varepsilon_i, 0)$.

Due to the sensitivity of the quadratic loss function to outliers (Huber, 1973; Catoni, 2012), the aforementioned LSE is particularly sensitive to the tails of the distribution of η_i, which correspond to the left tails of ε_i. From a non-asymptotic perspective, the L_2-error of the LSE exhibits an exponential-type deviation (high probability) bound under light-tailed noise distributions, while it only demonstrates a polynomial-type deviation bound under heavy-tailed distributions. Furthermore, in contrast to the parametric setting where LSEs achieve the same convergence rates in terms of mean squared error (MSE) under both (exponentially) light-tailed errors and errors with bounded p-th ($p \geq 2$) moments, recent studies have shown that heavy-tailed errors can degrade the convergence rate of nonparametric LSEs, resulting in a slower convergence

rate (Han and Wellner, 2019; Kuchibhotla and Patra, 2022; Fan et al., 2022). Therefore, the LSE $\widehat{g}_n$ may exhibit a slower convergence rate when the noise follows a heavy-tailed distribution.

To address this issue, we propose an alternative approach by replacing the quadratic loss with a robust loss function that exhibits both global Lipschitz continuity and local quadratic behavior near 0, ensuring insensitivity to heavy-tailed noises. Specifically, we employ the Huber loss (Huber, 1964), defined as

$$\ell_\tau(u) := \begin{cases} u^2/2 & \text{if} \quad |u| \le \tau \\ \tau|u| - \tau^2/2 & \text{if} \quad |u| > \tau \end{cases}. \tag{3.10}$$

Here, $\tau > 0$ is a robutification parameter that separates its quadratic and linear components. Then, given an initial estimator $\widehat{f}_n$ of f_0, a nonparametric robust ES regression estimator is defined as follows:

$$\begin{aligned} &\widehat{g}_{n,\tau} \in \operatorname*{argmin}_{g \in \mathscr{G}_n} \widehat{\mathscr{R}}_\tau(\widehat{f}_n, g) \\ &\text{where } \widehat{\mathscr{R}}_\tau(f,g) := \frac{1}{n}\sum_{i=1}^n \ell_\tau(S_\alpha(f(X_i), g(X_i); Y_i)) = \frac{1}{n}\sum_{i=1}^n \ell_\tau(Z_i(f) - \alpha g(X_i)), \end{aligned} \tag{3.11}$$

with S_α and Z_i defined in (3.6) and (3.7), respectively. When the class $\mathscr{G}_n$ consists of truncated deep ReLU neural networks, we refer to $\widehat{g}_{n,\tau}$ as the deep robust (Huber) ES regression (DRES) estimator.

The choice of the robustification parameter τ plays a crucial role in achieving a balance between robustness and bias (Zhou et al., 2018). To understand the impact of employing the Huber loss, we define the global minimizer of the population Huber loss as

$$g_{0,\tau} \in \operatorname*{argmin}_{\|g\|_\infty \le M_0} \mathscr{R}_\tau(f_0, g) := \mathbb{E}\ell_\tau(S_\alpha(f_0(X_i), g(X_i); Y_i)), \tag{3.12}$$

where the minimization is performed over all measurable functions g satisfying $\|g\|_\infty \le M_0$ for

some constant $M_0 > 0$. Denoting the quantile regression residual as $\varepsilon = Y - f_0(X)$, we can express

$$\begin{aligned}\ell_\tau(S_\alpha(f_0(X), g_0(X); Y)) &= \ell_\tau(\alpha f_0(X) + \{Y - f_0(X)\}\mathbb{1}\{Y \le f_0(X)\} - \alpha g_0(X)) \\ &= \ell_\tau(\varepsilon_- - \mathbb{E}(\varepsilon_-|X)),\end{aligned}$$

where $\varepsilon_- = \min(\varepsilon, 0)$. From this definition, $\varepsilon_- - \mathbb{E}(\varepsilon_-|X)$ is generally asymmetric (with respect to zero), leading to a deviation between $g_{0,\tau}$ and g_0. To quantify this deviation, we present the following proposition, which provides an upper bound for the robustification bias, defined as $\|g_{0,\tau} - g_0\|_2$.

Proposition 3.3.1. Assume that $\varepsilon_- = \min(\varepsilon, 0)$ satisfies $\mathbb{E}\{|\varepsilon_- - \mathbb{E}(\varepsilon_-|X)|^p|X\} \le \nu_p$ almost surely for some constant $\nu_p > 0$ and $p \ge 2$. For any $\tau \ge c_4 = 2\max\{4M_0, (2\nu_p)^{1/p}\}$, the global minimizer $g_{0,\tau}$ defined in (3.11) satisfies $\alpha\|g_{0,\tau} - g_0\|_2 \le 2^{p+1}\nu_p\tau^{1-p}$.

Proposition 3.3.1 reveals that the upper bound on bias depends on the robustification parameter τ and the moment index p. Thus, to mitigate the bias induced by using the Huber loss, it is necessary to employ a sufficiently large τ. However, a large value of τ will increase the statistical error, as demonstrated in Theorem 3.4.1. Therefore, it is crucial to carefully calibrate the value of τ in order to strike a balance between robustness and bias.

3.4 Statistical Theory

In this section, we analyze the statistical properties of the proposed nonparametric quantile and expected shortfall regression estimators using ReLU neural networks, with a focus on the latter. In the two-step approach, estimating ES involves the use of (surrogate) response variables that are not directly observable but need to be estimated from data in a preliminary step. The first challenge is characterizing their impact on the statistical properties of the ES estimator in the second stage. The second challenge arises when analyzing the robustified estimator for ES,

even when the "noise" variable is heavy-tailed and skewed, despite having a zero conditional mean. In this case, even with the oracle surrogate response variables incorporated into the procedure, the existing results and techniques from Farrell et al. (2021), Padilla et al. (2022), and Shen et al. (2021) do not apply.

In Section 3.4.1, we begin our analysis by deriving a generic upper bound on the estimation error for the robust ES estimator defined in (3.11). Our focus is on the case where the noise distribution has a finite p-th moment ($p \geq 2$). We consider both the DRES and DES estimators with various configurations of deep ReLU neural networks, as well as any quantile regression estimator $\widehat{f}_n$. We also derive non-asymptotic error bounds for the DRES estimators under light-tailed noise distributions. This demonstrates that using a proper robust estimator leads to minimal to no efficiency loss from a non-asymptotic perspective, in comparison to least squares estimators.

In Section 3.4.2, we revisit deep QR estimators given in (3.9) and examine their non-asymptotic statistical guarantees. Notably, we improve the existing results in the literature by employing different proof techniques and leveraging the new approximation result, Proposition 3.2.1; see Remark 3.4.2 for a comprehensive comparison with two existing related works. Finally, in Section 3.4.3, we combine the results from Sections 3.4.1 and 3.4.2 to establish the convergence rate of deep ES estimators when a DQR estimator is used to construct the surrogate responses. We specifically focus on the setting where both quantile and ES regression functions lie in hierarchical interaction models.

3.4.1 A generic upper bound of deep ES estimator

Before presenting our theoretical results, we impose the following conditions on the quantile regression residual $\varepsilon = Y - f_0(X)$ and its negative part $\varepsilon_- = \min(\varepsilon, 0)$. In this notation, we can equivalently express model (3.1) as $Y = f_0(X) + \varepsilon$, where the noise variable ε and the

conditional ES function g_0 satisfy

$$\mathbb{P}(\varepsilon \leq 0|X) = \alpha \text{ and } g_0(X) = f_0(X) + \frac{1}{\alpha}\mathbb{E}\{\varepsilon \mathbb{1}(\varepsilon \leq 0)|X\}.$$

Condition 1 (Noise distribution)**.** The conditional density function of ε given X, denoted by $p_{\varepsilon|X}$, exists and satisfies $p_{\varepsilon|X}(u) \leq \bar{p}$ for some constant $\bar{p} > 0$ almost surely (over X) for all $u \in \mathbb{R}$. Moreover, the negative part of the residual has uniformly bounded (conditional) p-th central moments for some $p \geq 2$, that is, there exists $\nu_p > 0$ such that $\mathbb{E}\{|\varepsilon_- - \mathbb{E}(\varepsilon_-|X)|^p|X\} \leq \nu_p$ almost surely over X.

Condition 1 requires the existence of a bounded conditional density of the response variable given covariates and that the negative part of the quantile residual ε has a bounded (conditional) p-th central moment.

Next, we recall the definition of the Pseudo dimension of a real-valued function class.

Definition 3.4.1 (Pseudo dimension (Anthony and Bartlett, 1999))**.** Let $\mathscr{F}$ be a set of real-valued functions on a domain $\mathscr{X}$. The pseudo dimension of $\mathscr{F}$, denoted by $\text{Pdim}(\mathscr{F})$, is defined to be the largest integer N for which there exist $\{x_1, x_2, \ldots, x_N\} \in \mathscr{X}^N$ and $\{r_1, r_2, \ldots, r_N\} \in \mathbb{R}^N$ such that for any $\mathbf{b} = (b_1, \ldots, b_N)^{\mathrm{T}} \in \{0,1\}^N$, there is a function $f \in \mathscr{F}$ with $\mathbb{1}\{f(x_i) \geq r_i\} = b_i$ for $1 \leq i \leq N$.

To quantify the estimation accuracy, for any $q \geq 1$, we use $\|\cdot\|_q$ to denote the function L_q-norm, that is, $\|h\|_q := \|h\|_{\mathbb{P}_X,q} = \{\mathbb{E}_{X\sim\mathbb{P}_X}|h(X)|^q\}^{1/q}$ for any function $h : [0,1]^d \to \mathbb{R}$.

Our first result is an oracle-type inequality that provides an upper bound on the L_2-error of the DRES estimator for any truncated deep ReLU network architecture, any robustification parameter $\tau \geq c_4$, and any QR estimator.

Theorem 3.4.1 (Oracle-type inequality for the DRES estimator)**.** Assume Condition 1 holds with $p \geq 2$, and $\max(\|f_0\|_\infty, \|g_0\|_\infty) \leq M_0$ for some $M_0 \geq 1$. Let $\tau \geq c_4$, $L, N \in \{3, 4, \ldots\}$ and

$\mathscr{G}_n = \mathscr{F}_{\mathrm{DNN}}(d, L, N, M_0)$. Given a class $\mathscr{F}_n$ of real-valued functions from $[0,1]^d$ to $[-M_0, M_0]$ with finite pseudo dimension, define

$$\begin{cases} \eta_{\mathrm{b}} = \frac{\nu_p}{(\tau/2)^{p-1}}, & \eta_{\mathrm{a}} = \inf_{g \in \mathscr{G}_n} \|g - g_0\|_2, \\ \eta_{\mathrm{s}} = (\nu_p^{1/p} + \sqrt{\tau})\, V_{n,\tau,\nu_p}, & \delta_{\mathrm{s}} = \sqrt{\mathrm{Pdim}(\mathscr{F}_n)\log(n)/n}, \end{cases} \tag{3.13}$$

and $V_{n,\tau,\nu_p} = LN\sqrt{\log(LN)\log(n^2\tau/\nu_p^{1/p})/n}$. For any real-valued function f, let $\mathscr{T}_{n,\tau}(\eta; f)$ be the set of approximate empirical Huber risk minimizers with optimization error $\eta \geq 0$, that is,

$$\mathscr{T}_{n,\tau}(\eta; f) = \left\{ g \in \mathscr{G}_n : \widehat{\mathscr{R}}_\tau(f, g) \leq \inf_{h \in \mathscr{G}_n} \widehat{\mathscr{R}}_\tau(f, h) + \eta^2 \right\}, \tag{3.14}$$

where $\widehat{\mathscr{R}}_\tau$ is defined in (3.11). Then, there exists a universal constant $c_5 > 0$ independent of $(N, L, n, \nu_p, p, d, f_0, g_0)$ such that, for any $\eta_{\mathrm{opt}} \geq 0$ and $u \geq 1$, the following bound

$$\sup_{g \in \mathscr{T}_{n,\tau}(\eta_{\mathrm{opt}}; \widehat{f}_n)} \|g - g_0\|_2 \leq \frac{c_5}{\alpha} \left\{ \eta_{\mathrm{s}} + \eta_{\mathrm{b}} + \eta_{\mathrm{a}} + \delta_{\mathrm{s}} + \delta_4^2 + \eta_{\mathrm{opt}} + (\nu_p^{1/p} + \sqrt{\tau}) \sqrt{\frac{u}{n}} \right\} \tag{3.15}$$

holds with probability at least $1 - Ce^{-u}$ conditioning on the event $\{\widehat{f}_n \in \mathscr{F}_0(\delta_4)\}$ for some $\delta_4 > 0$, where $\mathscr{F}_0(\delta) := \{f \in \mathscr{F}_n : \|f - f_0\|_4 \leq \delta\}$.

Theorem 3.4.1 establishes a non-asymptotic error bound for approximate DRES estimators using a plugged-in QR estimator $\widehat{f}_n$. This upper bound consists of six distinct terms: two stochastic error terms η_{s} and δ_{s} that correspond to the (conditional) quantile and ES estimation respectively, the bias η_{b} induced by the Huber loss, the neural network approximation error η_{a} for the underlying ES regression function g_0, the optimization error η_{opt}, and the squared L_4-error δ_4^2 for the QR estimator $\widehat{f}_n$.

Proposition 3.2.1 shows that increasing LN reduces the approximation error η_{a} when g_0 belongs to a hierarchical interaction model. However, this increase results in a larger stochastic error η_{s}. Together, these two terms highlight the trade-off between the complexity of the network

function class and its approximation power. Moreover, the term $\delta_s + \delta_4^2 + \eta_{\mathrm{b}} + \eta_{\mathrm{s}}$ explicitly reveals the impact of nonparametric QR estimation in stage one and the use of the Huber loss. The former is quantified by δ_{s} and δ_4^2. Thanks to the orthogonality condition (3.5), the squared L_4-error of the nonparametric QR estimator contributes to the L_2-error bound for the two-step ES estimator. Consequently, even if the QR estimator converges at a sub-optimal rate (under the L_4-norm), the ES estimator can still achieve the optimal convergence rate under the L_2-norm, as if the true quantile function f_0 were known. On the other hand, $\eta_{\mathrm{b}} + \eta_{\mathrm{s}}$ clarifies the role of the robustification parameter τ. A larger τ reduces bias, resulting in a smaller η_{b}. However, this reduction comes at the expense of compromising robustness, leading to a larger η_{s}. Therefore, it is crucial to properly tune the robustification parameter τ to balance bias and robustness.

As a benchmark method, we also derive non-asymptotic deviation bounds for DES estimators with any truncated deep ReLU network architecture and an initial nonparametric QR estimator.

Theorem 3.4.2 (Oracle-type inequality for the DES estimator)**.** Assume Condition 1 holds with $p \geq 2$, and $\max(\|f_0\|_\infty, \|g_0\|_\infty) \leq M_0$ for some $M_0 \geq 1$. Let $\mathscr{G}_n = \mathscr{F}_{\mathrm{DNN}}(d, L, N, M_0)$ with integers $L, N \geq 3$, and $\mathscr{F}_n$ be a class of functions from $[0,1]^d$ to $[-M_0, M_0]$ with a finite pseudo dimension. Define

$$\eta_{\mathrm{a}} = \inf_{g \in \mathscr{G}_n} \|g - g_0\|_2, \ \eta_{\mathrm{s}} = \nu_p^{1/p} V_n + \nu_p^{1/(2p)} V_n^{1-1/p} \ \text{ and } \ \delta_{\mathrm{s}} = \sqrt{\mathrm{Pdim}(\mathscr{F}_n)\log(n)/n}, \quad (3.16)$$

where $V_n = LN\sqrt{\log(LN)\log(n)/n}$. If n is sufficiently large so that $V_n \leq 1$, there exists some universal constant $c_6 > 0$ such that, for any $\eta_{\mathrm{opt}} \geq 0$ and $u \geq 1$, the following bound

$$\sup_{g \in \mathscr{T}_{n,\infty}(\eta_{\mathrm{opt}}; \widehat{f}_n)} \|g - g_0\|_2 \leq c_6 \alpha^{-1} \sqrt{u} \big(\eta_{\mathrm{s}} + \eta_{\mathrm{a}} + \delta_{\mathrm{s}} + \delta_4^2 + \eta_{\mathrm{opt}}\big) \quad (3.17)$$

holds with probability at least $1 - C(e^{-nV_n^2} + u^{-p})$ conditioning on the event $\{\widehat{f}_n \in \mathscr{F}_0(\delta_4)\}$, where $\mathscr{T}_{n,\infty}$ is defined in (3.14) by taking $\tau = \infty$.

In contrast to the result of Theorem 3.4.1, Theorem 3.4.2 demonstrates that the deviation bound of the DES estimator does not include the bias term. This is because the population Huber loss minimizer $g_{0,\tau}$, defined in (3.12), coincides with g_0 when $\tau = \infty$, resulting in $\eta_{\mathrm{b}} = 0$. However, since DES uses the L_2-loss, the corresponding estimator exhibits only a polynomial-type deviation bound, as shown in Theorem 3.4.2. This is in contrast to the exponential-type deviation bound achieved by the DRES estimator.

To complement our analysis, we investigate the non-asymptotic error bound of the DRES estimator under the presence of light-tailed noise distributions. Specifically, we assume that the negative part of the quantile residual ε follows a sub-Gaussian distribution as follows.

Condition 2 (Light-tailed noise)**.** The conditional density function of ε given X, denoted by $p_{\varepsilon|X}$, exists and satisfies $\sup_{u\in\mathbb{R}} p_{\varepsilon|X}(u) \le \bar{p}$ for some constant $\bar{p} > 0$ almost surely (over X). Moreover, there exists a constant $\sigma_0 > 0$ such that the negative part of the QR residual satisfies $\mathbb{E}[\exp(\{\varepsilon_- - \mathbb{E}(\varepsilon_-|X)\}^2/\sigma_0^2)|X] \le 2$ almost surely over X.

Theorem 3.4.3 (Oracle-type inequality for the DRES estimator with sub-Gaussian errors)**.** Assume Condition 2 holds for some $\sigma_0 > 0$, and $\max(\|f_0\|_\infty, \|g_0\|_\infty) \le M_0$ for some $M_0 \ge 1$. Let $L, N \in \{3,4,\dots\}, \mathscr{G}_n = \mathscr{F}_{\mathrm{DNN}}(d,L,N,M_0)$ and $\tau \ge c_7 := 2\max\{4M_0, (\log 4)^{1/2}\sigma_0\}$. Given a class $\mathscr{F}_n$ of real-valued functions from $[0,1]^d$ to $[-M_0, M_0]$ with finite pseudo dimension, define

$$\begin{cases} \eta_{\mathrm{b}} = 2(2M_0 + \sigma_0)e^{-\tau^2/(2\sigma_0^2)}, & \eta_{\mathrm{a}} = \inf_{g\in\mathscr{G}_n} \|g - g_0\|_2, \\ \eta_{\mathrm{s}} = \sigma_0 LN\sqrt{\log(LN)\log(n)/n}, & \delta_{\mathrm{s}} = \sqrt{\mathrm{Pdim}(\mathscr{F}_n)\log(n)/n}. \end{cases} \tag{3.18}$$

Then, there exists some universal constant $c_8 > 0$ such that for any $\delta_4 > 0$, $\eta_{\mathrm{opt}} \ge 0$ and $u \ge 1$, the following bound

$$\sup_{g\in\mathscr{T}_{n,\tau}(\eta_{\mathrm{opt}};\widehat{f}_n)} \|g - g_0\|_2 \le \frac{c_8}{\alpha}\left(\eta_{\mathrm{s}} + \eta_{\mathrm{b}} + \eta_{\mathrm{a}} + \delta_{\mathrm{s}} + \delta_4^2 + \eta_{\mathrm{opt}} + \sigma_0\sqrt{\frac{u}{n}}\right) \tag{3.19}$$

holds with probability at least $1 - Ce^{-u}$ conditioning on the event $\{\widehat{f}_n \in \mathscr{F}_0(\delta_4)\}$, where $\mathscr{T}_{n,\tau}$ is

defined in (3.14).

In contrast to the setting where ε_- only possesses a bounded (conditional) p-th central moment, Theorem 3.4.3 reveals that the bias term η_{b} decays exponentially in τ when ε_- is (conditional) sub-Gaussian. In particular, we have $\eta_{\mathrm{b}} \leq \sigma_0 n^{-1/2}$ as long as $\tau \geq \sigma_0\sqrt{\log n}$. Consequently, the impact of the robustification bias becomes negligible compared to the statistical error η_{s} in (3.18), which is unaffected by τ.

Remark 3.4.1 (Sample splitting and cross-fitting)**.** We can eliminate the statistical error term δ_{s}, induced by the estimation of the conditional QR function, from the error bounds of the proposed estimator by incorporating a sample-splitting algorithm.

Specifically, we first split the entire dataset into two parts: $\{(X_1,Y_1),\ldots,(X_{n_1},Y_{n_1})\}$ and $\{(X_{n_1+1},Y_{n_1+1}),\ldots,(X_n,Y_n)\}$, where $n_1 = \lceil n/2 \rceil$. The first subsample is used to train a QR estimator $\widehat{f}_n$, while the remaining subsample, together with $\widehat{f}_n$, is employed to compute the ES regression estimator $\widehat{g}_{\mathrm{split}}$. Following similar arguments as in the proofs of Theorems 3.4.1–3.4.3, it can be established that under the same conditions as outlined in Theorems 3.4.1–3.4.3, $\widehat{g}_{\mathrm{split}}$ satisfies concentration bounds that are similar to (3.15), (3.17) and (3.19), without the appearance of δ_{s}. As a result, the impact of QR estimation is only reflected by δ_4^2.

Nevertheless, using only half of the data to compute $\widehat{g}_{\mathrm{split}}$ may result in a loss of statistical efficiency. To mitigate this issue, the widely recognized approach is cross-fitting as discussed in Chernozhukov et al. (2018). Nonetheless, it remains uncertain whether the cross-fitting method improves the statistical efficiency over the basic sample-splitting method in our case. As pointed out by Foster and Syrgkanis (2023), establishing this improvement typically requires the demonstration of asymptotic normality or linear approximation of the nonparametric estimator in the literature. However, it remains an open question whether a DNN estimator exhibits an asymptotic linear approximation, which in turn leads to asymptotic normality. As a result, from a theoretical perspective, it remains unclear whether the use of cross-fitting can enhance the statistical efficiency over the basic sample-splitting in our setting.

3.4.2 Deep quantile regression estimator

In this section, we provide concentration bounds for the DQR estimator defined in (3.9), which is the nonparametric QR estimator obtained through empirical risk minimization over truncated ReLU neural networks using the check loss. As is common in QR literature, we begin by imposing certain regularity conditions on the conditional density function of ε given X.

Condition 3 (Conditional density)**.** The conditional density function of $\varepsilon = Y - f_0(X)$ given X, denoted by $p_{\varepsilon|X}$ exists and is continuous on its support. It satisfies

$$\underline{p} \leq p_{\varepsilon|X}(0) \leq \sup_{u\in\mathbb{R}} p_{\varepsilon|X}(u) \leq \bar{p}$$

almost surely (over X) for some $\bar{p} \geq \underline{p} > 0$. Moreover, there exists a constant $l_0 > 0$ such that $|p_{\varepsilon|X}(u_1) - p_{\varepsilon|X}(u_2)| \leq l_0|u_1 - u_2|$ for all $u_1, u_2 \in \mathbb{R}$ almost surely (over X).

Condition 3 is a standard assumption for the analysis of quantile regression estimators, especially from a non-asymptotic perspective. See, for example, Belloni and Chernozhukov (2011), Belloni et al. (2019), Pan and Zhou (2021) and Padilla et al. (2022).

We are now prepared to present an oracle-type error bound for the DQR estimator with an arbitrary ReLU neural network configuration. Recall that the empirical quantile loss $\widehat{\mathscr{Q}}_\alpha$ is defined as $\widehat{\mathscr{Q}}_\alpha(f) = n^{-1}\sum_{i=1}^n \rho_\alpha(Y_i - f(X_i))$ for any real-valued function f.

Theorem 3.4.4 (Oracle-type inequality for the DQR estimator)**.** Assume Condition 3 holds and $\|f_0\|_\infty \leq M_0$ for some $M_0 \geq 1$. Let $L, N \in \{3, 4, \dots\}$ and $\mathscr{F}_n = \mathscr{F}_{\text{DNN}}(d, L, N, M_0)$. Define

$$\delta_{\text{a}} = \inf_{f\in\mathscr{F}_n} \|f - f_0\|_2 \ \text{ and } \ \delta_{\text{s}} = LN\sqrt{\frac{\log(LN)\log n}{n}}.$$

Let $\mathscr{S}_n(\delta)$ be the set of approximate empirical (quantile) risk minimizers with the optimization

error $\delta > 0$, that is

$$\mathscr{S}_n(\delta) = \left\{ f \in \mathscr{F}_n : \widehat{\mathscr{Q}}_\alpha(f) \leq \inf_{\widetilde{f} \in \mathscr{F}_n} \widehat{\mathscr{Q}}_\alpha(\widetilde{f}) + \delta^2 \right\}. \tag{3.20}$$

Then, there exists some universal constant $c_9 > 0$ independent of (N, L, n, d, α) and f_0 such that for any $\delta_{\text{opt}} \geq 0$ and $u \geq 1$,

$$\mathbb{P}\left\{ \sup_{f \in \mathscr{S}_n(\delta_{\text{opt}})} \|f - f_0\|_2 \geq c_9 \left(\delta_{\text{s}} + \delta_{\text{a}} + \delta_{\text{opt}} + \sqrt{\frac{u}{n}} \right) \right\} \lesssim e^{-u}. \tag{3.21}$$

The non-asymptotic deviation bound, as presented in (3.21), comprises three main components: the stochastic error δ_{s}, the approximation error δ_{a} concerning f_0, and the optimization error δ_{opt}. Here, the statistical error term δ_{s} increases as the network hyper-parameters L and N grow, while the approximation error term δ_{a} decreases; see Proposition 3.2.1. Furthermore, it is important to note that exponential-type concentration inequalities naturally apply to non-parametric QR estimators even without requiring moment conditions on ε_i. However, specific regularity conditions on its (conditional) density function are still necessary. This underscores the robustness of quantile regression, particularly in handling the tails of the response variable.

By selecting suitable values for L and N to balance the stochastic and approximation errors, we demonstrate in the following result that the DQR estimator achieves optimal convergence rates when f_0 has a hierarchical interaction structure.

Theorem 3.4.5 (Convergence rate for the DQR estimator)**.** Assume Condition 3 holds and that $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure on $[0,1]^d$. Let γ^* be as in (3.2), and $L_0, N_0 \geq 3$ be such that $L_0 N_0 \asymp (n/\log^6 n)^{1/(4\gamma^*+2)}$. Consider the function class $\mathscr{F}_n = \mathscr{F}_{\text{DNN}}(d, L, N, M)$, where the depth and width are given by

$$L = c_1 \lceil L_0 \log L_0 \rceil \text{ and } N = c_2 \lceil N_0 \log N_0 \rceil, \tag{3.22}$$

respectively. Here, c_1 and c_2 are positive constants from Proposition 3.2.1. Then, for any $u \geq 1$ and $\delta_{\text{opt}} \leq \delta_n = (n/\log^6 n)^{-\gamma^*/(2\gamma^*+1)}$, it holds uniformly over $f_0 \in \mathscr{H}(d, l, M_0, \mathscr{P})$ and for all sufficiently large n that

$$\mathbb{P}\left[\sup_{f \in \mathscr{S}_n(\delta_{\text{opt}})} \|f - f_0\|_2 \geq c_{10}\left\{\left(\frac{\log^6 n}{n}\right)^{\gamma^*/(2\gamma^*+1)} + \sqrt{\frac{u}{n}}\right\}\right] \lesssim e^{-u},$$

where $c_{10} > 0$ is a universal constant depending polynomially on $t_{\max} = \max_{(t,\beta)\in\mathscr{P}} t$.

An immediate consequence of Theorem 3.4.5 is that

$$\|\widehat{f}_n - f_0\|_2 = \mathscr{O}_{\mathbb{P}}\Big(n^{-\gamma^*/(2\gamma^*+1)}(\log n)^{6\gamma^*/(2\gamma^*+1)}\Big).$$

By combining this upper bound with the following proposition, the DQR estimator, with an appropriately chosen network structure, achieves the minimax optimal convergence rate for the hierarchical interaction model, up to logarithmic terms. Recall the definition of t^* in (3.2).

Proposition 3.4.1 (Minimax lower bound for the hierarchical interaction model)**.** Assume $d \geq t^*$ and that Condition 3 holds. Then, it holds

$$\liminf_{n\to\infty} \inf_{\widetilde{f}_n} \sup_{\substack{f_0 \in \mathscr{H}(d,l,\mathscr{P},M_0) \\ X\sim\mathbb{P}_X}} n^{2\gamma^*/(2\gamma^*+1)} \mathbb{E}\|\widetilde{f}_n - f_0\|_2^2 > 0,$$

where the infimum is taken over all estimators constructed from the sample $\{(X_i, Y_i)\}_{i=1}^n$.

Remark 3.4.2 (Comparison to existing work on DNN estimators for quantile regression)**.** In recent years, there has been a growing interest in applying DNNs for nonparametric quantile regression due to its great success for solving classification and regression problems in general. When the true conditional quantile function has a compositional structure, Shen et al. (2021) derived upper bounds on a hybrid of L_1- and L_2-errors of the QR estimator using ReLU neural networks. Their analysis is restricted to the case where the smoothness of each component

function does not exceed 1. Moreover, assuming that the response variable, or equivalently, the regression error, has bounded p-th absolute moment, Shen et al. (2021) showed that

$$\mathbb{E}\Delta^2(\widehat{f}, f_0) \lesssim n^{-(2-2/p)\gamma^*/(2\gamma^*+1)} \log^2(n),$$

where $\Delta^2(f, f_0) = \mathbb{E}_{X\sim\mathbb{P}_X} \min\{|f(X) - f_0(X)|, |f(X) - f_0(X)|^2\}$, and γ^* plays a similar role as that defined in (3.2), which is the dimension-adjusted degree of smoothness. First, we note that the above bound does not imply an L_2-error bound but rather $\Delta^2(f, f_0) \leq \|f - f_0\|_2^2$. Furthermore, the convergence rate is inflated by a factor of $n^{(2/p)\gamma^*/(2\gamma^*+1)}$ under heavy-tailed errors compared to that under exponentially light-tailed errors, making it sub-optimal. This contradicts, however, the robustness nature of quantile regression in response to outliers in the response space.

Another recent work Padilla et al. (2022) also explored nonparametric QR estimators using deep ReLU neural networks and established optimal convergence rates for cases where the quantile function is compositional with Hölder smooth components or belongs to a Besov space. Our results differ from Padilla et al. (2022) in several aspects. First, Padilla et al. (2022) constrained their function class to sparse neural networks with bounded weights and biases, while the function class examined in this section does not have such restrictions. As a result, our approach is more practical, as implementing the restrictions mentioned in Padilla et al. (2022) necessitates various techniques like projection and dropout, as described in Goodfellow et al. (2016). Secondly, when the true quantile function is a composition of Hölder smooth functions, Theorem 2 in Padilla et al. (2022) requires the width of neural networks to increase as a power of n, and the depth L to be $L \asymp \log n$ to attain the optimal convergence rate. In contrast, Theorem 3.4.5 only necessitates an assumption regarding the product of the depth and width of neural networks, thereby offering flexibility in network design. This means that the optimal rate can be achieved with wide and shallow neural networks, thin and deep neural networks, or wide and deep neural networks as long as the product satisfies the assumption. Last but not least, the prefactor in the error bounds derived from Padilla et al. (2022) grows exponentially

with dimension, whereas the prefactor in Theorem 3.4.5 grows polynomially. Consequently, the dimension-dependent prefactor in Padilla et al. (2022) can dominate the error bound when the dimension is moderately large.

3.4.3 Convergence analysis of joint deep quantile and ES regression

Building on the results from Sections 3.4.1 and 3.4.2, in this section, we establish the convergence rates of two-step DRES and DES estimators using an initial DQR estimate. Based on the findings from the previous subsections, the key is to tune the hyper-parameters appropriately to achieve an optimal balance among the various error terms.

We first consider the DRES estimator defined in (3.11) in the presence of heavy-tailed noises. By combining Theorem 3.4.1, Theorem 3.4.4 and the neural network approximation result, Proposition 3.2.1, we establish the convergence rate of the DRES estimator as follows.

Theorem 3.4.6 (Convergence rate for the DRES estimator using a plugged-in DQR estimate)**.** Assume Conditions 1 and 3 hold with $p \geq 2$. Additionally, assume that $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure on $[0,1]^d$. Let γ^* be as in (3.2), and $L_0, N_0 \geq 3$ be such that

$$L_0 N_0 \asymp \left(\frac{n}{\log^6 n}\right)^{\zeta_p/(4\gamma^*+2\zeta_p)} \quad \text{with} \quad \zeta_p = 1 - \frac{1}{2p-1}.$$

Consider the function classes $\mathscr{F}_n = \mathscr{G}_n = \mathscr{F}_{\mathrm{DNN}}(d, L, N, M_0)$ with depth L and width N satisfying (3.22). Set

$$\eta_n^{\mathrm{AH}} \asymp \max\left(\nu_p^{1/p}, 1\right) \cdot \left(\frac{\log^6 n}{n}\right)^{\gamma^*\zeta_p/(2\gamma^*+\zeta_p)} \quad \text{and} \quad \tau \asymp \nu_p^{1/p}\left(\frac{n}{\log^6 n}\right)^{2\gamma^*(1-\zeta_p)/(2\gamma^*+\zeta_p)}.$$

Then, for any $u \geq 1$, $\delta_{\mathrm{opt}} \leq \eta_n^{\mathrm{AH}}$ and $\eta_{\mathrm{opt}} \leq \eta_n^{\mathrm{AH}}$, it holds uniformly over $f_0, g_0 \in \mathscr{H}(d, l, M_0, \mathscr{P})$

and for all sufficiently large n that

$$\mathbb{P}\left\{\sup_{f\in\mathscr{S}_n(\delta_{\mathrm{opt}}),\, g\in\mathscr{T}_{n,\tau}(\eta_{\mathrm{opt}};f)} \|g-g_0\|_2 \geq \frac{c_{11}}{\alpha}\left[\eta_n^{\mathrm{AH}} + \max\left\{\nu_p^{1/(2p)},1\right\}\sqrt{\frac{u}{n^{\zeta_p}}}\right]\right\} \lesssim e^{-u},$$

where $\mathscr{S}_n$ and $\mathscr{T}_{n,\tau}$ are defined in (3.20) and (3.14), respectively. Here, $c_{11}>0$ is independent of (n,u,p,ν_p) and depends polynomially on $t_{\max}=\max_{(t,\beta)\in\mathscr{P}} t$.

Next, we investigate the DES estimator defined in (3.8) in the presence of heavy-tailed noises. By combining Theorem 3.4.4, Proposition 3.2.1 and Theorem 3.4.2, we derive the convergence rate for the DES estimators as follows.

Theorem 3.4.7 (Convergence rate for the DES estimator using a plugged-in DQR estimate)**.** Under the same conditions as in Theorem 3.4.6, let $L_0, N_0 \geq 3$ be such that

$$L_0 N_0 \asymp \left(\frac{n}{\log^6 n}\right)^{\xi_p/(4\gamma^*+2\xi_p)} \quad \text{with } \xi_p = 1-\frac{1}{p}.$$

Consider the function classes $\mathscr{F}_n=\mathscr{G}_n=\mathscr{F}_{\mathrm{DNN}}(d,L,N,M_0)$ with depth L and width N satisfying (3.22). Set $\eta_n^{\mathrm{LS}} \asymp \max(\nu_p^{1/p},1)\cdot\{\log^6(n)/n\}^{\gamma^*\xi_p/(2\gamma^*+\xi_p)}$. Then, for any $u\geq 1$, $\delta_{\mathrm{opt}}\leq\eta_n^{\mathrm{LS}}$ and $\eta_{\mathrm{opt}}\leq\eta_n^{\mathrm{LS}}$, it holds uniformly for all $f_0,g_0\in\mathscr{H}(d,l,M_0,\mathscr{P})$ and for all sufficiently large n that

$$\mathbb{P}\left\{\sup_{f\in\mathscr{S}_n(\delta_{\mathrm{opt}}),\, g\in\mathscr{T}_{n,\tau}(\eta_{\mathrm{opt}};f)} \|g-g_0\|_2 \geq c_{12}\alpha^{-1}\sqrt{u}\,\eta_n^{\mathrm{LS}}\right\} \lesssim \frac{1}{u^p}.$$

Here, $c_{12}>0$ is independent of (n,u,p,ν_p) and depends polynomially on $t_{\max}$.

Given $\nu_p \asymp 1$, it is easy to see that η_n^{LS} is larger than η_n^{AH} because $\xi_p<\zeta_p$. Therefore, the DES estimator converges at a slower rate than the DRES estimator. More importantly, the deviation bounds in Theorem 3.4.6 and Theorem 3.4.7 confirm that, from a non-asymptotic perspective, the DRES estimator is significantly more robust against heavy tails.

Remark 3.4.3. When ε_- has a (conditional) bounded p-th ($p \geq 2$) moment and $f_0, g_0 \in \mathscr{H}(d,l,M_0,\mathscr{P})$, Theorem 3.4.6 and Theorem 3.4.7 establish that after selecting an appropriate robustification parameter and network structures, the two-step robust estimator $\widehat{g}_{n,\tau}$ satisfies

$$\alpha\|\widehat{g}_{n,\tau} - g_0\|_2 = \mathscr{O}_{\mathbb{P}}\Big(n^{-\gamma^*\zeta_p/(2\gamma^*+\zeta_p)}(\log n)^{6\gamma^*\zeta_p/(2\gamma^*+\zeta_p)}\Big), \tag{3.23}$$

and the two-step LSE $\widehat{g}_n$ achieves the following convergence rate

$$\alpha\|\widehat{g}_n - g_0\|_2 = \mathscr{O}_{\mathbb{P}}\Big(n^{-\gamma^*\xi_p/(2\gamma^*+\xi_p)}(\log n)^{6\gamma^*\xi_p/(2\gamma^*+\xi_p)}\Big), \tag{3.24}$$

respectively. We remark that when the function class $\mathscr{H}(d,l,M_0,\mathscr{P})$ satisfies $d \geq t^*$, these upper bounds are sharp up to a logarithmic factor of n. In detail, for given depth L and width N of neural networks, define

$$\mathscr{T}^{\mathrm{AH}}_{n,\tau}(\eta_{\mathrm{opt}}) := \Bigg\{ g \in \mathscr{F}_n(d,L,N,1) : \widehat{\mathscr{R}}_\tau(f_0,g) \leq \inf_{g\in\mathscr{F}_n(d,L,N,1)} \widehat{\mathscr{R}}_\tau(f_0,g) + n^{-100} \text{ or } \widehat{\mathscr{R}}_\tau(f_0,g) \leq \widehat{\mathscr{R}}_\tau(f_0,g_{0,\tau}) \vee \Big\{ \inf_{g\in\mathscr{F}_n(d,L,N,1)} \widehat{\mathscr{R}}_\tau(f_0,g) + C_1\eta^2_{\mathrm{opt}} \Big\} \Bigg\},$$

where $g_{0,\tau}$ is defined in (3.12). Furthermore, for a fixed function $f_0 : [0,1]^d \to \mathbb{R}$, and a function class $\mathscr{H} \subseteq \{g : \mathbb{R}^d \to [-1,1]\}$, define the family of data generating processes $\mathscr{U}(d,p,\mathscr{H})$ as follows: (i) Each coordinate of $X \in [0,1]^d$ follows the uniform distribution, (ii) $Y = f_0(X) + \varepsilon$ with $\mathbb{P}(\varepsilon \leq 0|X) = \alpha$, (iii) $e_\alpha(Y|X) = g_0(X) \in \mathscr{H}$, and (iv) $\mathbb{E}\{|\varepsilon_- - \mathbb{E}(\varepsilon_-|X)|^p|X\} \leq 1$. We denote $\eta_{n,*} \asymp n^{-\gamma^*\zeta_p/(2\gamma^*+\zeta_p)}(\log n)^{-\gamma^*(3\zeta_p+4)/(2\gamma^*+\zeta_p)}$ for a given $\mathscr{H}(d,l,1,\mathscr{P})$. Then, by combining Lemma 4.1 Fan et al. (2022) and Theorem 4.1 in Fan et al. (2022), we have

$$\liminf_{n\to\infty} \inf_{N,L\geq C_2,\tau\geq C_3} \sup_{(X,Y)\in\mathscr{U}(d,p,\mathscr{H})} \mathbb{P}\big\{\exists \widehat{g} \in \mathscr{T}^{\mathrm{AH}}_{n,\tau}(\eta_{n,*}) \text{ such that } \alpha\|\widehat{g} - g_0\|_2 \geq \eta_{n,*}\big\} = 1,$$

where $\mathscr{H} = \mathscr{H}(d,l,\mathscr{P},1)$ with $d \geq t^*$. Therefore, the L_2 error bound (3.23) for $\widehat{g}_{n,\tau}$ is sharp

up to logarithmic terms. In a similar manner, it can be shown that the bound (3.24) of $\widehat{g}_n$ is also sharp up to logarithmic terms by Theorem 4.2 in Fan et al. (2022).

Finally, we consider the case where the noise is sub-Gaussian. The following theorem shows that with a sufficiently large robustification parameter, the DRES estimator achieves the same convergence rate as the DES estimator.

Theorem 3.4.8 (Convergence rate for the DRES estimator using a plugged-in DQR estimate under sub-Gaussian noise)**.** Assume Conditions 2 and 3 hold. Moreover, assume that $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure on $[0,1]^d$. Let γ^* be as in (3.2), and $L_0, N_0 \geq 3$ be such that $L_0 N_0 \asymp \{n/\log^6(n)\}^{1/(4\gamma^*+2)}$. Consider the function classes $\mathscr{F}_n = \mathscr{G}_n = \mathscr{F}_{\mathrm{DNN}}(d, L, N, M_0)$, where the depth L and width N satisfy (3.22). Set $\tau \in [\max(c_7, \sigma_0\sqrt{\log n}), \infty]$ and $\eta_n^{\mathrm{subG}} \asymp \{\log^6(n)/n\}^{\gamma^*/(2\gamma^*+1)}$. Then, for any $u \geq 1$ and $\delta_{\mathrm{opt}}, \eta_{\mathrm{opt}} \leq \eta_n^{\mathrm{subG}}$, it holds uniformly over $f_0, g_0 \in \mathscr{H}(d, l, M_0, \mathscr{P})$ that

$$\mathbb{P}\left\{ \sup_{f \in \mathscr{S}_n(\delta_{\mathrm{opt}}),\, g \in \mathscr{T}_{n,\tau}(\eta_{\mathrm{opt}}; f)} \|g - g_0\|_2 \geq \frac{c_{13}}{\alpha} \max(\sigma_0, 1) \left(\eta_n^{\mathrm{subG}} + \sqrt{\frac{u}{n}} \right) \right\} \lesssim e^{-u},$$

where $c_{13} > 0$ is independent of (n, u, σ_0) and depends polynomially on $t_{\max}$.

Remark 3.4.4. In order to apply the oracle inequalities established in Section 3.4.1, it is necessary to establish an upper bound on the L_4-error of the employed DQR estimators. However, Theorem 3.4.4 only provides L_2-error bounds for DQR estimators. The current proof technique cannot directly control the L_4-error of a regression estimator using neural networks. Instead, we will use the following crude bound for a function f with $\|f\|_\infty \leq M_0$: $\|f\|_4^4 = \mathbb{E}_{X \sim \mathbb{P}_X}\{f^4(X)\} \leq M_0^2 \cdot \mathbb{E}_{X \sim \mathbb{P}_X}\{f^2(X)\} = M_0^2 \cdot \|f\|_2^2$. As a result, the L_2 convergence rate of ES estimators cannot be faster than that of the DQR estimator, even when an orthogonal score function is used. Due to the absence of results on tight L_4-error control for neural network estimators, it remains an open question whether our two-step ES estimators can achieve a faster convergence rate compared to DQR estimators when the conditional ES function is smoother than the conditional quantile

function. In general, it is natural to assume that the two functions share the same structure.

On the other hand, note that $\|f\|_4 \leq \|f\|_\infty$ for any real-valued function f. Therefore, the L_2 convergence rate of the two-step ES estimator may depend on $\|\widehat{f}_n - f_0\|_\infty^2$ instead, which is often sharper than $M_0\|\widehat{f}_n - f_0\|_2$. However, it is inherently more challenging to establish the L_∞-norm convergence rate for nonparametric QR estimators. This is achievable mostly for linear-type nonparametric estimators that display asymptotic linear approximations, such as the QR-series estimator (Belloni et al., 2019) and the (bias-corrected) kernel ridge regression estimator (Singh and Vijaykumar, 2023). The L_∞-norm convergence rate of a kernel ridge quantile regression estimator has not been investigated but is of independent interest. Recently, Imaizumi (2023) proposed a DNN estimator with a novel adversarial training scheme. The author not only derived a convergence rate for the L_∞-risk of the least squares estimator but also extended the analysis to accommodate more general loss functions, including the check loss. However, the convergence rate for the latter is sub-optimal, which leaves an open question regarding the attainment of optimal L_∞-norm convergence rates for DQR estimators. The construction of this estimator involves a preprocessing step where the output Y is transformed to yield a preprocessed output $\widehat{Y}$. Due to this preprocessing step, the L_∞ convergence rate of the proposed estimator cannot surpass that of $\widehat{Y}$; see Theorem 3 therein. Given the absence of estimators adaptable to hierarchical interaction models with an L_∞ convergence rate, the proposed estimator in Imaizumi (2023) is unsuitable for our context, where the true conditional quantile function belongs to a hierarchical interaction model.

3.5 Numerical Study

3.5.1 Monte Carlo experiments

In this section, we perform numerical studies to assess the performance of the proposed two-step deep ES regression estimator and its robust counterpart. We implement both methods in `Python` using the `PyTorch` module. We first obtain a DQR estimator $\widehat{f}_n$ by solving (3.9), and then compute the deep ES regression estimator $\widehat{g}_{n,\widehat{\tau}}$ by solving (3.11). The estimator $\widehat{g}_{n,\widehat{\tau}}$

involves a robustification parameter $\widehat{\tau} = \widehat{\tau}(n)$, which we select using a data-driven approach as follows.

Recall from Section 3.3.2 that ε is the quantile regression residual and that $\varepsilon_- = \varepsilon \wedge 0$. Assume that the (conditional) variance of ε_- is bounded by $v_2 > 0$ almost surely. In light of Theorem 3.4.6, ideally, $\widehat{\tau}$ should be selected to be of order $v_2^{1/2}(n/\log n)^{2\gamma^*/(6\gamma^*+2)}$. However, such a choice is practically infeasible because the intrinsic smoothness parameter γ^* defined in (3.2) is unknown. As a trade-off, we replace the exponent $2\gamma^*/(6\gamma^*+2)$ by 1/3, which serves as a good approximation provided that γ^* is sufficiently large. On the other hand, we use the sample variance estimator of the fitted negative QR residuals $\{\widehat{\varepsilon}_{i,-} := \min\{Y_i - \widehat{f}_n(X_i), 0\}\}_{i=1}^n$, denoted by $\widehat{v}_2$, as a proxy for the unknown noise scale v_2. Consequently, we propose a rule-of-thumb robustification parameter $\widehat{\tau} = \widehat{v}_2^{1/2}(n/\log n)^{1/3}$ that will be used throughout the numerical studies.

For DQR and two-step deep ES regression estimators, we employ fully connected ReLU neural networks with a depth of $L = 4$ and a width of $N = 256$. The network weights are optimized using the Adam optimizer (Kingma and Ba, 2014) for 200 epochs. We set the learning rate to 5×10^{-5} and use a batch size of 128. We do not employ any other regularization techniques except for early stopping, as described in Goodfellow et al. (2016). Specifically, we randomly split n i.i.d. samples into training set with n_{train} samples and validation set with $n_{\text{valid}} = n - n_{\text{train}}$ samples. We then train the neural network models on the training set with 200 epochs and select the model that minimizes the empirical L_2 error on the validation set. In our simulation, we set $n_{\text{valid}} = \lceil n/8 \rceil$.

We compare the proposed deep robust ES regression estimator (`DRES`) to several competitors: (i) the deep least squares ES estimator (`DES`) defined in (3.8); (ii) the oracle deep robust ES estimator (`oracle DRES`); (iii) the oracle deep least squares ES estimator (`oracle DES`); and (iv) the linear robust ES estimator (`LRES`) (He et al., 2023). In particular, `LRES` is an adaptive Huber linear regression estimator with surrogate response variables constructed using a plugged-in linear QR estimator. The oracle methods, `oracle DRES` and `oracle DES`, refer to

Table 3.1. The empirical mean integrated squared error $\widehat{\text{MISE}}$ (and standard error), when $d = 8, n = 1024/(5\alpha), \alpha = \{0.05, 0.1, 0.2\}$ and $\varepsilon_i \sim \mathcal{N}(0,1)$ or $\varepsilon_i \sim t_{2.5}/4$, averaged over 100 replications.

	$\varepsilon_i \sim \mathcal{N}(0,1)$			$\varepsilon_i \sim t_{2.5}/4$		
Methods	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
`DRES`	0.441 (0.006)	0.474 (0.008)	0.514 (0.007)	0.473 (0.007)	0.376 (0.005)	0.331 (0.005)
`DES`	0.463 (0.007)	0.493 (0.008)	0.527 (0.007)	0.581 (0.013)	0.450 (0.010)	0.377 (0.009)
`oracle DRES`	0.379 (0.004)	0.418 (0.006)	0.477 (0.008)	0.451 (0.009)	0.368 (0.008)	0.309 (0.005)
`oracle DES`	0.432 (0.006)	0.453 (0.006)	0.494 (0.007)	0.597 (0.016)	0.443 (0.011)	0.369 (0.010)
`LRES`	1.109 (0.003)	1.036 (0.003)	0.952 (0.003)	0.972 (0.004)	0.948 (0.003)	0.933 (0.003)

the two-step robust ES estimate (3.11) and the two-step LSE (3.8), respectively. Both methods use the true conditional quantile function f_0 to obtain the surrogate response variables. All DNN-based estimators are implemented under the same configurations as that of `DRES` described above. To assess the performance across different estimators $\widehat{g}$, we define the empirical mean integrated squared error (MISE) as

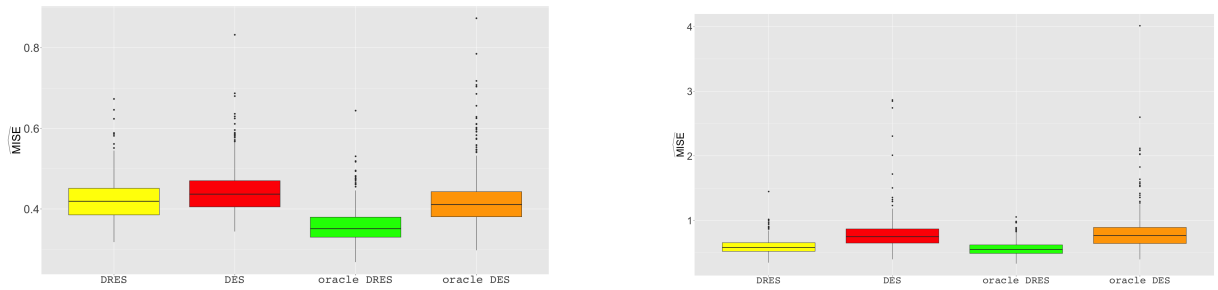
$$\widehat{\text{MISE}} = \frac{1}{T}\sum_{t=1}^{T}\left\{\widehat{g}(X_t^*) - g_0(X_t^*)\right\}^2,$$

computed using an independently generated testing set with $T = 10^5$ samples. The empirical MISE serves as an approximation to the squared L_2-error $\|\widehat{g} - g_0\|_2^2 = \mathbb{E}_{X^* \sim \mathbb{P}_X}\{|(\widehat{g} - g_0)(X^*)|^2\}$.

We generate the data $\{(X_i, Y_i)\}_{i=1}^n$ from the heteroscedastic model

$$Y_i = h_1(X_i) + h_2(X_i) \cdot \varepsilon_i,$$

where $X_i = (X_{i1}, \ldots, X_{i8})^{\mathrm{T}}$ with X_{ij} uniformly drawn from $[0,1]$, and the two functions h_1, h_2 :



(a) $\varepsilon_i \sim \mathcal{N}(0,1)$ with $n = 8192$ **(b)** $\varepsilon_i \sim t_{2.5}/4$ with $n = 8192$

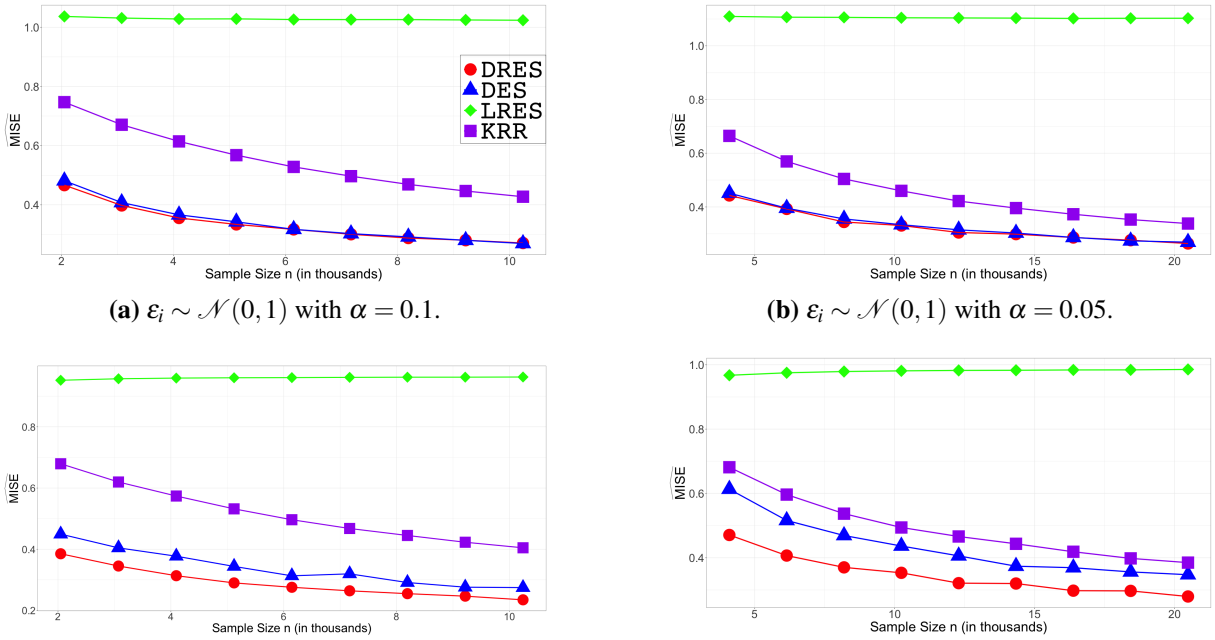
Figure 3.1. Boxplots of $\widehat{\text{MISE}}$ (based on 500 repetitions) for the four estimators (`DRES`, `DES`, `oracle DRES` and `oracle DES`) at quantile level $\alpha = 0.025$.

$\mathbb{R}^8 \to \mathbb{R}$ are defined as

$$h_1(\boldsymbol{x}) = \cos(2\pi x_1) + \frac{1}{1+e^{-x_2-x_3}} + \frac{1}{(1+x_4+x_5)^3} + \frac{1}{x_6+e^{x_7 x_8}},$$
$$h_2(\boldsymbol{x}) = \sin\left(\frac{\pi(x_1+x_2)}{2}\right) + \log(1+x_3^2 x_4^2 x_5^2) + \frac{x_8}{1+e^{-x_6-x_7}}.$$

We consider two different types of random noise ε_i: (i) $\mathcal{N}(0,1)$, the standard normal distribution (light-tailed); and (ii) $t_{2.5}/4$, the scaled t-distribution with 2.5 degrees of freedom (heavy-tailed). Since the function h_2 is nonnegative, the conditional α-level quantile and expected shortfall functions are $f_0(\boldsymbol{x}) = h_1(\boldsymbol{x}) + q_\alpha(\varepsilon) \cdot h_2(\boldsymbol{x})$ and $g_0(\boldsymbol{x}) = h_1(\boldsymbol{x}) + e_\alpha(\varepsilon) \cdot h_2(\boldsymbol{x})$, where $q_\alpha(\varepsilon)$ and $e_\alpha(\varepsilon)$ are the α-level quantile and expected shortfall of ε, respectively.

Simulation results for $n = \lceil 1024/(5\alpha) \rceil$ and the quantile level $\alpha \in \{0.05, 0.1, 0.2\}$, averaged over 100 repetitions, are reported in Table 3.1, for both random noise $\mathcal{N}(0,1)$ and $t_{2.5}/4$. We first observe the inferior performance of the linear estimator `LRES` compared to all nonparametric estimators. This performance difference is consistently observed across both light- and heavy-tailed models, regardless of the quantile level. Evidently, this discrepancy can be attributed to the misspecification of the linear model. When the noise is light-tailed, the performance of both `DRES` and `DES` remains consistent across different quantile levels. They exhibit analogous or slightly worse performance compared to the two oracle methods, which are not available in practice. However, in the presence of heavy-tailed errors, the proposed



(a) $\varepsilon_i \sim \mathcal{N}(0,1)$ with $\alpha = 0.1$.

(b) $\varepsilon_i \sim \mathcal{N}(0,1)$ with $\alpha = 0.05$.

(c) $\varepsilon_i \sim t_{2.5}/4$ with $\alpha = 0.1$.

(d) $\varepsilon_i \sim t_{2.5}/4$ with $\alpha = 0.05$.

Figure 3.2. Plots of empirical mean integrated squared error ($\widehat{\text{MISE}}$) versus sample size ranging from $\lceil 1024/(5\alpha) \rceil$ to $\lceil 1024/\alpha \rceil$ based on 100 repetitions, when ε_i follows $\mathcal{N}(0,1)$ or $t_{2.5}/4$ and $\alpha \in \{0.1, 0.05\}$.

robust estimator consistently outperforms `DES`. As a result, `DRES` demonstrates more reliable performance in the presence of heavy-tailed errors without compromising statistical efficiency under light-tailed noises.

To better demonstrate the robustness of `DES`, Figure 3.1 displays boxplots of $\widehat{\text{MISE}}$ for the DNN-based estimators (`DRES`, `DES`, `oracle DRES`, and `oracle DES`) at a quantile level of $\alpha = 0.025$, with noise following normal and t distributions. The boxplots clearly illustrate that when the noise distribution exhibits heavy tails, the least squares estimator `DES` experiences poor performance and high variability compared to the robust estimator `DRES`.

Thus far, we have compared various DNN-based estimators, with a specific focus on the DRES estimator, known for its robustness against heavy-tailed errors and efficiency in the presence of light-tailed errors. Note that after plugging in a QR estimate, in principle, any nonparametric LSE can be used to estimate the conditional ES function. In the subsequent

experiments, we implement an alternative two-step nonparametric ES estimator by combining DQR with kernel ridge regression (KRR) employing the radial basis function kernel, using the `Python` library `scikit-learn`. For simplicity, we refer to it as `KRR`. Under the same model as above, we increase the sample size from $1024/(5\alpha)$ to $1024/\alpha$ with $\alpha \in \{0.05, 0.1\}$. Figure 3.2 plots the empirical MISE versus sample size for the four ES regression estimators: `LRES`, `DES`, `DRES` and `KRR`. Due to model misspecification, the linear estimator fails to converge as the sample size increases, which is as expected. On the other hand, it is worth noting that both the least squares and robust DNN estimators consistently outperform the KRR estimator, regardless of the error distribution.

3.5.2 Upper Tail Average of Precipitation at Continental United States

The El Niño–Southern Oscillation (ENSO) is an irregular climate phenomenon characterized by periodic variations in winds and sea surface temperatures across the tropical eastern Pacific Ocean. The Climate Prediction Center in the United States (US) defines *El Niño conditions* (or *La Niño conditions*) when the sea surface temperature in the Niño-3.4 region of the equatorial Pacific Ocean deviates more than 0.5°C above (or below) the normal temperature for the same period. Substantial anomalies in seasonal precipitation have been associated with its warm (El Niño) and cool (La Niña) phases (Kahya and Dracup, 1993; Ropelewski and Halpert, 1986, 1996). Specifically, recent research indicates that ENSO may be associated with regional increased rainfall variability (Yun et al., 2021). Consequently, it is important to understand the relationship between ENSO and the upper tail average of precipitation.

We analyze the influence of El Niño on the upper tail average of precipitation across the continental US. To this end, we apply our proposed methodology to the US precipitation reanalysis data set (Slivinski et al., 2019). This data set comprises of daily precipitation measurements in millimeters, which are derived from reanalysis by integrating a wide range of observational data and numerical modeling. The data set covers 819 grid points within the continental US at a $1^\circ \times 1^\circ$ spatial resolution from year 1950 to year 2015. Subsequently, we pre-process the data by

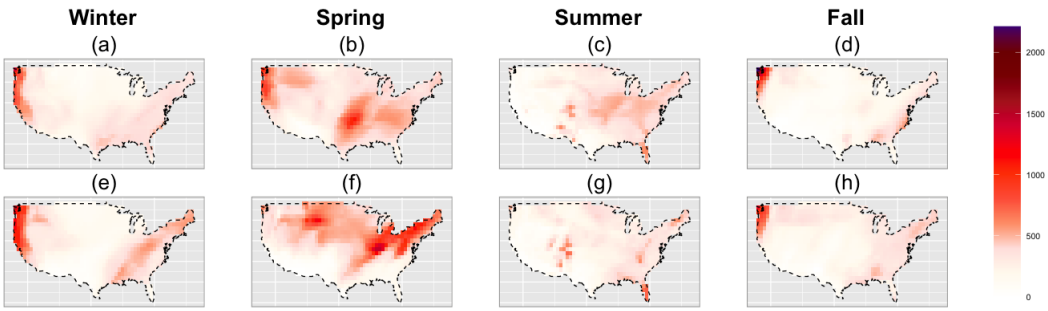


Figure 3.3. Subfigures (a)–(d) illustrate the predicted precipitation during periods of El Niño event in year 2010. Subfigures (e)–(h) display the predicted precipitation when El Niño is not in progress in year 2010.

summing the daily precipitation to obtain a monthly total precipitation. This pre-processed data is then grouped into four seasonal categories: winter (December to February), spring (March to May), summer (June to August), and fall (September to November). In the case of the winter season, the pre-processed dataset comprises 161,343 data points, while each of the other seasonal categories contains 162,162 data points.

For each seasonal dataset, we fit the proposed robust expected shortfall regression at $\alpha = 0.9$ (upper tail), where the robustification parameter is tuned by the procedure described in Section 3.5.1. The covariate of interest is the Niño-3.4 index, and we adjust for other variables including the year, and latitude and longitude for each location in the continental US. To calculate the Niño-3.4 index, we compute monthly averages of sea surface temperatures in Niño-3.4 region, subtract the annual mean temperature, and subsequently normalize the data to have mean zero and standard deviation one. We note that estimating conditional upper tail averages at $\alpha = 0.9$ is equivalent to fitting the proposed `DRES` method at level $1 - \alpha$ after flipping the sign of the response. We estimate the conditional quantile function at level α via a DQR estimator, which solves (3.9). For the implementation in `Python`, we employ fully connected ReLU neural networks with a depth of $L = 6$ and a width of $N = 256$ using `PyTorch`. We set the learning rate to 10^{-3} and use a batch size of 256 for 500 epochs. Similar to the approach in Section 3.5.1, we utilize the early stopping with the number of validation data $n_{\text{valid}} = \lceil n/8 \rceil$.

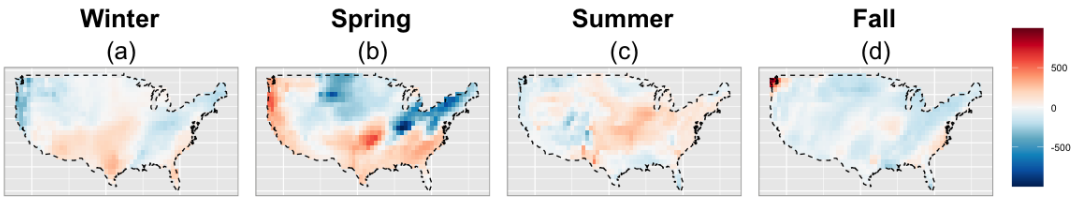


Figure 3.4. The discrepancies between predicted ES precipitation during periods of El Niño event and those in the absence of El Niño conditions for each season.

Figure 3.3 presents the predicted (upper tail) conditional ES of precipitation values for each location in year 2010 at level $\alpha = 0.9$. Subfigures (a)–(d) in the figure present the predicted precipitation during El Niño events, with the Niño-3.4 index set to 2.0. The remaining subfigures show predictions for non-El Niño periods, with the Niño-3.4 index set to 0. We see from Figure 3.3 that the West Coast is predicted to experience significantly higher levels of precipitation compared to other regions in most seasons.

To understand the impact of El Niño on the (upper tail) ES of precipitation, we calculate the discrepancies between predicted ES precipitation during periods of El Niño event and those in the absence of El Niño conditions for each season, illustrated in Figure 3.4. Our results reveal that the impact of El Niño exhibits spatial and seasonal variation. In particular, during the winter and spring seasons, we observe that in the presence of El Niño, the north region is predicted to experience drier weather compared to the normal condition, while conversely, the south region becomes wetter. We remark that this pattern during the winter and spring aligns with a well-known teleconnection known as the north-south seesaw in precipitation in climate literature (Becker, Berbery and Higgins, 2009; Cayan, Redmond and Riddle, 1999; Dettinger et al., 1998; Mo and Higgins, 1998). We enhance the current findings by offering a detailed description of how El Niño affects the upper percentile of precipitation averages.

To further illustrate the predictive capability of our estimators for heavy rainfall, we examine a case study involving the devastating floods in Texas and Oklahoma in May 2015. This particular month marked a historical record as the wettest May and the all-time wettest month in the United States, based on 121 years of recorded data (Terti et al., 2019), which leads

to deadly flash floods. Specifically, in Dallas, Texas, May 2015 witnessed a total precipitation of 1175 mm, a substantial departure from the average spring precipitation of 379 mm from the year 1950 to the year 2010. It's important to note that the Niño-3.4 index for May 2015 measured 1.60. Our DRES estimator predicts 1070 mm during El Niño events, and 449 mm for non-El Niño periods. This result suggests that the extreme rainfall in Dallas in May 2015 could be largely associated with the El Niño events. Concurrently, we also implemented deep least squares (mean) regression using the same neural network configuration. The deep mean regression estimator predicts 760 mm during El Niño events, and 362 mm for non-El Niño events. This result demonstrates the importance of using the (conditional) upper-tail average to predict extreme rainfall events, in contrast to the use of least squares regression methods that only focus on centrality. Consequently, estimating the (conditional) upper-tail average is a more effective method for predicting extreme rainfall events, enabling local water management to take early precautions to mitigate flooding effectively.

3.6 Acknowledgements

This chapter, in full, is currently being prepared for submission for publication of the material. Yu, Myeonghun; Tan, Kean Ming; Huixia Judy Wang; Zhou, Wen-Xin. The dissertation author was the primary investigator and author of this material.

Chapter 4, in full, is currently being prepared for submission for publication of the material. Yu, Myeonghun; Wang, Yue; Xie, Siyu; Tan, Kean Ming; Zhou, Wen-Xin. The dissertation author was the primary investigator and author of this material.

Chapter 4

Estimation and Inference for Nonparametric Expected Shortfall Regression over RKHS

4.1 Introduction

Since the seminal work of Koenker and Bassett (1978), quantile regression (QR) has become a valuable statistical tool that provides deeper insights into distributional properties beyond traditional mean-based regression models. This approach is particularly crucial when the tails, either left or right, of the response distribution carry significant implications, such as low birth weight, high precipitation or temperature, and low earnings or test scores. By examining multiple quantiles, analysts gain a more comprehensive understanding of the factors influencing different segments of the data. This enhances the robustness of statistical inferences and facilitates a more comprehensive interpretation of complex relationships in various fields, including finance, economics, and social sciences. We refer to Koenker (2005) and Koenker et al. (2017) for a comprehensive overview of quantile regression methods, theory, computation, and various applications.

In the application of quantile regression across diverse domains, practitioners commonly execute a series of models at predefined quantile levels, such as 5%, 10%, 25%, 50%, 75%, 90%, and 95%, wherein they report the estimated coefficients along with the associated p-values

or 95% confidence intervals. Despite the wealth of information provided by these summary statistics, surpassing the insights offered by least squares estimates, their interpretation often remains somewhat ambiguous. Illustratively, consider a study conducted by Coronese et al. (2019) investigating the impact of natural disasters on economic damages, where the response variable denotes yearly economic damages in US dollars (USD billion). In this context, if the projected 95% single-event damage for the year 2010 is estimated at $10 billion, it signifies a 5% probability of surpassing the $10 billion damage threshold. However, no specific quantitative details are available regarding the extent of damages beyond this threshold. Consequently, an additional question arises: How can we model and forecast the average damages incurred by the most severely affected 5% of cases? Another potential concern arises when the statistical significance of a predictor varies across different quantiles. For instance, a predictor that demonstrates 5% statistical significance in a conditional 90% quantile model may lose this significance at conditional 87.5% and 92.5% quantile models. This variability can result in inconclusive findings, highlighting the sensitivity of the results to the chosen quantile levels.

To address the limitations associated with quantiles, we consider a set of functionals that extend beyond mean and quantile measures, providing coverage for flexible prespecified regions within a distribution. Let $Y \in \mathbb{R}$ represent a generic dependent variable of interest, which could represent earnings, test scores, precipitation, temperature, and economic damage, among others. We denote its cumulative distribution function (CDF) and quantile function as $F_Y(y) = \mathbb{P}(Y \leq y)$, for $y \in \mathbb{R}$, and $q_Y(\tau) = \inf\{y \in \mathbb{R} : F_Y(y) \geq \tau\}$, for $\tau \in (0,1)$. It is worth recalling that $\mathbb{E}(Y) = \int_{-\infty}^{\infty} y \mathrm{d}F_Y(y) = \int_0^1 q_Y(u)\mathrm{d}u$. At level $\tau \in (0,1)$, we define the left-tail average of Y as

$$e_Y(\tau) = \frac{1}{\tau}\int_0^\tau q_Y(u)\mathrm{d}u. \tag{4.1}$$

Under this notation, the right-tail average of Y (at level $1-\tau$) is $\frac{1}{\tau}\int_{1-\tau}^1 q_Y(u)\mathrm{d}u$, which, through a change of variable, is equivalent to $-e_{-Y}(\tau)$, the negative left-tail average of Y at level τ. In

particular, $e_Y(1) = \mathbb{E}(Y)$. The functional $e_Y(\tau)$ is also recognized as the expected shortfall (ES) or conditional value-at-risk, a widely employed risk measure in operations research (Rockafellar and Uryasev, 2000) and quantitative risk management (McNeil et al., 2015), with applications in banking, insurance and actuarial science. We refer to $e_Y(\tau)$ in (4.1) as the τ-th ES in this work, which is a natural coupling to the τ-th quantile $q_Y(\tau)$ because if F_Y is continuous at $q_Y(\tau)$, $e_Y(\tau)$ can be equivalently written as $e_Y(\tau) = \mathbb{E}\{Y|Y \leq q_Y(\tau)\}$.

In the presence of explanatory variables $X \in \mathbb{R}^d$, a variety of methods–ranging from parametric and semiparametric to nonparametric–have been developed for estimating and inferring the conditional ES of Y given X. Notable contributions include those by Dimitriadis and Bayer (2019), Patton et al. (2019), Taylor (2019), Barendse (2020), Guillen, Bermúdez and Pitarque (2021), Peng and Wang (2023) and He et al. (2023), which offer (semi-)parametric approaches under either joint parametric quantile and ES models or specific families of response distributions. In the context of nonparametric ES regression, we refer to Scaillet (2005), Cai and Wang (2008), Kato (2012), Linton and Xiao (2013), Martins-Filho, Yao and Torero (2018), Olma (2021) and Fissler, Merz and Wüthrich (2023), among others. It is worth noting that a majority of prevailing nonparametric conditional ES estimators rely on Nadaraya-Watson and local linear methods, along with various adaptations. Consequently, these methods are particularly well-suited for low-dimensional settings, such as $1 \leq d \leq 3$. To effectively address covariates of moderate dimensionality, Fissler, Merz and Wüthrich (2023) proposed joint quantile and ES regression estimators using a joint loss function and deep neural networks (DNNs). On the downside, DNNs often require large amounts of training data to generalize well and are prone to overfitting. The process of finding the optimal combination of hyperparameters for DNNs may require extensive experimentation. From a different perspective, Chetverikov, Liu and Tsyvinski (2022) considered a semiparametric model in which the conditional ES function is linear. Meanwhile, the nuisance conditional CDF of Y given X is estimated nonparametrically by a version of the random forest (RF) method. To establish the asymptotic normality of the two-step linear ES estimator, the preliminary conditional CDF estimator must satisfy a high-level

uniform consistency requirement. However, the justification for this requirement, particularly regarding the RF estimator, remains unclear. For technical reasons, their approach also relies on sample splitting, meaning the conditional CDF and ES regression coefficients are estimated on different subsamples.

This paper aims to propose efficient estimation and inference methods for nonparametric expected shortfall regression, helping bridge the gap between model flexibility and complexity. To address the practical concerns associated with local polynomial and DNN regressions mentioned earlier, we focus on nonparametric regression over reproducing kernel Hilbert spaces (RKHSs) (Schölkopf and Smola, 2002). Specifically, we assume that both the conditional quantile function and the conditional ES function of the response variable Y given the input covariates $X \in \mathbb{R}^d$ belong to RKHSs. RKHS regression provides flexibility and nonlinear modeling, as different choices of kernel functions capture various types of nonlinear relationships in the data. The "kernel trick" implicitly maps the input data into a higher-dimensional space without explicitly computing the transformed features. Moreover, proper tuning of the ridge regularization parameter helps prevent overfitting and improves generalization ability.

Motivated by the use of an orthogonal score for parametric ES regression (Barendse, 2020), we propose a fully nonparametric two-step method for fitting nonlinear conditional ES functions. In the first step, we estimate the conditional quantile function through kernel ridge regression (KRR) with the check loss and derive finite-sample, high probability bounds for the resulting estimator in both L_2 and RKHS norms; see Theorem 4.3.1. These intermediate results complement existing asymptotic convergence results for quantile KRR (Li et al., 2007; Lian, 2022) and are of independent interest. In the second step, we apply least squares KRR to estimate the conditional ES function, using the quantile KRR estimates as surrogate response variables. Both steps involve only convex optimization, and there is no need for sample splitting to facilitate the corresponding theoretical analysis.

Subject to a high-level restriction on the accuracy of the quantile KRR estimator, we establish finite-sample convergence rates for the two-step ES KRR estimator in Theorem 4.3.2.

More specifically, we provide these rates in terms of exponential-type deviation bounds. For conducting inference, we further establish a non-asymptotic functional Bahadur representation (Theorem 4.3.3), which allows for an explicit characterization of Gaussian approximation error bounds as functions of the effective dimension, sample size, regularization parameters, and QR estimation error; see Theorem 4.3.4. As a byproduct, we also present parallel results for the oracle two-step ES KRR estimator obtained by inserting the true conditional quantile function. These results not only demonstrate that the impact of nonparametric QR estimation is first-order negligible but also improve upon the best available results in KRR inference theory (Shang and Cheng, 2013). Due to space constraints, we provide instantiations of the general bounds on estimation error and Gaussian approximation error for various RKHSs in the supplementary material.

Due to the complex nature of the asymptotic variance, the Gaussian approximation results mentioned above are instructive but not directly applicable in practice. To address this limitation, we employ a multiplier/weighted bootstrap procedure to construct pointwise confidence intervals and provide rigorous theoretical guarantees for its validity; see Theorems 4.3.5 and 4.3.6. Guided by bootstrap approximation theory, we propose a reduced-form bootstrap statistic. This eliminates the need for solving weighted KRR repeatedly, thus considerably reducing the computational cost.

We apply our method to medical expense data created by Lantz (2013), which uses demographic statistics from the U.S. Census Bureau. The goal is to examine a key observation in insurance claim size modeling that covariates may have different effects on the claim size distribution. For instance, the age of the beneficiary enrolled in an insurance plan may be a crucial variable in explaining systematic effects on large medical expenses charged to the plan but may be irrelevant in describing such effects in average charges. Through this relatively small dataset, we demonstrate the potential of nonparametric ES regression techniques for flexible insurance claim size modeling, as opposed to relying on the popular gamma model. Our objective is not to replace quantile regression but to provide an additional regression tool and insights to

be incorporated into a broader system of risk analysis and tail learning.

The rest of this paper is organized as follows. We begin with a brief introduction to the joint quantile and ES regression framework in Section 4.2. After providing some preliminaries on RKHSs in Section 4.2.1, we propose the two-step ES estimator employing an orthogonal score function in Section 4.2.2. The multiplier bootstrap procedure for statistical inference is introduced in Section 4.2.3. In Section 4.3, we begin our analysis by deriving exponential-type deviation bounds for the proposed estimators under general RKHSs. In Section 4.3.2, we derive functional Bahadur representations and Berry-Esseen bounds for the proposed estimator, and establish the theoretical validity of the bootstrap procedure. Section 4.4 examines the finite-sample performance and usefulness of the proposed estimator through numerical experiments and a real data demonstration. We conclude the paper with a discussion in Section 4.5. Instantiations of the general bounds for various RKHSs in Section 4.3 and the proofs of all theoretical results are deferred to the supplementary materials.

NOTATION. We use $c_1, c_2, \ldots$ to denote the global constants employed in the statements and proofs of theorems, propositions, corollaries, and lemmas. On the other hand, $C_1, C_2, \ldots$ denote local intermediate constants within the proofs and may vary from one line to another. For two sequences of real numbers $\{a_i\}_{i\geq 1}, \{b_i\}_{i\geq 1}$, we write $a_i \lesssim b_i$ if there exists a constant $C > 0$ independent of i such that $a_i \leq C b_i$ for all $i \geq 1$, and $a_i \gtrsim b_i$ if $b_i \lesssim a_i$. Moreover, we write $a_i \asymp b_i$ if $a_i \lesssim b_i$ and $a_i \gtrsim b_i$. For the sample size, we assume $n \geq 3$ throughout the paper, ensuring $\log n \geq 1$.

4.2 Model Setup and Methodologies

Let $\{(Y_i, X_i)\}_{i=1}^n$ be n independent random samples from $(Y, X) \in \mathbb{R} \times \mathscr{X}$, where $\mathscr{X} \subseteq \mathbb{R}^d$ is a compact subset. Here, Y is a real-valued response variable and $X \in \mathbb{R}^d$ is a d-dimensional vector of random covariates. For simplicity, we assume that $\mathscr{X} = [0,1]^d$ is the unit cube in $\mathbb{R}^d$. Let $F_Y(\cdot|X)$ be the conditional CDF of Y given X. Given a $\tau \in (0,1)$, the conditional τ-th

quantile and expected shortfall of Y given X are written, respectively, as

$$q_Y(\tau|X) = \inf\{y \in \mathbb{R} : F_Y(y|X) \geq \tau\} \quad \text{and} \quad e_Y(\tau|X) = \mathbb{E}\{Y|Y \leq q_Y(\tau|X), X\}.$$

We consider the following nonparametric models for the conditional quantile and ES:

$$q_{Y_i}(\tau|X_i) = f_0(X_i) \text{ and } e_{Y_i}(\tau|X_i) = g_0(X_i), \tag{4.2}$$

where $f_0, g_0 : [0,1]^d \to \mathbb{R}$ are two unknown functions satisfying $\mathbb{P}\{Y \leq f_0(X)|X\} = \tau$ almost surely and $g_0(x) = \mathbb{E}\{Y|Y \leq f_0(X), X = x\}$ for $x \in [0,1]^d$. Under this assumption, it is well-known that f_0 minimizes the population check loss objective $\mathscr{Q}_\tau(f) - \mathscr{Q}_\tau(0)$ over all functions $f : \mathscr{X} \to \mathbb{R}$, where $\mathscr{Q}_\tau(f) := \mathbb{E}\{\rho_\tau(Y - f(X))\}$ and ρ_τ is the check function defined as $\rho_\tau(u) = u\{\tau - \mathbb{1}(u \leq 0)\}$ (Koenker and Bassett, 1978). In addition, provided that $\mathbb{E}(Y^2) < \infty$, the true conditional ES function g_0 minimizes the expected truncated squared error loss $\mathbb{E}[(Y - g(X))^2 \mathbb{1}\{Y \leq f_0(X)\}]$ over all functions $g : \mathscr{X} \to \mathbb{R}$.

Our main objective is to develop inference methods for the nonparametric ES regression function g_0. Specifically, we aim to construct (asymptotically) $100 \cdot (1-\alpha)\%$ (e.g., $\alpha = 0.05$) pointwise confidence intervals $\mathscr{C}_\alpha(x) = [\widehat{g}^l(x), \widehat{g}^u(x)]$ for $g_0(x)$, satisfying that $\mathbb{P}\{g_0(x) \in \mathscr{C}_\alpha(x)\} \to 1-\alpha$ as $n \to \infty$ for each $x \in \mathscr{X}$. Here, the probability is taken with respect to the training sample $\{(X_i, Y_i)\}_{i=1}^n$ used to construct the confidence intervals.

4.2.1 Preliminaries on RKHS

Assume that the marginal distribution $\mathbb{P}_X$ of $X \in \mathscr{X}$ is non-degenerate. Let $L_2(\mathbb{P}_X)$ be the Hilbert space of functions $f : \mathscr{X} \to \mathbb{R}$ square-integrable with respect to $\mathbb{P}_X$, that is, $L_2(\mathbb{P}_X) = \{f : \mathscr{X} \to \mathbb{R} \,|\, \int_{\mathscr{X}} f^2(x) \mathrm{d}\mathbb{P}_X(x) < \infty\}$. Denote by $\|\cdot\|_2 = \|\cdot\|_{L_2(\mathbb{P}_X)}$ the L_2-norm in the space $L_2(\mathbb{P}_X)$ induced by the inner product $\langle f, g\rangle_2 = \langle f, g\rangle_{L_2(\mathbb{P}_X)} = \int_{\mathscr{X}} f(x)g(x)\mathrm{d}\mathbb{P}_X(x)$.

Let $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be a continuous, symmetric, and positive semidefinite kernel

function, known as a Mercer kernel. Define the function $K_x : \mathscr{X} \to \mathbb{R}$ as $K_x(x') = K(x,x')$ for any $x,x' \in \mathscr{X}$. A reproducing kernel Hilbert space (RKHS) $\mathscr{H} = \mathscr{H}_K$ associated with the Mercer kernel K is defined as the completion of the linear span of $\{K_x : x \in \mathscr{X}\}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$, satisfying $\langle K_x, K_{x'} \rangle_{\mathscr{H}} = K(x,x')$ (Aronszajn, 1950). Denote by $\| \cdot \|_{\mathscr{H}}$ the RKHS norm induced by $\langle \cdot, \cdot \rangle_{\mathscr{H}}$. For any $r > 0$, let $\mathbb{B}_{\mathscr{H}}(r)$ be the ball of radius r with respect to the RKHS norm, i.e., $\mathbb{B}_{\mathscr{H}}(r) = \{f \in \mathscr{H} : \|f\|_{\mathscr{H}} \leq r\}$. For every $f \in \mathscr{H}$ and $x \in \mathscr{X}$, the well-known reproducing property states that $f(x) = \langle K_x, f \rangle_{\mathscr{H}}$. Moreover, define the integral operator $T_K : L_2(\mathbb{P}_X) \to L_2(\mathbb{P}_X)$

$$T_K(f)(x) = \int_{\mathscr{X}} K(x,x') f(x') \mathrm{d}\mathbb{P}_X(x'), \quad \forall f \in L_2(\mathbb{P}_X), x \in \mathscr{X}. \tag{4.3}$$

Throughout the manuscript, we impose the following boundedness condition on K.

Condition 4.2.1. The kernel function is uniformly bounded, that is, $\sup_{x \in \mathscr{X}} \sqrt{K(x,x)} \leq 1$.

By compactness of $\mathscr{X}$ and boundedness of the kernel, the Mercer's theorem ensures the existence of a sequence of eigenfunctions $\{\phi_j\}_{j \geq 1}$ that form an orthonormal basis of $L_2(\mathbb{P}_X)$, and an associated set of non-negative eigenvalues $\{\mu_j\}_{j \geq 1}$ such that

$$K(x,x') = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(x') \text{ and } T_K(\phi_j) = \mu_j \phi_j(x), \quad j = 1, 2, \ldots, \tag{4.4}$$

where the convergence of the infinite series holds absolutely and uniformly on $\mathscr{X} \times \mathscr{X}$. Without loss of generality, we assume that $\{\mu_j\}_{j \geq 1}$ is non-increasing. With this Mercer expansion, the squared RKHS norm takes the form $\|f\|^2_{\mathscr{H}} = \sum_{j=1}^{\infty} f_j^2 / \mu_j$, where $f_j = \int_{\mathscr{X}} f(x) \phi_j(x) \mathrm{d}\mathbb{P}_X(x) = \langle f, \phi_j \rangle_2$. Consequently, the RKHS $\mathscr{H}$ can be written as $\mathscr{H} = \{f = \sum_{j=1}^{\infty} f_j \phi_j \,|\, \sum_{j=1}^{\infty} f_j^2 / \mu_j < \infty\}$. By the spectral decomposition of $K(\cdot,\cdot)$, we have $T_K(f)(x) = \sum_{j=1}^{\infty} \mu_j \langle \phi_j, f \rangle_2 \, \phi_j(x)$ for any $x \in \mathscr{X}$ and $f \in L_2(\mathbb{P}_X)$. For any $r \geq 0$, we define the r-th power of T_K as $T_K^r(f)(x) = \sum_{j=1}^{\infty} \mu_j^r \langle \phi_j, f \rangle_2 \, \phi_j(x)$.

For the purposes of estimation and inference, we define $\mathscr{F}$ as $\{f = f' + b : f' \in \mathscr{H}, b \in$

$\mathbb{R}\}$. Since the commonly used Gaussian RKHS does not include non-zero constant functions (Minh, 2010), the use of $\mathscr{F}$ allows for a more flexible setting. Throughout the following discussion, we will use the notation f' for any function that belongs to an RKHS without an additional intercept term. Following the approach of Zhang, Liu and Wu (2016), we assume that each f in $\mathscr{F}$ can be uniquely decomposed as $f' + b$, with $f' \in \mathscr{H}$ and $b \in \mathbb{R}$.

4.2.2 Expected shortfall regression in RKHS

We propose a two-step procedure for estimating the conditional ES function $g_0 : \mathscr{X} \to \mathbb{R}$ in (4.2) in the context of RKHS by treating the conditional quantile function f_0 as a nuisance parameter (Barendse, 2020; He et al., 2023). Let $S_\tau(q,e;Y) = \tau q + \tau(Y-q)\mathbb{1}(Y \le q) - \tau e$ with $q,e \in \mathbb{R}$ be a score function that satisfies the moment condition $\mathbb{E}\{S_\tau(f_0(X), g_0(X); Y)|X\} = 0$ almost surely. Assuming that the conditional distribution of Y given X, denoted as $F_{Y|X}$, is continuous, we have the following orthogonality property:

$$\frac{\partial}{\partial q}\mathbb{E}\{S_\tau(q,e;Y)|X\}\bigg|_{q=f_0(X)} = \tau - F_{Y|X}(f_0(X)) = 0. \tag{4.5}$$

At the first step, we estimate f_0 nonparametrically via a *kernel ridge regression*:

$$\widehat{f} = \widehat{f}_n(\lambda_q) \in \underset{f \in \mathscr{F}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - f(X_i)) + \lambda_q \|f'\|^2_{\mathscr{H}} \right\}, \tag{4.6}$$

where $\lambda_q > 0$ is a regularization parameter. We refer to $\widehat{f}$ as the quantile KRR (Q-KRR) estimator (Takeuchi et al., 2006; Li et al., 2007). By the representer theorem (Kimeldorf and Wahba, 1971), it suffices to consider output functions that belong to the span of the fundamental functions defined by the kernel K and the training sample, i.e., $\{K(X_i,\cdot)\}_{i=1}^n$, possibly including an intercept term. Using the parameterization $f(\cdot) = f'(\cdot) + b = \sum_{i=1}^n \alpha_i K(X_i,\cdot) + b$, optimization

problem (4.6) can be reformulated as

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^n, b\in\mathbb{R}}{\text{minimize}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(Y_i - b - \sum_{j=1}^{n} \alpha_j K(X_i, X_j)\right) + \lambda_q \cdot \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha} \right\}, \tag{4.7}$$

where $\mathbf{K} = (K(X_i, X_j))_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$ is the kernel matrix and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$.

At the second step, we construct surrogate response variables $\widehat{Z}_i = Z_i(\widehat{f})$ using the nonparametric quantile estimate obtained at the first step, where

$$Z_i(f) = \tau f(X_i) + \{Y_i - f(X_i)\}\mathbb{1}\{Y_i \le f(X_i)\} \tag{4.8}$$

for any function $f : \mathscr{X} \to \mathbb{R}$. We then propose a two-step ES kernel ridge regression (ES-KRR) estimator $\widehat{g} = \widehat{g}_n(\lambda_e)$ of g_0 defined as

$$\widehat{g} \in \underset{g\in\mathscr{F}}{\operatorname{argmin}} \widehat{\mathscr{L}_n}(\widehat{f}, g) \quad \text{with} \quad \widehat{\mathscr{L}_n}(f, g) = \frac{1}{n}\sum_{i=1}^{n} \{Z_i(f)/\tau - g(X_i)\}^2 + \lambda_e \|g'\|^2_{\mathscr{H}}, \tag{4.9}$$

where $\lambda_e > 0$ is a second regularization parameter. Similarly, (4.9) can be rewritten as

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^n, b\in\mathbb{R}}{\text{minimize}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{Z}_i/\tau - b - \sum_{j=1}^{n} \alpha_j K(X_i, X_j)\right)^2 + \lambda_e \cdot \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha} \right\}. \tag{4.10}$$

Thus, we have $\widehat{g}(\cdot) = \widehat{b} + \sum_{j=1}^{n} \widehat{\alpha}_j K(X_j, \cdot)$ where $(\widehat{\boldsymbol{\alpha}}, \widehat{b})$ is the solution to (4.10).

4.2.3 Pointwise inference with multiplier bootstrap

In this section, we propose a framework for conducting pointwise inference on the conditional ES function g_0 in (4.2). Specifically, we propose a multiplier bootstrap procedure to construct asymptotically valid confidence intervals for g_0 at any predetermined $x_0 \in \mathscr{X}$.

Let $W_1, \ldots, W_n \sim W$ be independently generated random weights that satisfy $\mathbb{E}(W) = 1$

and $\mathrm{Var}(W) = 1$. We define the bootstrap ES-KRR estimator $\widehat{g}^\flat = \widehat{g}^\flat_n(\lambda_e)$ as

$$\widehat{g}^\flat \in \operatorname*{argmin}_{g \in \mathscr{F}} \widehat{\mathscr{L}^\flat_n}(\widehat{f}, g) \quad \text{with} \quad \widehat{\mathscr{L}^\flat_n}(f,g) = \frac{1}{n}\sum_{i=1}^{n} W_i \{Z_i(f)/\tau - g(X_i)\}^2 + \lambda_e \|g'\|^2_{\mathscr{H}}. \tag{4.11}$$

Denote by $\mathbb{P}^*$ and $\mathbb{E}^*$ the conditional probability and expectation given $\mathscr{D}_n$, respectively, and note that $\mathbb{E}^*\{\widehat{\mathscr{L}^\flat_n}(\widehat{f}, g)\} = \widehat{\mathscr{L}_n}(\widehat{f}, g)$. Therefore, $\widehat{\mathscr{L}^\flat_n}$ is a conditionally unbiased "estimate" of $\widehat{\mathscr{L}_n}$ in the bootstrap world, so that $\widehat{g}^\flat$ can be viewed as the bootstrap estimator of $\widehat{g}$.

To preserve the convexity of the loss function $\widehat{\mathscr{L}^\flat_n}$, non-negative weights are typically preferred. Two commonly used choices are $W_i \sim \mathrm{Exp}(1)$, an exponential distribution with rate 1, and $W_i \sim 1 + e_i$, where $e_i \sim \mathrm{Unif}(\{-1, 1\})$ follows the Rademacher distribution. Using a similar re-parametrization as in (4.10), we can compute $\widehat{g}^\flat$ by solving a quadratic program. Then, for a prescribed nominal level $\alpha \in (0,1)$, we can construct confidence intervals for $g_0(x_0)$ via the percentile, pivotal, or normal-based methods.

The computational complexity of the algorithm in Saunders, Gammerman and Vovk (1998) for solving KRR problems scales as $\mathscr{O}(n^3)$. Consequently, the complexity of computing $\widehat{g}^\flat$ increases to $\mathscr{O}(Bn^3)$, where B is the number of bootstrap samples. To mitigate the computational cost, various algorithms such as the divide-and-conquer (Zhang, Duchi and Wainwright, 2013), Nyström method (Williams and Seeger, 2000), and randomized sketches (Yang, Pilanci and Wainwright, 2017), can be used to approximate the solution.

In Section 4.3.2, we will establish a functional Bahadur representation of the ES-KRR estimator $\widehat{g}$ around the population (penalized) risk minimizer g_{λ_e}, defined as

$$g_{\lambda_e} = \operatorname*{argmin}_{g \in \mathscr{F}} \Big[\mathbb{E}\{Z_i(f_0)/\tau - g(X_i)\}^2 + \lambda_e \|g'\|^2_{\mathscr{H}}\Big]. \tag{4.12}$$

Specifically, Theorem 4.3.3 provides an upper bound on the difference between $\tau(\widehat{g} - g_{\lambda_e})$ and $n^{-1}\sum_{i=1}^n \{Z_i(f_0) - \tau g_0(X_i)\}(T_K + \lambda_e I)^{-1} K_{X_i}$ under the supremum norm, where T_K is the integral operator defined in (4.3), I is the identity operator, and $K_{X_i}(\cdot) = K(X_i, \cdot)$. Motivated by this

Algorithm 2. Pointwise confidence interval construction using weighted bootstrap

Input: Training data $\{(Y_i, X_i)\}_{i=1}^n$, number of bootstrap samples B, regularization parameters λ_q and λ_e, and nominal level $\alpha \in (0,1)$.

1: Compute the Q-KRR and ES-KRR estimators defined in (4.6) and (4.9), respectively.
2: Calculate the n-vector $\mathbf{v}_{x_0} = (v_{x_0,1}, \ldots, v_{x_0,n})^{\mathrm{T}}$ defined in (4.13).
3: **for** $b = 1, 2, \ldots, B$ **do**
4: Generate independent random weights $W_1^{(b)}, \ldots, W_n^{(b)} \sim W$.
5: Compute bootstrap statistics $\mathfrak{B}_b^\flat(x_0) = (1/n)\sum_{i=1}^n (W_i^{(b)} - 1)\{Z_i(\widehat{f})/\tau - \widehat{g}(X_i)\} v_{x_0,i}$.
6: **end for**
7: Compute the upper $(\alpha/2)$-th and $(1-\alpha/2)$-th sample quantiles of bootstrap statistics $\{\mathfrak{B}_b^\flat(x_0)\}_{b=1}^B$, denoted by $\widehat{u}_{\alpha/2}^\flat$ and $\widehat{u}_{1-\alpha/2}^\flat$, respectively.

Output: α-level confidence interval $[\widehat{g}(x_0) - \widehat{u}_{\alpha/2}^\flat, \widehat{g}(x_0) - \widehat{u}_{1-\alpha/2}^\flat]$.

representation, for any $x_0 \in \mathscr{X}$, we can bypass the need to solve the quadratic program in (4.11) for each bootstrap iteration by approximating the distribution of $\widehat{g}(x_0) - g_0(x_0)$ with the following quantity when the bias $g_{\lambda_e}(x_0) - g_0(x_0)$ is negligible:

$$\mathfrak{B}^\flat(x_0) := \frac{1}{n}\sum_{i=1}^n (W_i - 1)\{Z_i(\widehat{f})/\tau - \widehat{g}(X_i)\}(\widehat{T} + \lambda_e I)^{-1} K_{X_i}(x_0).$$

Here, $\widehat{T} = (1/n)\sum_{i=1}^n K_{X_i} \otimes K_{X_i}$ is the empirical integral operator satisfying $\mathbb{E}(\widehat{T}) = T_K$, where the tensor product $K_{X_i} \otimes K_{X_i} : \mathscr{H} \to \mathscr{H}$ is defined as $(K_{X_i} \otimes K_{X_i})(h) := \langle K_{X_i}, h\rangle_{\mathscr{H}} K_{X_i} = h(X_i) K_{X_i}$.

To implement the proposed procedure, motivated by Singh and Vijaykumar (2023), we provide an equivalent representation of $\mathfrak{B}^\flat(x_0)$ in Proposition 4.2.1. Let

$$\mathbf{v}_{x_0} = (v_{x_0,1}, \ldots, v_{x_0,n})^{\mathrm{T}} = (\mathbf{K}/n + \lambda_e \mathbf{I}_n)^{-1} \mathbf{k}_{x_0} \in \mathbb{R}^n, \tag{4.13}$$

where $\mathbf{k}_{x_0} = (K(X_1, x_0), \ldots, K(X_n, x_0))^{\mathrm{T}} \in \mathbb{R}^n$, and $\mathbf{I}_n$ is the identity matrix of size n.

Proposition 4.2.1 (Equivalent representation of $\mathfrak{B}^\flat(x_0)$)**.** Let $\mathbf{v}_{x_0} = (v_{x_0,1}, \ldots, v_{x_0,n})^{\mathrm{T}}$ be the n-vector defined in (4.13). The above bootstrap statistic $\mathfrak{B}^\flat(x_0)$ can be written as

$$\mathfrak{B}^\flat(x_0) = \frac{1}{n}\sum_{i=1}^n (W_i - 1)\{Z_i(\widehat{f})/\tau - \widehat{g}(X_i)\} v_{x_0,i}. \tag{4.14}$$

For any $\alpha \in (0,1)$, we construct a $100*(1-\alpha)\%$ confidence interval of $g_0(x_0)$ as

$$\mathscr{I}_\alpha^\flat(x_0) := \left[\widehat{g}(x_0) - u_{\alpha/2}^\flat, \widehat{g}(x_0) - u_{1-\alpha/2}^\flat\right], \tag{4.15}$$

where $u_\alpha^\flat := u_\alpha^\flat(x_0, \mathscr{D}_n) = \inf\{u \in \mathbb{R} : \mathbb{P}^*(\mathfrak{B}^\flat(x_0) > u) \leq \alpha\}$ denotes the upper α-quantile of $\mathfrak{B}^\flat(x_0)$ under $\mathbb{P}^*$. In practice, once the vector $\mathbf{v}_{x_0}$ is obtained, we can compute the quantity $u_\alpha^\flat$ with arbitrary precision using Monte Carlo simulations; see Algorithm 2.

4.3 Statistical Theory: General Results

For notational convenience, we omit the intercept term throughout the statistical analysis. We define the *effective dimension* of the operator T_K as

$$\mathfrak{D}_\lambda := \mathrm{Tr}((T_K + \lambda I)^{-1} T_K) = \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda}$$

for any $\lambda > 0$ (Zhang, 2002; Caponnetto and De Vito, 2007). The effective dimension plays a central role in determining both the convergence rate of kernel ridge regression estimators and the dependence of regularization on the sample size. Note that T_K is a trace-class operator, i.e., the sum of all the eigenvalues of T_K is finite. To see this, Condition 4.2.1 ensures that $K(X_i, X_i) = \sum_{j=1}^{\infty} \mu_j \phi_j^2(X_i) \leq 1$. Since ϕ_j's are orthonormal, taking the expectation over X_i yields $\mathrm{Tr}(T_K) = \sum_{j=1}^{\infty} \mu_j \leq 1$. Therefore, $\mathfrak{D}_\lambda$ is well-defined for any $\lambda > 0$, satisfying $\mathfrak{D}_\lambda \leq \lambda^{-1} \sum_{j=1}^{\infty} \mu_j \leq \lambda^{-1}$. Moreover, $\mathfrak{D}_\lambda$ is increasing as λ decreases. Note that $\mathfrak{D}_\lambda \geq 1/2$ for any $\lambda \leq \mu_1$. Therefore, without loss of generality, we assume $\mathfrak{D}_\lambda \gtrsim 1$ throughout the paper. We refer the reader to Section B of the supplementary material for explicit upper bounds on $\mathfrak{D}_\lambda$ for three commonly used kernels.

4.3.1 Convergence analysis

This section presents non-asymptotic deviation bounds for the proposed estimators. We begin by providing the convergence rate of the Q-KRR estimator in (4.6). This serves as an intermediate result for the analysis of the ES-KRR estimator, for which we provide a self-contained proof. Under model (4.2), let $\varepsilon_i = Y_i - f_0(X_i)$ be the QR residuals that satisfy $\mathbb{P}(\varepsilon_i \leq 0|X_i) = \tau$ almost surely. Consistent with common practice in the QR literature, we impose some regularity conditions on the conditional distribution of ε_i given X_i in Condition 4.3.1. High probability error bounds under the $\|\cdot\|_2$ and $\|\cdot\|_{\mathscr{H}}$ norms for the Q-KRR estimator (4.6) are presented in Theorem 4.3.1.

Condition 4.3.1 (Conditional density)**.** The conditional density function of ε_i given X_i, denoted by $p_{\varepsilon_i|X_i}$ exists and is continuous on its support. Moreover, there exists absolute constants $\underline{p}, l_0 > 0$ such that $\min_{|u|\leq l_0} p_{\varepsilon_i|X_i}(u) \geq \underline{p}$ almost surely (over X_i).

Theorem 4.3.1 (Convergence rates for quantile KRR)**.** Assume Conditions 4.2.1 and 4.3.1 hold, and $f_0 = T_K^{r_q} f^*$ for some $0 \leq r_q \leq 1/2$ and $f^* \in \mathscr{H}$. For any $t > 0$, let $\lambda_q > 0$ be such that $\lambda_q \geq (\mathfrak{D}_{\lambda_q} + t)/n$ and $\lambda_q^{r_q}\|f^*\|_{\mathscr{H}} \leq 1$. Then, there exist constants $c_1, c_2 > 0$, independent of (n, λ_q, t, f_0) and $\mathscr{H}$, such that with probability at least $1 - e^{-t}$, the Q-KRR estimator $\widehat{f} = \widehat{f}_n(\lambda_q)$ satisfies $\|\widehat{f} - f_0\|_2 \leq c_1\{\lambda_q^{r_q+1/2}\|f^*\|_{\mathscr{H}} + \sqrt{(\mathfrak{D}_{\lambda_q} + t)/n}\}$ and $\|\widehat{f} - f_0\|_{\mathscr{H}} \leq c_2\{\lambda_q^{r_q}\|f^*\|_{\mathscr{H}} + \lambda_q^{-1/2}\sqrt{(\mathfrak{D}_{\lambda_q} + t)/n}\}$.

The non-asymptotic L_2-error bound in Theorem 4.3.1 consists of two components: the regularization bias $\lambda_q^{r_q+1/2}\|f^*\|_{\mathscr{H}}$ and the error term $\sqrt{(\mathfrak{D}_{\lambda_q} + t)/n}$. The bias term arises due to the use of ridge penalty and grows proportionally with λ_q when r_q is fixed. In contrast, the variance term shrinks as λ_q increases. To determine the optimal convergence rate, we need to choose a suitable value of λ_q to balance the trade-off between bias and variance. The assumption that $f_0 = T_K^{r_q} f^*$ for some $0 \leq r_q \leq 1/2$ and $f^* \in \mathscr{H}$ is referred to as the source condition (see, e.g., Chapter 3 in Engl, Hanke and Neubauer (1996)), also viewed as a smoothness assumption.

As r_q increases, the true quantile function becomes "smoother", and the bias error term decreases when $\lambda_q < 1$. Finally, the Q-KRR estimator satisfies an exponential-type concentration inequality without requiring any moment conditions on ε_i. This highlights the robustness of QR against heavy-tailed response distributions.

Next, we shift our focus to the two-step ES-KRR estimator constructed from nonparametrically generated surrogate response variables. To this end, we impose additional assumptions on the QR residuals and eigenfunctions of the RKHS.

Condition 4.3.2 (Sub-Gaussian random noise)**.** The conditional density function of ε given X is uniformly bounded from above, that is, $\sup_{u\in\mathbb{R}} p_{\varepsilon|X}(u) \leq \bar{p}$ for some constant $\bar{p} > 0$. Moreover, there exists $\sigma_0 > 0$ such that the negative part of the QR residual, $\varepsilon_- := \min(\varepsilon, 0)$, satisfies $\log \mathbb{E}_X[e^{t\{\varepsilon_- - \mathbb{E}_X(\varepsilon_-)\}}] \leq \sigma_0^2 t^2/2$ for all $t \in \mathbb{R}$ almost surely (over X), where $\mathbb{E}_X(\cdot) = \mathbb{E}(\cdot|X)$ denotes the conditional expectation given X.

Condition 4.3.3 (Uniformly bounded eigenfunctions)**.** The eigenfunctions $\{\phi_j\}_{j\geq 1}$ are uniformly bounded, that is, $\sup_{j\geq 1} \|\phi_j\|_\infty \leq C_\phi < \infty$ for some universal constant $C_\phi \geq 1$.

The sub-Gaussian assumption is common in the literature on nonparametric statistics. Under this assumption, nonparametric least squares estimators have nice properties, such as rate-optimality. Referring back to (4.8), where $Z_i(f_0) = \tau f_0(X_i) + \varepsilon_{i,-}$ with $\varepsilon_{i,-} = \min(\varepsilon_i, 0)$, we observe that it satisfies $\mathbb{E}\{Z_i(f_0)|X_i\} = \tau g_0(X_i)$ or, equivalently, $Z_i(f_0) = \tau g_0(X_i) + \varepsilon_{i,-} - \mathbb{E}_{X_i}(\varepsilon_{i,-})$. Therefore, we impose the above moment condition on ε_-. The uniform boundedness stated in Condition 4.3.3 dates back to Mendelson and Neeman (2010) and plays a crucial role in our analysis of the two-step ES estimator. It facilitates the establishment of a non-trivial error bound in the supremum norm for the Q-KRR estimator.

Our next result establishes non-asymptotic errors bound under the $\|\cdot\|_2$ and $\|\cdot\|_{\mathscr{H}}$ norms for the ES-KRR estimator (4.9), conditioning on the event that the nuisance estimator $\widehat{f} \in \mathscr{H}$ falls within a local neighborhood of f_0.

Theorem 4.3.2 (Convergence rates for expected shortfall KRR)**.** Assume that Conditions 4.2.1, 4.3.2 and 4.3.3 hold, and that $g_0 = T_K^{r_e} g^*$ for some $0 \le r_e \le 1/2$ and $g^* \in \mathscr{H}$. For any $t > 0$, let λ_e satisfy $\lambda_e \gtrsim (t + \log \mathfrak{D}_{\lambda_e})/n$ and $n \ge C_\phi^2 \mathfrak{D}_{\lambda_e} \log n$, and define the event $\mathscr{E}(\delta_2, \lambda_q) := \{\widehat{f} \in \mathscr{H} : \|\widehat{f} - f_0\|_2^2 + \lambda_q \|\widehat{f} - f_0\|_{\mathscr{H}}^2 \le \delta_2^2\}$ for $\delta_2, \lambda_q > 0$. Moreover, define

$$\gamma_{\mathrm{b}} = \tau \lambda_e^{r_e + 1/2} \|g^*\|_{\mathscr{H}}, \ \gamma_{\mathrm{s}} = \sigma_0 \sqrt{\frac{\mathfrak{D}_{\lambda_e} + t}{n}}, \ \delta_{\mathrm{s}} = C_\phi \, \delta_2 \, \mathfrak{D}_{\lambda_q}^{1/2} \sqrt{\frac{D_{\lambda_e} + t}{n}}.$$

Then, there exist constants $c_3, c_4 > 0$, independent of $(n, \lambda_q, \lambda_e, t, f_0, g_0)$ and $\mathscr{H}$, such that with probability at least $1 - 6e^{-t}$ conditioned on the event $\mathscr{E}(\delta_2, \lambda_q)$, the ES-KRR estimator $\widehat{g} = \widehat{g}_n(\lambda_e)$ satisfies

$$\tau \|\widehat{g} - g_0\|_2 \le c_3 \big(\gamma_{\mathrm{b}} + \gamma_{\mathrm{s}} + \delta_{\mathrm{s}} + \bar{p} \|\widehat{f} - f_0\|_4^2\big) \text{ and } \tau \|\widehat{g} - g_0\|_{\mathscr{H}} \le c_4 \frac{\gamma_{\mathrm{b}} + \gamma_{\mathrm{s}} + \delta_{\mathrm{s}} + \bar{p} \|\widehat{f} - f_0\|_4^2}{\sqrt{\lambda_e}}.$$

Theorem 4.3.2 establishes a non-asymptotic error bound for ES-KRR estimators using a plugged-in QR estimator $\widehat{f}$. The upper bound comprises four terms. Similar to the error bounds for Q-KRR estimators, γ_{b} captures the bias arising from the use of the ridge penalty, while γ_{s} corresponds to the variance of the estimator. When r_e is fixed, an increase in λ_e results in a larger bias and a smaller variance, highlighting the trade-off between bias and variance. Furthermore, when $\lambda_e < 1$, the bias term decreases as r_e increases.

The estimation error associated with the plugged-in estimate $\widehat{f}$ is characterized by two components: δ_{s} and $\bar{p} \|\widehat{f} - f_0\|_4^2$. The term δ_{s} arises when bounding the suprema of certain product empirical processes. Notably, due to the orthogonal property (4.5) of the score function, the squared L_4-error of the nonparametric QR estimator contributes to the L_2-error bound for the ES-KRR estimator. Therefore, even if the QR estimator converges at a slower rate under the L_4-norm, the ES estimator can still achieve the optimal convergence rate, as if the true conditional quantile function f_0 is known. While Theorem 4.3.1 quantifies the accuracy of QR estimation using L_2 and RKHS norms, an upper bound on the L_4-norm can be derived from the

inequality $\|h\|_4^4 \le \|h\|_\infty^2 \mathbb{E}_{\mathbb{P}_X}\{h^2(X)\} = \|h\|_\infty^2 \|h\|_2^2$. As a result, it is reasonable to anticipate that the L_4-error $\|\widehat{f} - f_0\|_4^2$ could be significantly smaller than $\|\widehat{f} - f_0\|_2$, depending on upper bounds for $\|\widehat{f} - f_0\|_\infty$.

4.3.2 Theoretical guarantees for pointwise inference

The two-step ES-KRR estimator, denoted as $\widehat{g}$, minimizes the penalized empirical risk $g \mapsto \widehat{\mathscr{L}_n}(\widehat{f}, g)$ with nonparametrically generated response variables. As a crucial step towards deriving the asymptotic distribution of $\widehat{g}$ via Gaussian approximation, the following theorem provides a non-asymptotic Bahadur representation of $\widehat{g}$. To assess the influence of QR estimation on the inference for the ES function, we also establish a similar Bahadur representation for the "oracle" estimator, denoted by $\widehat{g}_{\text{ora}} = \operatorname{argmin}_{g \in \mathscr{H}} \widehat{\mathscr{L}_n}(f_0, g)$. This oracle estimator is obtained by plugging in the true conditional quantile function f_0 in the empirical risk. Recall that the population (penalized) risk minimizer $g_{\lambda_e} = \operatorname{argmin}_{g \in \mathscr{H}} \mathbb{E}\{\widehat{\mathscr{L}_n}(f_0, g)\}$ defined in (4.12). Since $\mathbb{E}\{Z_i(f_0) - \tau g_0(X_i)\} = 0$, from Proposition 1 in Caponnetto and De Vito (2007) we see that g_{λ_e} is uniquely determined by $g_{\lambda_e} = (T_K + \lambda_e I)^{-1} T_K g_0$.

In addition to the QR residuals $\varepsilon_i = Y_i - f_0(X_i)$, define the zero-mean random variables $\omega_i = \varepsilon_{i,-} - \mathbb{E}_{X_i}(\varepsilon_{i,-})$ with $\varepsilon_{i,-} = \min\{\varepsilon_i, 0\}$, which can be viewed as ES residuals in the sense that $Z_i(f_0) = \tau g_0(X_i) + \omega_i$.

Theorem 4.3.3 (Functional Bahadur representations)**.** Assume that Conditions 4.2.1–4.3.3 hold, and $(f_0, g_0) = (T_K^{r_q} f^*, T_K^{r_e} g^*)$ for some $0 \le r_q, r_e \le 1/2$ and $f^*, g^* \in \mathscr{H}$. For any $t > 0$, let $n \ge 64 C_\phi^2 \mathfrak{D}_{\lambda_e}(t + \log n)$, $\lambda_q \ge (\mathfrak{D}_{\lambda_q} + t)/n$, $\lambda_q^{r_q} \|f^*\|_{\mathscr{H}} \le 1$ and $\lambda_e \gtrsim (t + \log \mathfrak{D}_{\lambda_e})/n$. Define $\delta_n := \delta_n(\lambda_q, n, t)$ and $\gamma_n := \gamma_n(\lambda_e, n, t)$ as

$$\delta_n = \lambda_q^{r_q + 1/2} \|f^*\|_{\mathscr{H}} + \sqrt{\frac{\mathfrak{D}_{\lambda_q} + t}{n}} \quad \text{and} \quad \gamma_n = \tau \lambda_e^{r_e + 1/2} \|g^*\|_{\mathscr{H}} + \sigma_0 \sqrt{\frac{\mathfrak{D}_{\lambda_e} + t}{n}}.$$

Then, with probability at least $1-8e^{-t}$, the two-step ES-KRR estimator satisfies

$$\left\| \tau(\widehat{g} - g_{\lambda_e}) - \frac{1}{n}\sum_{i=1}^{n} \omega_i (T_K + \lambda_e I)^{-1} K_{X_i} \right\|_{\infty} \leq c_5 \mathfrak{D}_{\lambda_e}^{1/2} \{\Delta_1(\lambda_e) + \Delta_2(\lambda_q, \lambda_e)\}, \tag{4.16}$$

where $\Delta_1(\lambda_e) = \mathfrak{D}_{\lambda_e}^{1/2} \gamma_n \sqrt{(t+\log n)/n}$, $\Delta_2(\lambda_q, \lambda_e) = \mathfrak{D}_{\lambda_q}^{1/2} \delta_n \{\delta_n + \sqrt{(\mathfrak{D}_{\lambda_e} + t)/n}\}$, and $c_5 = c_5(C_\phi) > 0$. Moreover, with probability at least $1-6e^{-t}$, the two-step oracle ES-KRR estimator satisfies

$$\left\| \tau(\widehat{g}_{\mathrm{ora}} - g_{\lambda_e}) - \frac{1}{n}\sum_{i=1}^{n} \omega_i (T_K + \lambda_e I)^{-1} K_{X_i} \right\|_{\infty} \leq c_6 \mathfrak{D}_{\lambda_e}^{1/2} \Delta_1(\lambda_e), \tag{4.17}$$

where $c_6 = c_6(C_\phi) > 0$.

Under Conditions 4.2.1, 4.3.2 and 4.3.3, for any $x_0 \in \mathscr{X}$, it follows from the reproducing property of K, the orthonomality and boundedness of ϕ_j that

$$\begin{aligned}
&\mathbb{E}\big[\omega_i^2 \{(T_K + \lambda_e I)^{-1} K_{X_i}(x_0)\}^2\big] \lesssim \sigma_0^2 \, \mathbb{E}\{(T_K + \lambda_e I)^{-1} K_{X_i}(x_0)\}^2 \\
&= \sigma_0^2 \, \mathbb{E}\left\{ \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda_e} \phi_j(X_i) \phi_j(x_0) \right\}^2 = \sigma_0^2 \sum_{j=1}^{\infty} \left(\frac{\mu_j}{\mu_j + \lambda_e} \right)^2 \phi_j^2(x_0) \\
&\leq C_\phi^2 \sigma_0^2 \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda_e} = C_\phi^2 \sigma_0^2 \mathfrak{D}_{\lambda_e}.
\end{aligned} \tag{4.18}$$

This indicates $(1/n)\sum_{i=1}^n \omega_i (T_K + \lambda_e I)^{-1} K_{X_i}(x_0) = \mathscr{O}_{\mathbb{P}}(\sqrt{\mathfrak{D}_{\lambda_e}/n})$. From (4.16) and (4.17) we see that $\widehat{g}$ and $\widehat{g}_{\mathrm{ora}}$ are first-order equivalent, both well approximated by $(1/n)\sum_{i=1}^n \omega_i (T_K + \lambda_e I)^{-1} K_{X_i}(x_0)$, provided that $\Delta_1(\lambda_e) + \Delta_2(\lambda_q, \lambda_e) = o(n^{-1/2})$. It is noteworthy that the functional Bahadur representation provided in Theorem 4.3.3 improves the existing results, making it of independent interest; see Section A in the supplementary material for details.

Building upon the functional Bahadur representations in Theorem 4.3.3, we establish Berry-Esseen bounds for $\widehat{g}(x_0)$ and $\widehat{g}_{\mathrm{ora}}(x_0)$ with $x_0 \in \mathscr{X}$ fixed. In particular, the following theorem demonstrates that as long as $\Delta_1(\lambda_e) + \Delta_2(\lambda_q, \lambda_e) = o(n^{-1/2})$, the two-step estimator

$\widehat{g}(x_0)$ is asymptotically equivalent to the oracle estimator $\widehat{g}_{\text{ora}}(x_0)$. The Berry-Esseen bound quantifies the accuracy of Gaussian approximations, and therefore directly implies the asymptotic normality under certain conditions.

Theorem 4.3.4 (Berry-Esseen bounds for ES-KRR estimators)**.** Assume that the same set of conditions as in Theorem 4.3.3 holds. For any $x_0 \in \mathscr{X}$, let

$$\rho^2_{\lambda_e}(x_0) := \frac{\mathbb{E}\{\omega_i(T_K+\lambda_e I)^{-1}K_{X_i}(x_0)\}^2}{\mathfrak{D}_{\lambda_e}} > 0. \tag{4.19}$$

Then, the two-step ES-KRR estimator $\widehat{g}(x_0)$ and the oracle estimator $\widehat{g}_{\text{ora}}$ satisfy

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left\{\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}\frac{\tau}{\rho_{\lambda_e}(x_0)}(\widehat{g}-g_{\lambda_e})(x_0)\le u\right\}-G(u)\right| \le c_7\sqrt{n}\big\{\Delta_1(\lambda_e)+\Delta_2(\lambda_q,\lambda_e)\big\}+8e^{-t}$$

and $\sup_{u\in\mathbb{R}}|\mathbb{P}\{\sqrt{n/\mathfrak{D}_{\lambda_e}}\tau(\widehat{g}_{\text{ora}}-g_{\lambda_e})(x_0)/\rho_{\lambda_e}(x_0)\le u\}-G(u)| \lesssim c_8\sqrt{n}\Delta_1(\lambda_e)+6e^{-t}$, respectively, where $c_7 = c_7(C_\phi,\rho_{\lambda_e}(x_0),\sigma_0) > 0$, $c_8 = c_8(C_\phi,\rho_{\lambda_e}(x_0),\sigma_0) > 0$ and $G(\cdot)$ denotes the standard normal CDF.

Corollary 4.3.1 (Pointwise asymptotic normality)**.** Assume that the same conditions as in Theorem 4.3.3 hold with $t = \log n$, and that $\rho^2_{\lambda_e}(x_0) \to \rho^2(x_0)$ for some $\rho^2(x_0) > 0$ as $n \to \infty$. Moreover, let (λ_q,λ_e,n) be such that $\Delta_1(\lambda_e)+\Delta_2(\lambda_q,\lambda_e) = o(n^{-1/2})$. Then, as $n\to\infty$, $\tau\sqrt{n/\mathfrak{D}_{\lambda_e}}(\widehat{g}-g_{\lambda_e})(x_0) \xrightarrow{\text{d}} \mathscr{N}(0,\rho^2(x_0))$, where $\xrightarrow{\text{d}}$ indicates "convergence in distribution". The same result holds for $\widehat{g}_{\text{ora}}$ when (λ_e,n) satisfy $\Delta_1(\lambda_e) = o(n^{-1/2})$.

From (4.18), we see that $\rho^2_{\lambda_e}(x_0) \le C^2_\phi\sigma^2_0$. Since λ_e is essentially a function of n, we make the high-level assumption that the variance sequence $\rho_{\lambda_e}(x_0)$ has a positive limit as $n\to\infty$. Similar conditions are often assumed in the literature (Shang and Cheng, 2013; Zhao, Liu and Shang, 2021) to obtain asymptotic distributions of KRR estimators. To ensure that the bias term $g_{\lambda_e}(x_0)-g_0(x_0)$ is negligible asymptotically, we prefer using smaller λ_e values that correspond to undersmoothing. This is a common procedure in nonparametric inference (Hall,

1992). Specifically, we choose a regularization parameter that yields a rate-suboptimal estimator but with an asymptotically centered normal distribution.

Remark 4.3.1 (Confidence interval construction based on the asymptotic normality)**.** Based on the asymptotic results in Corollary 4.3.1, we may consider the (approximate) $100*(1-\alpha)\%$ confidence interval $\widehat{g}(x_0) \pm \tau^{-1} z_{\alpha/2} \cdot \rho_{\lambda_e}(x_0) \sqrt{\mathfrak{D}_{\lambda_e}/n}$, where $z_{\alpha/2}$ is the upper $(\alpha/2)$-percentile of the standard normal distribution. The variance term $\rho^2_{\lambda_e}(x_0)$, or equivalently, $\mathbb{E}[\omega_i^2\{(T_K + \lambda_e I)^{-1} K_{X_i}(x_0)\}^2]$, still needs to be estimated in practice. Assuming $\mathrm{Var}(\varepsilon_{i,-}|X_i) = \sigma^2$ for some constant $\sigma^2 > 0$, similar to (4.18) it can be calculated that $\rho^2_{\lambda_e}(x_0) = \frac{\sigma^2}{\mathfrak{D}_{\lambda_e}} \sum_{j=1}^{\infty} \frac{\mu_j^2}{(\mu_j+\lambda_e)^2} \phi_j^2(x_0)$. In some cases, such as when the RKHS is the periodic Sobolev space, an explicit formulation of $\lim_{\lambda_e \to 0} \mathfrak{D}_{\lambda_e}^{-1} \sum_{j=1}^{\infty} \mu_j^2 \phi_j^2(x_0)/(\mu_j + \lambda_e)^2$ can be calculated (see, e.g., Lemma 6.1 in Shang and Cheng (2013)). It then suffices to estimate σ^2. However, the above homoscedasticity condition is fairly restrictive and neglects the heterogeneity in X at different quantile levels of the response distribution. Without this condition, consistently estimating $\rho^2_{\lambda_e}(x_0)$ becomes a challenging task.

To examine the validity of bootstrap-based confidence interval constructions described in Section 4.2.3, we consider the bootstrap statistic $\mathfrak{B}^\flat(x_0)$ given in (4.14), with the random weights W_i satisfying the following condition.

Condition 4.3.4 (Sub-Gaussian bootstrap weights)**.** The random weights $\{W_i\}_{i=1}^n$ are independent copies of some random variable W satisfying $\mathbb{E}(W) = 1$ and $\mathrm{Var}(W) = 1$. Moreover, there exists a constant $\sigma_W > 0$ such that $\log \mathbb{E}\{e^{t(W-1)}\} \le \sigma_W^2 t^2/2$ for all $t \in \mathbb{R}$.

Condition 4.3.4 is satisfied by commonly used Rademacher weights and Gaussian weights. Recall that $\mathbb{P}^*(\cdot) = \mathbb{P}(\cdot|\mathscr{D}_n)$ denotes the conditional probability given $\mathscr{D}_n = \{(Y_i, X_i)\}_{i=1}^n$. The next result shows that the conditional distribution of $\mathfrak{B}^\flat(x_0)$ approximates the distribution of $\widehat{g}(x_0) - g_0(x_0)$ well under suitable conditions.

Theorem 4.3.5 (Bootstrap approximation error)**.** Assume that Conditions 4.2.1–4.3.4 hold, $(f_0, g_0) = (T_K^{r_q} f^*, T_K^{r_e} g^*)$ for some $0 \le r_q, r_e \le 1/2$ and $f^*, g^* \in \mathscr{H}$. For any $t > 0$, let $n \ge$

$64C_\phi^2 \mathfrak{D}_{\lambda_e}(t+\log n)\log n$, $\lambda_q \geq (\mathfrak{D}_{\lambda_q}+t)/n$, $\lambda_q^{r_q}\|f^*\|_{\mathscr{H}} \leq 1$ and $\lambda_e \gtrsim (t+\log \mathfrak{D}_{\lambda_e})/n$. Then, with probability (over the independent sample $\mathscr{D}_n$) at least $1-16e^{-t}$,

$$\begin{aligned}&\sup_{u\in\mathbb{R}}\left|\mathbb{P}\{(\widehat{g}-g_0)(x_0)\leq u\}-\mathbb{P}^*\{\mathfrak{B}^\flat(x_0)\leq u\}\right| \\ &\quad\leq c_9\sqrt{n}\big\{\Delta_1(\lambda_e)+\Delta_2(\lambda_q,\lambda_e)+\tau\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}\big\}+13e^{-t}, \end{aligned} \tag{4.20}$$

where $c_9 = c_9(C_\phi, \sigma_0, \sigma_W, \rho_{\lambda_e}(x_0)) > 0$ and $\rho_{\lambda_e}(x_0)$ is as in (4.19).

For any $\alpha \in (0,1)$, recall the definition of $u_\alpha^\flat$ in (4.14). As a direct consequence of Theorem 4.3.5, the following result provides a theoretical guarantee for the confidence interval $\mathscr{I}_\alpha^\flat(x_0)$ constructed using the multiplier bootstrap, as defined in (4.15).

Theorem 4.3.6 (Validity of bootstrap approximation)**.** Under the same set of conditions as in Theorem 4.3.5, there exists a constant $c_{10} = c_{10}(C_\phi, \sigma_0, \sigma_W, \rho_{\lambda_e}(x_0))$ such that, for any $\alpha \in (0,1)$, $|\mathbb{P}\{g_0(x_0) \in \mathscr{I}_\alpha^\flat(x_0)\} - (1-\alpha)| \leq \mathrm{Err}^\flat(n,t)$, where $\mathrm{Err}^\flat(n,t) = c_{10}\sqrt{n}\{\Delta_1(\lambda_e)+\Delta_2(\lambda_q,\lambda_e)+\tau\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}\}+62e^{-t}$.

Since the convergence rates and conditions ensuring asymptotic normality vary considerably among different kernel types, we defer instantiations of the general bounds on estimation error and Gaussian approximation error for RKHSs to Section B in the supplementary material.

4.4 Numerical and Empirical Studies

The two-step nature of the proposed method offers advantages not only in terms of desirable statistical properties, achieved through the use of an orthogonal score, but also in facilitating practical implementations. The computation of the Q-KRR estimator involves reformulating (4.7) into a quadratic program (Takeuchi et al., 2006):

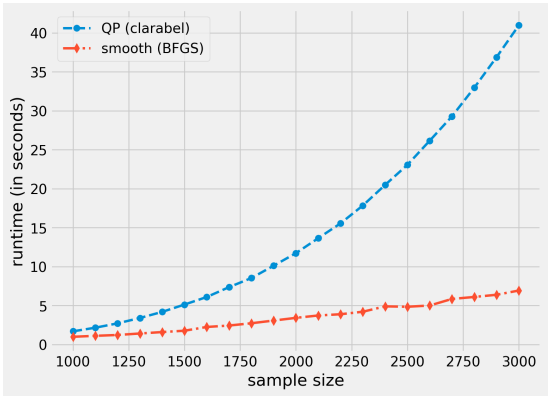
$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha}-\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{y},\ \text{ subject to }\ C_q(\tau-1)\leq\alpha_i\leq C_q\tau, 1\leq i\leq n, \sum_{i=1}^n\alpha_i=0, \tag{4.21}$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $C_q = 1/(2\lambda_q n)$. Let $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)^{\mathrm{T}}$ be the optimal solution from (4.21) and let $\widehat{b}$ be the τ-th sample quantile of $\mathbf{y} - \mathbf{K}\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^n$. The Q-KRR estimator $\widehat{f}$ is then computed as $\widehat{f}(x) = \widehat{b} + \sum_{i=1}^n \widehat{\alpha}_i K(X_i, x)$. The quadratic program (4.21) can be efficiently solved using off-the-shelf solvers. For relatively small to moderate sample sizes, typically ranging from 500 to 2000, our `Python` implementation employs `Clarabel` (https://oxfordcontrol.github.io/ClarabelDocs), an interior point numerical solver for convex optimization problems that using a novel homogeneous embedding. In the second step, the estimation of g_0 is performed through the least squares KRR, using the generated surrogate response variables.

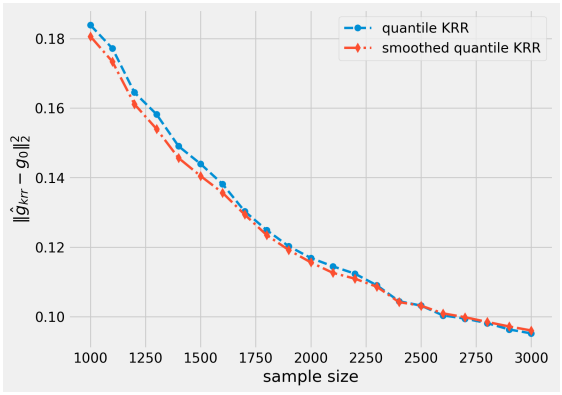
The primary computational effort of our proposed two-step procedure arises from the kernel ridge regression with the check/quantile loss. To enhance the efficiency of handling large-scale datasets, we propose a different approach combining convolution smoothing and quasi-Newton methods, as recently advocated in He et al. (2023). Given a smoothing parameter/bandwidth $h > 0$ and a nonnegative, symmetric kernel function $H(\cdot)$ that integrates to 1, the convolution-smoothed check loss is defined as $\rho_{\tau,h}(u) = \rho_\tau \circ H_h = (1/h)\int_{-\infty}^{\infty} \rho(v) H((v-u)/h)\mathrm{d}v$, where $H_h(u) = H(u/h)/h$. To solve the smoothed version of problem (4.7), in which the check loss ρ_τ is replaced by $\rho_{\tau,h}$, we use the `L-BFGS-B` method, a limited-memory version of the BFGS algorithm, in the `minimize` function from the `scipy.optimize` module. Figure 4.1 presents a runtime comparison for computing the Q-KRR estimator using `Clarabel` and its smoothed variant employing `L-BFGS-B`. In our implementation, the default smoothing parameter h is set to $\max\{10^{-4}, \widehat{\sigma} n^{-1/3}\}$, where $\widehat{\sigma}$ denotes the sample standard deviation of the fitted KRR residuals $\{Y_i - \widehat{m}_{\mathrm{krr}}(X_i)\}_{i=1}^n$.

4.4.1 Synthetic data experiments

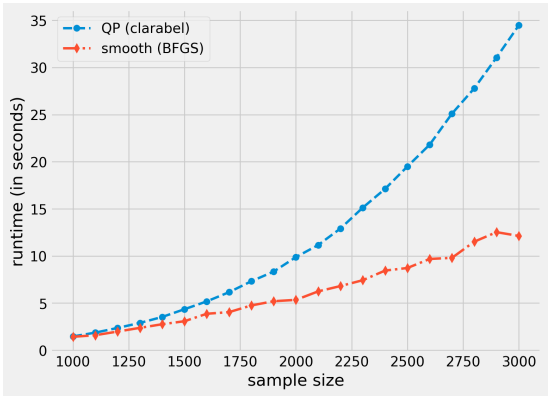
We conduct numerical studies to assess the out-of-sample performance of the two-step expected shortfall KRR estimator and the finite-sample performance evaluation of the proposed bootstrap inference procedure. We generate the data $\{(Y_i, X_i)\}_{i=1}^n$ from the location-scale model



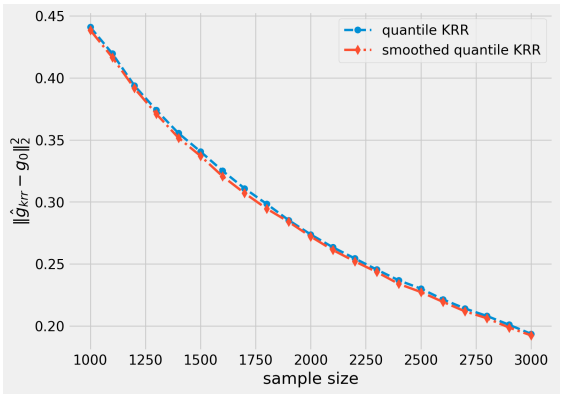
(a) Runtime comparison under Model (4.22)



(b) Error comparison under Model (4.22)



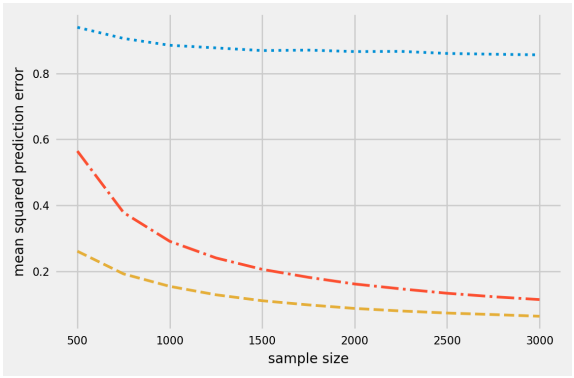
(c) Runtime comparison under Model (4.23)



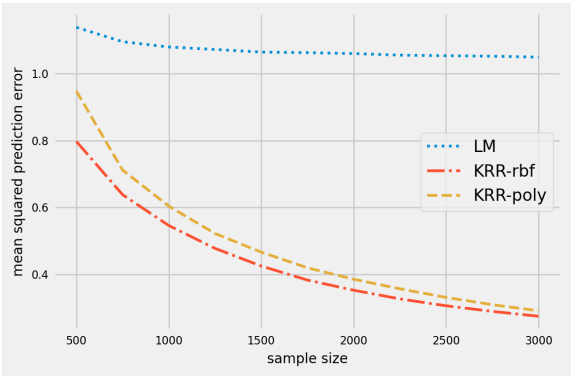
(d) Error comparison under Model (4.23)

Figure 4.1. Numerical comparison between Q-KRR and smoothed Q-KRR when $\tau = 0.1$ and $\lambda_q = 1/(2n)$. The former relies on a QP reformulation solved by the `Clarabel` solver, while the latter uses the `BFGS` algorithm. Data are generated from Models (4.22) and (4.23) with n ranging from 1000 to 3000. Left panels: average running time (in seconds) versus sample size. Right panels: mean squared error (in-sample) versus sample size.

$Y_i = m(X_i) + s(X_i)\eta_i$, where $X_i \in \mathbb{R}^d$ follows the uniform distribution on the unit cube $[0,1]^d$ and $\eta_i \sim \mathcal{N}(0,1)$. Here, $m : [0,1]^d \to \mathbb{R}$ is the conditional mean function and $s : [0,1]^d \to (0,\infty)$ corresponds to the heterogenous noise scale. For a given quantile level $\tau \in (0,1)$, the associated conditional quantile and ES functions of Y_i given $X_i = x$ are $f_0(x) = m(x) + s(x)q_{\mathcal{N}(0,1)}(\tau)$ and $g_0(x) = m(x) + s(x)e_{\mathcal{N}(0,1)}(\tau)$, respectively. Throughout our numerical studies, we consider two



(a) MSPE vs sample size under Model (4.22)



(b) MSPE vs sample size under Model (4.23)

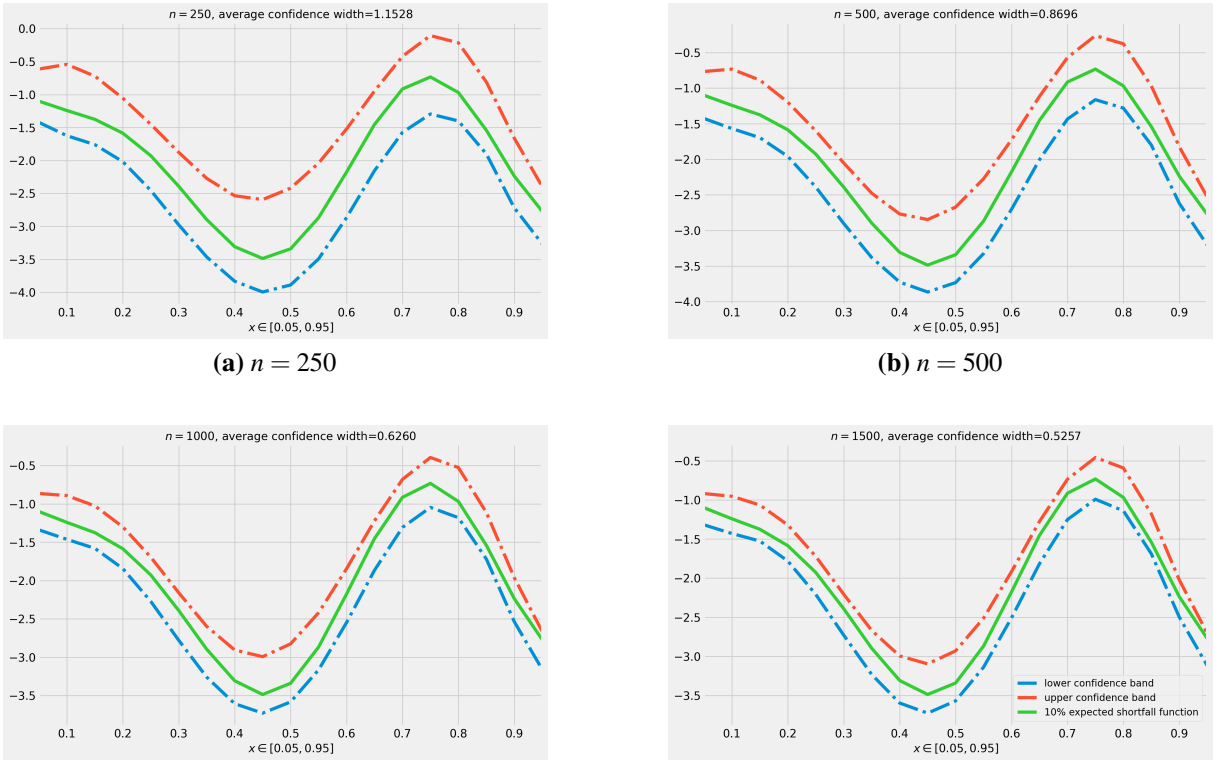
Figure 4.2. The mean squared prediction error with n ranging from 500 to 3000, averaged over 200 replications, for two 10%-level ES-KRR estimators using RBF and polynomial kernels, and the linear ES estimator under Models (4.22) and (4.23).

different nonparametric models with dimensions $d = 5$ and $d = 8$, respectively:

$$d = 5 \quad \begin{cases} m(x) = -(x_1 + x_2 + x_3)^3 - \tanh(x_1 + x_3 + x_5), \\ s(x) = 1 + (x_4 - 0.5)^2, \end{cases} \tag{4.22}$$

$$d = 8 \quad \begin{cases} m(x) = \cos(2\pi x_1) + \frac{1}{1+e^{-x_2-x_3}} + \frac{1}{(1+x_4+x_5)^3} + \frac{1}{x_6+e^{x_7 x_8}}, \\ s(x) = \sin(\pi(x_1+x_2)/2) + \log(1 + x_3^2 x_4^2 x_5^2) + \frac{x_8}{1+e^{-x_6-x_7}}. \end{cases} \tag{4.23}$$

We compute the two-step ES-KRR estimator using the radial basis function (RBF) kernel $K(x,x') = \exp(-\|x-x'\|_2^2)$ (`KRR-rbf`) and the polynomial kernel $K(x,x') = (\langle x,x'\rangle + 1)^3$ (`KRR-poly`), $x, x' \in \mathbb{R}^d$. Both kernels are employed with regularization parameters $\lambda_e = 2\lambda_q = 1/n$. For demonstrative purposes only, we also implement the two-step linear ES regression estimator (`LM`) proposed in He et al. (2023). The out-of-sample performance is assessed using the mean squared prediction error (MSPE) on a test set $\{(Y_i^{\text{test}}, X_i^{\text{test}})\}_{i=1}^{n_{\text{test}}}$ with a size of $n_{\text{test}} = 10000$, that is, $n_{\text{test}}^{-1} \sum_{i=1}^{n_{\text{test}}} \{g_0(X_i^{\text{test}}) - \widehat{g}(X_i^{\text{test}})\}^2$. In Figure 4.2, we report the MSPE of different methods under Models (4.22) and (4.23) at the quantile level $\tau = 0.1$, averaged over 200 replications, for $n \in \{500, 750, 1000, \ldots, 3000\}$. The MSPEs of KRR estimators under different kernels decrease



(a) $n = 250$ **(b)** $n = 500$

(c) $n = 1000$ **(d)** $n = 1500$

Figure 4.3. 95% pointwise bootstrap confidence bands for the true 10%-level ES regression function g_0 at $x_0 \in \{0.05, \ldots, 0.95\}$ with $n \in \{250, 500, 1000, 1500\}$. Normal weights $W_i \sim \mathcal{N}(1,1)$ are used and the number of bootstrap samples is fixed at $B = 1000$.

as the sample size increases, while that of the LM method remains at a constant level due to model misspecification. Compared to local polynomial-type ES estimators, which are mostly applicable for uni- and bi-variate cases, the two-step kernel ridge regression method showcases notable efficiency and accuracy in moderate-dimensional settings.

Next, we examine the effectiveness of the proposed multiplier bootstrap for constructing pointwise confidence bands. We consider a univariate heteroscedastic model $Y = 2X \cdot \sin(3.5\pi X) + \{0.5 + |\sin(\pi X)|\}\eta$, where $X \sim \text{Unif}(0,1)$ and $\eta \sim \mathcal{N}(0,1)$. Fixing $\tau = 0.1$, we generate training samples of size $n \in \{250, 500, 1000, 1500\}$ and construct pointwise 95% confidence intervals for $g_0(x_0)$, $x_0 \in \{0.05, 0.1, \ldots, 0.9, 0.95\}$ using Algorithm 2 with $W_i \sim \mathcal{N}(1,1)$, $B = 1000$ and $\alpha = 0.05$. Both Q-KRR and ES-KRR estimators use the polynomial kernel

$K(x,x') = (1+xx')^9$ and the ridge penalties are set to $\lambda_e = 2\lambda_q = 10^{-4}/n$. The upper and lower confidence bands, averaged over 200 replications, are shown in Figure 4.3. Compared to the computation of Q-KRR and ES-KRR estimators, the additional cost of bootstrapping is negligible. Nevertheless, this rapid bootstrap method ensures good coverage, and the confidence interval narrows with an increasing sample size n.

4.4.2 An Application to Medical Expense Data

We demonstrate the applicability of nonparametric ES regression on a simulated dataset containing medical expenses for patients in the United States. This data was created by Lantz (2013) using demographic statistics from the U.S. Census Bureau and thus approximately reflects real-world conditions. In total, the dataset contains 1,338 beneficiaries enrolled in some insurance plan, with features indicating characteristics of the patient, as well as the total medical expenses charged to the plan for the calendar year. Let Y_i be the insurance charges (in \$1000) that range from 1.12 to 63.77. Figure 4.4 shows the kernel density estimate of the insurance charges as well as two vertical lines indicating the sample mean and sample 90% quantile. The empirical density appears to be bimodal, with one mode occurring in the upper tail. This makes the prediction of the upper tail average particularly relevant. The available features include `age`, `gender` (male or female), `BMI` (body mass index), `children` (the number of children/dependents covered by the plan), `smoker` (yes or no, depending on whether the insured regularly smokes tobacco), and `region` (place of residence, divided into northeast, southeast, southwest, and northwest).

For our analysis, we partition the data into a training set of size 1003 (75% of the total) and a validation set of size 335. We first fit standard KRR and 90%-quantile KRR on the training set, with regularization parameters selected to minimize the mean squared error and the check loss on the validation set, respectively. By plugging in fitted 90% quantiles $\widehat{f}(X_i)$, we apply the two-step approach to fit an upper 10%-level ES-KRR. Both methods employ a polynomial kernel with a degree of 5.

To compare KRR for predicting average charges and ES-KRR for predicting the average

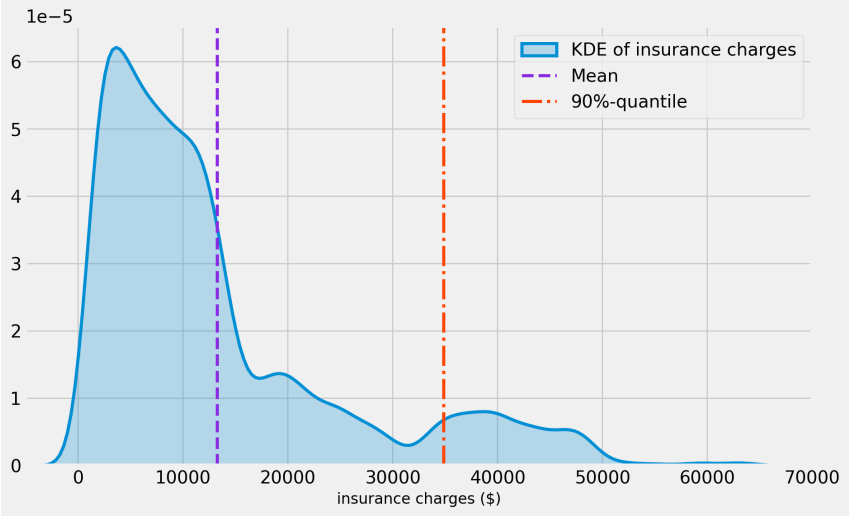


Figure 4.4. Kernel density estimate of the insurance charges (in $).

of high chargers (those exceeding the 90% quantile), we use the variable permutation importance (VPI) to measure the importance of individual feature components (Breiman, 2001). The VPI is calculated by first randomly permuting one feature component at a time across the entire training sample and then measuring the relative increase in loss obtained using the features with the permuted component. For KRR, the mean squared error loss is used, and for ES-KRR, the mean squared error loss is defined using the surrogate response variables given in (4.8), with f replaced by $\widehat{f}$. Figure 4.5 displays the VPIs for the mean (left) and the upper 10% ES (right). The bars display relative increases in losses, ordered by their magnitudes. The feature `smoker` is most significant in both cases and is the overwhelmingly dominating factor for predicting average charges. The ordering of the remaining features changes drastically. With KRR regression, `bmi` and `age` arise after `smoker`, whose VPIs are much higher than the rest. With upper 10% ES-KRR, `region` and `age` become the second tier, showing the importance of age and spatial effect on higher insurance charges. Comparably, `bmi` and `children` are less important but are not completely negligible. These results exactly reflect the potentially different effects of the patients' characteristics in average and large charges, bringing new insights into insurance pricing.

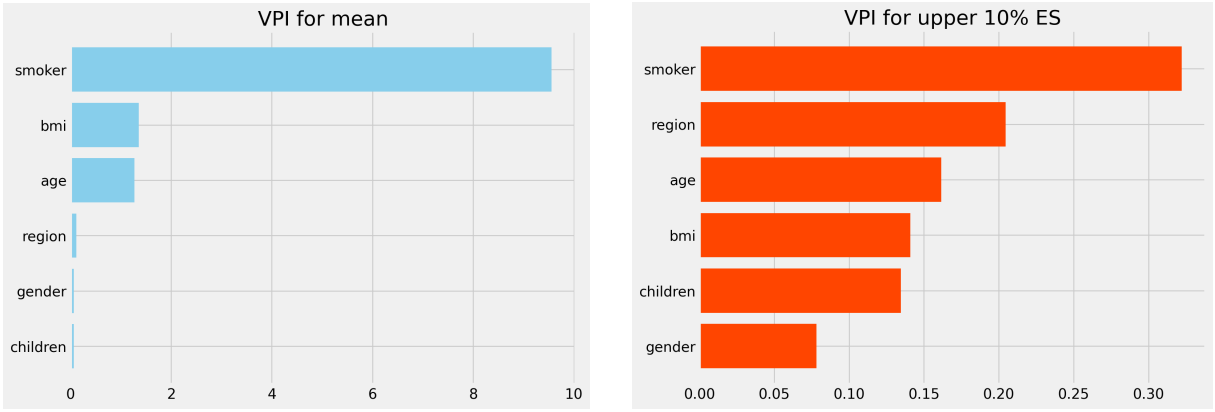


Figure 4.5. Variable permutation importance for mean regression with KRR (left), and 10% upper ES regression using the proposed two-step method (right).

4.5 Conclusion and Discussions

In this paper, we propose a two-step nonparametric method for estimating expected shortfall regression and introduce efficient bootstrap procedures to construct pointwise confidence intervals. We establish a finite-sample theoretical framework for this two-step method, including high probability bounds, functional Bahadur representation, pointwise Gaussian approximations, and bootstrap validity. Numerical and empirical studies further confirm the efficacy and usefulness of nonparametric ES regression in RKHS.

While our main focus is on pointwise inference, there is a natural interest in developing a uniform confidence band for the conditional ES function with theoretical guarantees. That is, for any $\alpha \in (0,1)$, construct a confidence band $\{\mathscr{C}_n(x) = [\widehat{g}^L(x), \widehat{g}^U(x)] : x \in \mathscr{X}\}$ from $\{(Y_i, X_i)\}_{i=1}^n$, satisfying that $\mathbb{P}\{g_0(x) \in \mathscr{C}_n(x) \text{ for all } x \in \mathscr{X}\} \to 1-\alpha$ as $n \to \infty$. Drawing on ideas from Singh and Vijaykumar (2023), we consider the maximum bootstrap statistic $\mathfrak{M}^\flat := \sup_{x\in\mathscr{X}} |\mathfrak{B}^\flat(x)|$ with $\mathfrak{B}^\flat(x)$ given in (4.14). We then construct an approximate $100*(1-\alpha)\%$ uniform confidence band of g_0 as $\{\mathscr{C}_n(x) = [\widehat{g}(x) - t_\alpha^\flat, \widehat{g}(x) + t_\alpha^\flat] : x \in \mathscr{X}\}$, where $t_\alpha^\flat$ denotes the upper α-quantile of $\mathfrak{M}^\flat$ under $\mathbb{P}^*$. While the non-asymptotic theory developed in this work and a collection of technical results provide a foundation for validating this method, a comprehensive theoretical examination, including the derivation of Gaussian approximation inequalities for empirical

processes, requires significant additional effort. This task is deferred to future research.

4.6 Acknowledgements

This chapter, in full, is currently being prepared for submission for publication of the material. Yu, Myeonghun; Wang, Yue; Xie, Siyu; Tan, Kean Ming; Zhou, Wen-Xin. The dissertation author was the primary investigator and author of this material.

Appendix A

Supplementary Material for Chapter 1

A.1 Proofs of matrix sensing

A.1.1 Proof of Theorem 1.3.1

For simplicity, we write $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}_{\tau,\lambda}$ and $\mathscr{E} = \mathscr{E}(s,l,\kappa)$ with the parameters (s,l,κ) satisfying (1.12). We prove the theorem by contradiction. Assume that $\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} > C\sqrt{\rho}\,(\lambda/\kappa)^{1-q/2}$ for some $C > 0$ to be determined. Then there exists some $\eta \in (0,1)$ such that $\widetilde{\mathbf{\Theta}}_\eta = \mathbf{\Theta}^* + \eta(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)$ satisfies $\|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}^*\|_{\mathrm{F}} = C\sqrt{\rho}\,(\lambda/\kappa)^{1-q/2}$. Applying Lemma F.2 in Fan *et al.* (2018) to the loss function $\widehat{L}_\tau(\cdot)$, we have

$$\langle \nabla \widehat{L}_\tau(\widetilde{\mathbf{\Theta}}_\eta) - \nabla \widehat{L}_\tau(\mathbf{\Theta}^*), \widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}^* \rangle \le \eta \langle \nabla \widehat{L}_\tau(\widehat{\mathbf{\Theta}}) - \nabla \widehat{L}_\tau(\mathbf{\Theta}^*), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \rangle. \tag{A.1}$$

To bound the right-hand side of (A.1), the first-order necessary condition for the convex optimization problem (1.7) implies that

$$\langle \nabla \widehat{L}_\tau(\widehat{\mathbf{\Theta}}) + \lambda \widehat{\boldsymbol{Z}}, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \rangle \le 0,$$

where $\widehat{\boldsymbol{Z}} \in \partial \|\widehat{\mathbf{\Theta}}\|_*$ satisfies $\langle \widehat{\boldsymbol{Z}}, \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}} \rangle \le \|\mathbf{\Theta}^*\|_* - \|\widehat{\mathbf{\Theta}}\|_*$. Whenever $\lambda \ge 2\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2$, it

follows that

$$\langle \nabla \widehat{L}_\tau(\widehat{\boldsymbol{\Theta}}) - \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*), \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^* \rangle \leq \lambda(\|\boldsymbol{\Theta}^*\|_* - \|\widehat{\boldsymbol{\Theta}}\|_*) + \frac{\lambda}{2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_* \\ \leq \frac{3\lambda}{2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_*, \tag{A.2}$$

where the last step uses the triangular inequality. To bound $\|\widehat{\boldsymbol{\Delta}}\|_* = \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_*$, we follow the proof of Corollary 2 in Negahban and Wainwright (2011). Let $\boldsymbol{\Theta}^* = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ be an SVD of $\boldsymbol{\Theta}^*$, where $\boldsymbol{U} \in \mathbb{R}^{d_1 \times d_2}, \boldsymbol{V} \in \mathbb{R}^{d_2 \times d_2}$ and the diagonals of $\boldsymbol{\Sigma} \in \mathbb{R}^{d_2 \times d_2}$ are in descending order. For some integer $r \leq d_2$ to be specified, define $\boldsymbol{U}^r \in \mathbb{R}^{d_1 \times r}$ and $\boldsymbol{V}^r \in \mathbb{R}^{d_2 \times r}$, whose columns correspond to the first r columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Moreover, define

$$\mathbb{M} = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \mathrm{row}(\boldsymbol{\Theta}) \subset \mathrm{col}(\boldsymbol{V}^r), \mathrm{col}(\boldsymbol{\Theta}) \subset \mathrm{col}(\boldsymbol{U}^r)\},$$
$$\overline{\mathbb{M}}^{\perp} = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \mathrm{row}(\boldsymbol{\Theta}) \perp \mathrm{col}(\boldsymbol{V}^r), \mathrm{col}(\boldsymbol{\Theta}) \perp \mathrm{col}(\boldsymbol{U}^r)\},$$

where $\mathrm{col}(\cdot)$ and $\mathrm{row}(\cdot)$ denote the column space and row space, respectively. For any $\boldsymbol{\Delta} \in \mathbb{R}^{d_1 \times d_2}$ and a closed subspace $\mathscr{W}$ of $\mathbb{R}^{d_1 \times d_2}$, let $\boldsymbol{\Delta}_{\mathscr{W}}$ be the projection of $\boldsymbol{\Delta}$ onto $\mathscr{W}$. Applying Lemma 1 in Negahban et al. (2012) to $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$, we obtain

$$\|\widehat{\boldsymbol{\Delta}}_{\overline{\mathbb{M}}^{\perp}}\|_* \leq 3\|\widehat{\boldsymbol{\Delta}}_{\overline{\mathbb{M}}}\|_* + 4\sum_{j \geq r+1} \sigma_j(\boldsymbol{\Theta}^*).$$

Since $\mathrm{rank}(\widehat{\boldsymbol{\Delta}}_{\overline{\mathbb{M}}}) \leq 2r$, the above inequality implies

$$\|\widehat{\boldsymbol{\Delta}}\|_* \leq \|\widehat{\boldsymbol{\Delta}}_{\overline{\mathbb{M}}}\|_* + \|\widehat{\boldsymbol{\Delta}}_{\overline{\mathbb{M}}^{\perp}}\|_* \leq 4\|\widehat{\boldsymbol{\Delta}}_{\overline{\mathbb{M}}}\|_* + 4\sum_{j \geq r+1} \sigma_j(\boldsymbol{\Theta}^*) \leq 4\sqrt{2r}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}} + 4\sum_{j \geq r+1} \sigma_j(\boldsymbol{\Theta}^*). \tag{A.3}$$

Set a threshold $t = \lambda/\kappa$, we choose

$$r = |\{j \in \{1, 2, \ldots, d_2\} : \sigma_j(\boldsymbol{\Theta}^*) \geq t\}|.$$

Since $0 \leq q \leq 1$, we have

$$\sum_{j \geq r+1} \sigma_j(\mathbf{\Theta}^*) = t \sum_{j \geq r+1} \frac{\sigma_j(\mathbf{\Theta}^*)}{t} \leq t \sum_{j \geq r+1} \left(\frac{\sigma_j(\mathbf{\Theta}^*)}{t} \right)^q = t^{1-q} \sum_{j \geq r+1} \sigma_j(\mathbf{\Theta}^*)^q \leq t^{1-q} \rho.$$

By the definition of r, we also have $\rho \geq \sum_{j \leq r} \sigma_j(\mathbf{\Theta}^*)^q \geq r t^q$, and hence $r \leq \rho t^{-q}$. Combining these two bounds with (A.3) and the assumption on $\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}$ (i.e. $\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} > C\sqrt{\rho}\,(\lambda/\kappa)^{1-q/2}$) yields

$$\|\widehat{\mathbf{\Delta}}\|_* \leq 4\sqrt{2\rho t^{-q}} \|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} + 4t^{1-q}\rho \tag{A.4}$$

$$\leq (4\sqrt{2} + 4/C)\sqrt{\rho} \left(\frac{\kappa}{\lambda} \right)^{q/2} \|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}. \tag{A.5}$$

Consequently, $\widetilde{\mathbf{\Theta}}_\eta \in \mathbf{\Theta}^* + \mathbb{B}(s) \cap \mathbb{C}(l)$ provided $l \geq (4\sqrt{2} + 4/C)\sqrt{\rho}\,(\kappa/\lambda)^{q/2}$. Conditioning on $\mathscr{E}$, it follows from (A.1), (A.2) and (A.5) that

$$\kappa \|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}^*\|_{\mathrm{F}}^2 \leq \frac{3\lambda}{2} \cdot (4\sqrt{2} + 4/C)\sqrt{\rho} \left(\frac{\kappa}{\lambda} \right)^{q/2} \cdot \eta \|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}.$$

Since $\eta \|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} = \|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}^*\|_{\mathrm{F}} = C\sqrt{\rho}\,(\lambda/\kappa)^{1-q/2}$ by construction, canceling out $\|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}^*\|_{\mathrm{F}}$ from both sides yields

$$C\sqrt{\rho}\,(\lambda/\kappa)^{1-q/2} = \|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}^*\|_{\mathrm{F}} \leq (6\sqrt{2} + 6/C)\sqrt{\rho}\,(\lambda/\kappa)^{1-q/2}.$$

Based on the above analysis, we choose $C = 9.15$ so that $6\sqrt{2} + 6/C \approx 9.14 < C$, which is a contradiction. We thus conclude that $\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} \leq 9.15\sqrt{\rho}\,(\lambda/\kappa)^{1-q/2}$ conditioned on $\mathscr{E}$.

Combining this Frobenius norm error rate with (A.4) proves the claimed error bound under nuclear norm. The proof is complete . □

A.1.2 Proof of Proposition 1.3.1

To begin with, define the zero-mean random matrix $\mathbf{\Gamma} = \nabla \widehat{L}_\tau(\mathbf{\Theta}^*) - \mathbb{E}\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)$ so that

$$\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2 \le \|\mathbf{\Gamma}\|_2 + \|\mathbb{E}\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2. \tag{A.6}$$

First, we bound $\|\mathbf{\Gamma}\|_2$ via a standard covering argument: there exist a $(1/4)$-net $\mathscr{N}_1$ of $\mathbb{S}^{d_1-1}$ and a $(1/4)$-net $\mathscr{N}_2$ of $\mathbb{S}^{d_2-1}$ with $|\mathscr{N}_1| \le 9^{d_1}$ and $|\mathscr{N}_2| \le 9^{d_2}$ such that

$$\|\mathbf{\Gamma}\|_2 \le 2 \max_{\boldsymbol{u}\in\mathscr{N}_1} \max_{\boldsymbol{v}\in\mathscr{N}_2} \boldsymbol{u}^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{v} = 2 \max_{\boldsymbol{u}\in\mathscr{N}_1} \max_{\boldsymbol{v}\in\mathscr{N}_2} \frac{1}{n}\sum_{i=1}^{n} \{\xi_i \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v} - \mathbb{E}(\xi_i \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v})\}, \tag{A.7}$$

where $\xi_i = \psi_\tau(\varepsilon_i)$. Since $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v} = \mathrm{vec}(\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}})^{\mathrm{T}}\mathrm{vec}(\boldsymbol{X}_i)$ is sub-exponential and $\|\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\|_{\mathrm{F}} = 1$, for $k = 2,3,\ldots$ we have

$$\mathbb{E}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}|^k = \nu_0^k \cdot k\int_0^\infty u^{k-1}\mathbb{P}(|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}| \ge \nu_0 u)\mathrm{d}u \le 2\nu_0^k \cdot k\int_0^\infty u^{k-1}e^{-u}\mathrm{d}u = 2k!\nu_0^k. \tag{A.8}$$

It follows that

$$\mathbb{E}(\xi_i \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v})^2 \le \sigma_0^2 \cdot 4\nu_0^2, \quad \text{and} \quad \mathbb{E}(|\xi_i \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}|^k) \le 2\tau^{k-2}\sigma_0^2 \cdot k!\nu_0^k = \frac{k!}{2}(4\sigma_0^2\nu_0^2)(\tau\nu_0)^{k-2}.$$

Applying Bernstein's inequality, we see that for every $x > 0$,

$$\mathbb{P}\left(\boldsymbol{u}^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{v} \ge 2\sqrt{2}\nu_0\sigma_0\sqrt{\frac{x}{n}} + \nu_0\tau\frac{x}{n}\right) \le e^{-x}.$$

Taking the union bound over all vectors $\boldsymbol{u} \in \mathscr{N}_1$ and $\boldsymbol{v} \in \mathscr{N}_2$, and setting $x = 3d + z \ge \log(9^{d_1}) + \log(9^{d_2}) + z$ with $d := d_1 + d_2$, it follows from (A.7) that

$$\mathbb{P}\left(\|\mathbf{\Gamma}\|_2 \ge 4\sqrt{2}\nu_0\sigma_0\sqrt{\frac{3d+z}{n}} + 2\nu_0\tau\frac{3d+z}{n}\right) \le e^{-z}. \tag{A.9}$$

For the second term $\|\mathbb{E}\nabla\widehat{L}_\tau(\mathbf{\Theta}^*)\|_2$ in (A.6), note that

$$\|\mathbb{E}\nabla\widehat{L}_\tau(\mathbf{\Theta}^*)\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^{d_1-1},\boldsymbol{v}\in\mathbb{S}^{d_2-1}} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\xi_i \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}).$$

Recall that $\mathbb{E}(\varepsilon_i|\boldsymbol{X}_i)=0$ and $|\xi_i-\varepsilon_i| = |\ell'_\tau(\varepsilon_i)-\varepsilon_i| \le \varepsilon_i^2/\tau$. Therefore, for each $\boldsymbol{u}\in\mathbb{S}^{d_1-1}$ and $\boldsymbol{v}\in\mathbb{S}^{d_2-1}$, applying (A.8) with $k=2$ gives

$$\mathbb{E}(\xi_i\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}) \le \frac{1}{\tau}\mathbb{E}(\varepsilon_i^2|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}|) \le 2\nu_0\frac{\sigma_0^2}{\tau}.$$

Combining this with (A.6) and (A.9), we conclude that with probability at least $1-e^{-z}$,

$$\|\nabla\widehat{L}_\tau(\mathbf{\Theta}^*)\|_2 \le 4\sqrt{2}\nu_0\sigma_0\sqrt{\frac{3d+z}{n}} + 2\nu_0\tau\frac{3d+z}{n} + 2\nu_0\frac{\sigma_0^2}{\tau}.$$

Therefore, for any $\sigma\ge\sigma_0$, taking $\tau=\sigma\sqrt{n/(3d+z)}$ yields

$$\|\nabla\widehat{L}_\tau(\mathbf{\Theta}^*)\|_2 \le 10\nu_0\cdot\sigma\sqrt{\frac{3d+z}{n}}$$

with probability at least $1-e^{-z}$. This proves the claimed bound. □

A.1.3 Proof of Proposition 1.3.2

Given $s,l>0$, define the local neighborhood of $\mathbf{\Theta}^*$

$$\mathbf{\Lambda} = \mathbf{\Lambda}(s,l) = \{\mathbf{\Theta}\in\mathbb{R}^{d_1\times d_2} : \mathbf{\Theta}\in\mathbf{\Theta}^* + \mathbb{B}(s)\cap\mathbb{C}(l)\},$$

where $\mathbb{B}(s)$ and $\mathbb{C}(l)$ are defined in (1.11). Since the Huber loss is convex and differentiable, we have

$$\begin{aligned} D(\boldsymbol{\Theta}) &:= \langle \nabla \widehat{L}_\tau(\boldsymbol{\Theta}) - \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*), \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \rangle \\ &= \frac{1}{n} \sum_{i=1}^n \{ \psi_\tau(y_i - \langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle) - \psi_\tau(y_i - \langle \boldsymbol{X}_i, \boldsymbol{\Theta} \rangle) \} \langle \boldsymbol{X}_i, \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \rangle \\ &\geq \frac{1}{n} \sum_{i=1}^n \{ \psi_\tau(y_i - \langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle) - \psi_\tau(y_i - \langle \boldsymbol{X}_i, \boldsymbol{\Theta} \rangle) \} \langle \boldsymbol{X}_i, \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \rangle \mathbb{1}_{\mathscr{E}_i}, \end{aligned}$$

where $\mathbb{1}_{\mathscr{E}_i}$ is the indicator function of the event

$$\mathscr{E}_i = \big\{ |\varepsilon_i| \leq \tau/2 \big\} \cap \big\{ |\langle \boldsymbol{X}_i, \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \rangle| \leq (\tau/2s) \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \big\}.$$

Noting that $|y_i - \langle \boldsymbol{X}_i, \boldsymbol{\Theta} \rangle| \leq \tau$ for all $\boldsymbol{\Theta} \in \boldsymbol{\Lambda}$ on $\mathscr{E}_i$, and $\ell''_\tau(u) = 1$ for $|u| \leq \tau$, we further obtain

$$D(\boldsymbol{\Theta}) \geq \frac{1}{n} \sum_{i=1}^n \langle \boldsymbol{X}_i, \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \rangle^2 \mathbb{1}_{\mathscr{E}_i}. \tag{A.10}$$

To lower bound the right-hand side of (A.10), we first introduce the following Lipschitz continuous functions. For any given $R > 0$, define the function

$$\varphi_R(x) = \begin{cases} x^2 & \text{if } |x| \leq R/2, \\ \{x - R\mathrm{sign}(x)\}^2 & \text{if } \frac{R}{2} \leq |x| \leq R, \\ 0 & \text{if } |x| > R. \end{cases}$$

It is easy to see that φ_R is R-Lipschitz continuous and satisfies

$$\varphi_{cR}(cx) = c^2 \varphi_R(x) \ \text{ for any } c > 0, \ \text{ and } \ x^2 \mathbb{1}(|x| \leq R/2) \leq \varphi_R(x) \leq x^2 \mathbb{1}(|x| \leq R). \tag{A.11}$$

Therefore,

$$\frac{D(\boldsymbol{\Theta})}{\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^*\|_{\mathrm{F}}^2} \geq \frac{1}{n}\sum_{i=1}^{n} \varphi_{\tau/(2s)}(\langle \boldsymbol{X}_i, \boldsymbol{\Theta}-\boldsymbol{\Theta}^*\rangle / \|\boldsymbol{\Theta}-\boldsymbol{\Theta}^*\|_{\mathrm{F}})\chi_i, \tag{A.12}$$

where $\chi_i = \mathbb{1}(|\varepsilon_i| \leq \tau/2)$. Write $\boldsymbol{\Delta} = \boldsymbol{\Theta}-\boldsymbol{\Theta}^*$ and $\omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) = \varphi_{\tau/(2s)}(\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle / \|\boldsymbol{\Delta}\|_{\mathrm{F}})\chi_i$. In the following, we bound the expectation $\mathbb{E}\omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i)$ and the random fluctuation term

$$\Omega = \sup_{\boldsymbol{\Delta}\in\mathbb{C}(l)} \left| \frac{1}{n}\sum_{i=1}^{n} \{\omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) - \mathbb{E}\omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i)\} \right|,$$

respectively.

Fix $\boldsymbol{\Delta} = \boldsymbol{\Theta}-\boldsymbol{\Theta}^*$ for any $\boldsymbol{\Theta} \in \boldsymbol{\Lambda}$ and write $\bar{\boldsymbol{\Delta}} = \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}$. By (A.11) and Markov's inequality,

$$\begin{aligned}
&\mathbb{E}\omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \\
&\quad \geq \mathbb{E}\langle \boldsymbol{X}_i, \bar{\boldsymbol{\Delta}}\rangle^2 - \mathbb{E}\{\langle \boldsymbol{X}_i, \bar{\boldsymbol{\Delta}}\rangle^2 \mathbb{1}(|\langle \boldsymbol{X}_i, \bar{\boldsymbol{\Delta}}\rangle| \geq \tau/(4s))\} - \mathbb{E}\{\langle \boldsymbol{X}_i, \bar{\boldsymbol{\Delta}}\rangle^2 \mathbb{1}(|\varepsilon_i| \geq \tau/2)\} \\
&\quad \geq c_l - \left\{ \left(\frac{4s}{\tau}\right)^2 \mathbb{E}\langle \boldsymbol{X}_i, \bar{\boldsymbol{\Delta}}\rangle^4 + \left(\frac{2}{\tau}\right)^2 \mathbb{E}(\langle \boldsymbol{X}_i, \bar{\boldsymbol{\Delta}}\rangle^2 \varepsilon_i^2) \right\}.
\end{aligned}$$

Since $\|\bar{\boldsymbol{\Delta}}\|_{\mathrm{F}} = 1$, letting $\tau \geq 4\nu_0\sqrt{(2\sigma_0^2 + 96\nu_0^2 s^2)/c_l}$ and applying the moment bound in (A.8) yields

$$\mathbb{E}\omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \geq c_l - \frac{16\cdot 48\nu_0^4 s^2 + 16\nu_0^2\sigma_0^2}{\tau^2} \geq \frac{1}{2}c_l \ \text{ for all } \ \boldsymbol{\Delta} \in \mathbb{B}(s). \tag{A.13}$$

Next, we evaluate the random fluctuation term Ω which is the supremum of an empirical process indexed by $\boldsymbol{\Delta} \in \mathbb{C}(l)$. By the definition of $\omega_{\boldsymbol{\Delta}}$, we have

$$0 \leq \omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \leq \left(\frac{\tau}{4s}\right)^2, \quad \text{and} \quad \mathbb{E}\omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i)^2 \leq \mathbb{E}\langle \boldsymbol{X}_i, \bar{\boldsymbol{\Delta}}\rangle^4 \leq 48\nu_0^4.$$

By Bousquet's inequality (see, e.g. Theorem 12.5 in Boucheron, Lugosi and Massart (2013)),

for any $z > 0$ it holds with probability at least $1 - e^{-z}$ that

$$\begin{aligned}\Omega &\le \mathbb{E}\Omega + (\mathbb{E}\Omega)^{1/2}\frac{\tau}{2s}\sqrt{\frac{z}{n}} + 4\nu_0^2\sqrt{\frac{6z}{n}} + \frac{\tau^2}{16s^2}\frac{z}{3n} \\ &\le 2\mathbb{E}\Omega + 4\nu_0^2\sqrt{\frac{6z}{n}} + \frac{\tau^2}{16s^2}\frac{4z}{3n}, \end{aligned} \tag{A.14}$$

where the last inequality follows from the elementary inequality that $ab \le a^2 + b^2/4$ for all $a, b \in \mathbb{R}$.

To bound the expectation $\mathbb{E}\Omega$, applying Rademacher symmetrization we get

$$\mathbb{E}\Omega \le 2\mathbb{E}\left\{ \sup_{\boldsymbol{\Delta}\in\mathbb{C}(l)} \frac{1}{n}\sum_{i=1}^{n} e_i \cdot \omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \right\},$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables. Since $\chi_i = \mathbb{1}(|\varepsilon_i| \le \tau/2) \in \{0, 1\}$, we can write $\omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) = \varphi_{\tau/(2s)}(\chi_i \langle \boldsymbol{X}_i, \bar{\boldsymbol{\Delta}} \rangle)$. By the Lipschitz continuity of φ_R, for each sample $(\boldsymbol{X}_i, \varepsilon_i)$ and for any $\boldsymbol{\Delta}, \boldsymbol{\Delta}' \in \mathbb{R}^{d_1 \times d_2}$,

$$|\omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) - \omega_{\boldsymbol{\Delta}'}(\boldsymbol{X}_i, \varepsilon_i)| \le \frac{\tau}{2s} |\chi_i \langle \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}} \rangle - \chi_i \langle \boldsymbol{X}_i, \boldsymbol{\Delta}'/\|\boldsymbol{\Delta}'\|_{\mathrm{F}} \rangle|.$$

Moreover, $\omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) = 0$ whenever $\chi_i \langle \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}} \rangle = 0$. Define the subset $\mathbb{T} \subseteq \mathbb{R}^n$ as

$$\mathbb{T} = \{ \boldsymbol{t} = (t_1, \ldots, t_n)^{\mathrm{T}} : t_i = \chi_i \langle \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}} \rangle, i = 1, \ldots, n, \text{ and } \boldsymbol{\Delta} \in \mathbb{C}(l) \},$$

and the contraction $\phi : \mathbb{R} \to \mathbb{R}$ as $\phi(t) = (2s/\tau) \cdot \varphi_{\tau/(2s)}(t)$. The Lipschitz continuity of φ ensures that $\phi(\cdot)$ is 1-Lipschitz. Applying Talagrand's contraction principle (see, e.g., Theorem 4.12 and

(4.20) in Ledoux and Talagrand (1991)), we have

$$\begin{aligned}
\mathbb{E}\Omega &\le 2\mathbb{E}\left\{ \sup_{\boldsymbol{\Delta}\in\mathbb{C}(l)} \frac{1}{n}\sum_{i=1}^{n} e_i \cdot \omega_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \right\} \\
&= \frac{\tau}{s}\mathbb{E}\left\{ \sup_{t\in T} \frac{1}{n}\sum_{i=1}^{n} e_i \phi(t_i) \right\} \\
&\le \frac{\tau}{s}\mathbb{E}\left(\sup_{t\in T} \frac{1}{n}\sum_{i=1}^{n} e_i t_i \right) \\
&= \frac{\tau}{s}\mathbb{E}\left(\sup_{\boldsymbol{\Delta}\in\mathbb{C}(l)} \frac{1}{n}\sum_{i=1}^{n} e_i \langle \chi_i \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}} \rangle \right) \\
&\le \frac{\tau l}{sn}\mathbb{E}\Big\| \underbrace{\sum_{i=1}^{n} e_i \chi_i \boldsymbol{X}_i}_{=:\boldsymbol{M}} \Big\|_2 ,
\end{aligned} \tag{A.15}$$

where the last inequality follows from the cone constraint that $\|\boldsymbol{\Delta}\|_* \le l\|\boldsymbol{\Delta}\|_{\mathrm{F}}$. It thus remains to bound the spectral norm of the random matrix $\boldsymbol{M}$.

By the variational characterization of the operator norm, we can write

$$\|\boldsymbol{M}\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^{d_1-1}} \sup_{\boldsymbol{v}\in\mathbb{S}^{d_2-1}} \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{v} = \sup_{\boldsymbol{u}\in\mathbb{S}^{d_1-1}} \sup_{\boldsymbol{v}\in\mathbb{S}^{d_2-1}} \sum_{i=1}^{n} e_i \chi_i \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}.$$

By a standard covering argument, there exist a $(1/4)$-net $\mathscr{N}_1$ of $\mathbb{S}^{d_1-1}$ and a $(1/4)$-net $\mathscr{N}_2$ of $\mathbb{S}^{d_2-1}$ with $|\mathscr{N}_1| \le 9^{d_1}$ and $|\mathscr{N}_2| \le 9^{d_2}$ such that

$$\|\boldsymbol{M}\|_2 \le 2 \max_{\boldsymbol{u}\in\mathscr{N}_1} \max_{\boldsymbol{u}\in\mathscr{N}_2} \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{v} = 2 \max_{\boldsymbol{u}\in\mathscr{N}_1} \max_{\boldsymbol{u}\in\mathscr{N}_2} \sum_{i=1}^{n} e_i \chi_i \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}.$$

Therefore,

$$\frac{1}{2\nu_0}\mathbb{E}\|\boldsymbol{M}\|_2 \le \mathbb{E} \max_{\boldsymbol{u}\in\mathscr{N}_1, \boldsymbol{u}\in\mathscr{N}_2} \underbrace{\sum_{i=1}^{n} e_i \chi_i \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}/\nu_0}_{=:M_{\boldsymbol{u},\boldsymbol{v}}} . \tag{A.16}$$

For any $(\boldsymbol{u},\boldsymbol{v}) \in \mathscr{N}_1 \times \mathscr{N}_2$, the moment bound (A.8) implies

$$\mathbb{E}(e_i\chi_i\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}/v_0)^2 \leq 4 \quad \text{and} \quad \mathbb{E}|e_i\chi_i\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{v}/v_0|^k \leq 2\cdot k! = \frac{k!}{2}\cdot 4\cdot 1^{k-2},$$

for $k = 3,4,\ldots$. Following the proof of Theorems 2.10 and 2.5 in Boucheron, Lugosi and Massart (2013), it can be shown that for all $\lambda \in (0,1/c)$,

$$\log \mathbb{E}e^{\lambda M_{\boldsymbol{u},\boldsymbol{v}}} \leq \psi(\lambda) := \frac{v\lambda^2}{2(1-c\lambda)}$$

and hence

$$\mathbb{E}\max_{\boldsymbol{u}\in\mathscr{N}_1,\boldsymbol{u}\in\mathscr{N}_2} M_{\boldsymbol{u},\boldsymbol{v}} \leq \inf_{\lambda\in(0,1/c)} \frac{\log 9^{d_1+d_2}+\psi(\lambda)}{\lambda} = \sqrt{2v\log 9^d} + c\log 9^d,$$

where $v = 4n$, $c = 1$ and $d = d_1 + d_2$. Substituting this bound into (A.16) and then (A.15) yields

$$\mathbb{E}\|\boldsymbol{M}\|_2 \leq 8.4v_0\sqrt{dn} + 4.4v_0 d, \text{ and } \mathbb{E}\Omega \leq \frac{v_0\tau l}{s}\left(8.4\sqrt{\frac{d}{n}} + 4.4\frac{d}{n}\right).$$

Together with the concentration inequality (A.14), this implies that with probability at least $1-e^{-z}$, $\Omega \leq c_l/4$ as long as $n \geq C(v_0/c_l)^2(\tau/s)^2(l^2d+z)$ for some universal constant $C > 0$. Combining this with (A.12) and (A.13) proves Proposition 1.3.2. □

A.1.4 Proof of Theorem 1.3.2

By Proposition 1.3.1, let $\lambda \asymp \sigma_0\sqrt{(d+z)/n}$ and $\tau \asymp \sigma_0\sqrt{n/(d+z)}$ with $d = d_1+d_2$ so that $\lambda \geq 2\|\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2$ with probability at least $1-e^{-z}$. Next, choose $s \asymp \tau$ and $l \asymp (\rho/\sigma_0^q)^{1/2}(n/d)^{q/4}$. Then both (1.12) and (1.14) are satisfied under the sample size scaling $n \gtrsim \max\{(\rho/\sigma_0^q)^{2/(2-q)},1\}(d+z)$. Applying Proposition 1.3.2, we conclude that the local RSC event $\mathscr{E}(s,l,c_l/4)$ occurs with probability at least $1-e^{-z}$. The claimed bounds then follow immediately from Theorem 1.3.1. □

A.2 Proofs of matrix completion

A.2.1 Proof of Proposition 1.3.3

Recall that $\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*) = (1/n)\sum_{i=1}^n \xi_i \boldsymbol{X}_i$, where $\xi_i = \psi_\tau(\varepsilon_i)$. Similarly to the proof of Proposition 1.3.1, we will bound $\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*) - \mathbb{E}\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)$ and $\mathbb{E}\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)$, respectively.

First, we bound $\|\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*) - \mathbb{E}\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2$ using the matrix Bernstein inequality. For any fixed $\boldsymbol{u} = (u_1, \ldots, u_{d_1}) \in \mathbb{S}^{d_1-1}$, we have

$$\boldsymbol{u}^{\mathrm{T}}(\mathbb{E}\xi_i^2 \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}})\boldsymbol{u} = \frac{1}{d_1 d_2}\sum_{j=1}^{d_1}\sum_{k=1}^{d_2} \mathbb{E}\{\xi_i^2 | \boldsymbol{X}_i = \boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\mathrm{T}}(d_2)\}u_j^2 \le \frac{\sigma_0^2}{d_1}\sum_{j=1}^{d_1} u_j^2 = \frac{\sigma_0^2}{d_1}.$$

Taking the supremum over $\boldsymbol{u} \in \mathbb{S}^{d_1-1}$ yields $\|\mathbb{E}(\xi_i^2 \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}})\|_2 \le \sigma_0^2/d_1$. Similarly, it can be shown that $\|\mathbb{E}(\xi_i^2 \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i)\|_2 \le \sigma_0^2/d_2$. Write $\boldsymbol{A}_i = \xi_i \boldsymbol{X}_i - \mathbb{E}(\xi_i \boldsymbol{X}_i) \in \mathbb{R}^{d_1 \times d_2}$. Recall that $d_1 \ge d_2$. We have

$$\max\left(\|\mathbb{E}\boldsymbol{A}_i \boldsymbol{A}_i^{\mathrm{T}}\|_2, \|\mathbb{E}\boldsymbol{A}_i^{\mathrm{T}} \boldsymbol{A}_i\|_2\right) \le \max\left(\|\mathbb{E}\xi_i^2 \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}\|_2, \|\mathbb{E}\xi_i^2 \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i\|_2\right) \le \frac{\sigma_0^2}{d_2}.$$

On the other hand, $\|\boldsymbol{A}_i\|_2 \le 2\tau$ due to the fact that $|\ell_\tau'(u)| \le \tau$ for all $u \in \mathbb{R}$. Applying the matrix Bernstein inequality (see, e.g. Theorem 6.1.1 in Tropp (2015)) with modifications, we obtain that for any $x > 0$,

$$\|\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*) - \mathbb{E}\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2 \le \sigma_0\sqrt{\frac{2x}{d_2 n}} + \frac{2\tau x}{3n}$$

with probability at least $1 - (d_1 + d_2)e^{-x}$. Taking $x = \log d + z$ with $d = d_1 + d_2$, we obtain that with probability at least $1 - e^{-z}$,

$$\|\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*) - \mathbb{E}\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2 \le \sigma_0\sqrt{\frac{2(\log d + z)}{d_2 n}} + \frac{2\tau}{3}\frac{\log d + z}{n}. \tag{A.17}$$

For the second term $\|\mathbb{E}\xi_i \boldsymbol{X}_i\|_2$, note that for any $\boldsymbol{u} = (u_1, \ldots, u_{d_1}) \in \mathbb{S}^{d_1-1}$ and $\boldsymbol{v} =$

$(v_1,\ldots,v_{d_2}) \in \mathbb{S}^{d_2-1}$,

$$\begin{aligned}\boldsymbol{u}^{\mathrm{T}}\mathbb{E}(\xi_i \boldsymbol{X}_i)\boldsymbol{v} &= \frac{1}{d_1 d_2}\sum_{j=1}^{d_1}\sum_{k=1}^{d_2}\mathbb{E}\{\xi_i|\boldsymbol{X}_i = \boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\mathrm{T}}(d_2)\}u_j v_k \\ &\le \frac{1}{d_1 d_2}\sum_{j=1}^{d_1}\sum_{k=1}^{d_2}\frac{1}{\tau}\mathbb{E}\{\varepsilon_i^2|\boldsymbol{X}_i = \boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\mathrm{T}}(d_2)\}|u_j v_k| \\ &\le \frac{\sigma_0^2}{\tau d_1 d_2}\sum_{j=1}^{d_1}\sum_{k=1}^{d_2}|u_j v_k| \le \frac{\sigma_0^2}{\tau\sqrt{d_1 d_2}}.\end{aligned}$$

Taking the supremum over $\boldsymbol{u} \in \mathbb{S}^{d_1-1}$ and $\boldsymbol{v} \in \mathbb{S}^{d_2-1}$, it follows that

$$\|\mathbb{E}\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2 \le \frac{\sigma_0^2}{\tau\sqrt{d_1 d_2}} \le \frac{\sigma_0^2}{\tau d_2}.$$

For any $\sigma \ge \sigma_0$, setting $\tau = \sigma\sqrt{n/\{d_2(z+\log d)\}}$ in (A.17) and the above inequality leads to the conclusion. □

A.2.2 Proof of Proposition 1.3.4

We follow the same notations as in the proof of Proposition 1.3.2. For $\boldsymbol{\Delta} \in \mathbb{R}^{d_1\times d_2}$, define

$$\pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i,\varepsilon_i) = \varphi_{\frac{\tau}{2s}\|\boldsymbol{\Delta}\|_{\mathrm{F}}}(\langle\boldsymbol{X}_i,\boldsymbol{\Delta}\rangle)\chi_i \ \text{ with } \ \chi_i = \mathbb{1}(|\varepsilon_i| \le \tau/2).$$

In view of (A.12), we will show that with high probability,

$$\frac{1}{n}\sum_{i=1}^{n}\pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i,\varepsilon_i) \ge \frac{1}{4d_1 d_2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 - 513l^2\frac{d_1(z+\log d)}{n}\|\boldsymbol{\Delta}\|_\infty^2 \tag{A.18}$$

holds uniformly for all $\boldsymbol{\Delta} \in \mathbb{A}(s,l)$ whenever τ and n are sufficiently large. The claimed bound then follows immediately.

Write $\boldsymbol{\Delta} = (\Delta_{jk})_{1\le j\le d_1, 1\le k\le d_2}$. By (A.11) and Markov's inequality, we have

$$\begin{aligned}
&\mathbb{E}\pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \\
&\ge \mathbb{E}\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle^2 - \mathbb{E}\left\{ \langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle^2 \mathbb{1}\left(|\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle| \ge \frac{\tau}{4s}\|\boldsymbol{\Delta}\|_{\mathrm{F}} \right) \right\} - \mathbb{E}\{ \langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle^2 \mathbb{1}(|\varepsilon_i| \ge \tau/2) \} \\
&\ge \frac{1}{d_1 d_2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 - \left\{ \left(\frac{4s}{\tau\|\boldsymbol{\Delta}\|_{\mathrm{F}}} \right)^2 \mathbb{E}\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle^4 + \left(\frac{2}{\tau} \right)^2 \mathbb{E}(\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle^2 \varepsilon_i^2) \right\}.
\end{aligned}$$

Note that

$$\mathbb{E}\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle^4 = \frac{1}{d_1 d_2} \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \Delta_{jk}^4 \le \frac{1}{d_1 d_2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \|\boldsymbol{\Delta}\|_{\infty}^2,$$

and $\mathbb{E}(\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle^2 \varepsilon_i^2) \le \sigma_0^2 \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2/(d_1 d_2)$. Provided $\tau^2 \ge 16\max[ns^2/\{l^2 d_1^2 d_2(z+\log d)\}, \sigma_0^2]$ with $d = d_1 + d_2$, we have

$$\mathbb{E}\pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \ge \frac{3}{4}\frac{1}{d_1 d_2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 - l^2 \frac{d_1(z+\log d)}{n}\|\boldsymbol{\Delta}\|_{\infty}^2.$$

Consequently, to prove (A.18), it suffices to show that for all $\boldsymbol{\Delta} \in \mathbb{A}(s,l)$, it follows with high probability that

$$\left| \frac{1}{n}\sum_{i=1}^{n} \pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) - \mathbb{E}\pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \right| \le \frac{1}{2d_1 d_2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 + 512\frac{l^2 d_1(z+\log d)}{n}\|\boldsymbol{\Delta}\|_{\infty}^2. \tag{A.19}$$

To prove (A.19), we extend the arguments from the proof of Lemma 12 in Klopp (2014) to deal with the more complex function class $\{\pi_{\boldsymbol{\Delta}} : \mathbb{R}^{d_1 \times d_2} \times \mathbb{R} \to [0,\infty)\}_{\boldsymbol{\Delta} \in \mathbb{A}(s,l)}$. First, set $\eta = 8\sqrt{(z+\log d)/n}$ and define the constrain set

$$\mathbb{D}(l) = \left\{ \boldsymbol{\Delta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Delta}\|_{\infty} = 1, \frac{\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2}{d_1 d_2} \ge \eta \ \text{ and } \ \|\boldsymbol{\Delta}\|_* \le l\|\boldsymbol{\Delta}\|_{\mathrm{F}} \right\}.$$

By (A.11), it holds that $c^2 \pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) = \pi_{c\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i)$ for any $c > 0$. This implies that if $\boldsymbol{\Delta}$ satis-

fies (A.19), then $c\boldsymbol{\Delta}$ would also satisfy (A.19) for any $c>0$. Therefore, we only need to control the probability of the event

$$\mathscr{E}(l) := \Bigg\{ \exists \boldsymbol{\Delta} \in \mathbb{D}(l) \ \text{ such that } \\ \left| \frac{1}{n}\sum_{i=1}^{n} \pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) - \mathbb{E}\pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \right| > \frac{1}{2d_1 d_2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 + 512\frac{l^2 d_1 (z+\log d)}{n} \Bigg\}.$$

We estimate the probability of event $\mathscr{E}(l)$ by a standard peeling argument. For some $\gamma > 1$ to be determined, define the subsets

$$\mathbb{D}_m(l) = \left\{ \boldsymbol{\Delta} \in \mathscr{D}(l) : \gamma^{m-1}\eta \le \frac{1}{d_1 d_2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \le \gamma^m \eta \right\}, \ \ m = 1, 2, \ldots. \tag{A.20}$$

On the event $\mathscr{E}(l)$, there exists some $m \ge 1$ such that $\boldsymbol{\Delta} \in \mathbb{D}_m(l)$ and hence

$$\begin{aligned} \left| \frac{1}{n}\sum_{i=1}^{n} \pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) - \mathbb{E}\pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \right| &> \frac{1}{2d_1 d_2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 + 512\frac{l^2 d_1 (z+\log d)}{n} \\ &\ge \frac{1}{2}\gamma^{m-1}\eta + 512\frac{l^2 d_1 (z+\log d)}{n} \\ &= \frac{1}{2\gamma}\gamma^m \eta + 512\frac{l^2 d_1 (z+\log d)}{n}. \end{aligned} \tag{A.21}$$

Moreover, define the events

$$\mathscr{E}_m(l) = \Bigg\{ \exists \boldsymbol{\Delta} \in \mathbb{D}_m(l) \text{ such that } \\ \left| \frac{1}{n}\sum_{i=1}^{n} \pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) - \mathbb{E}\pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \right| > \frac{1}{2\gamma}\gamma^m \eta + 512\frac{l^2 d_1 (z+\log d)}{n} \Bigg\}, \ \ m \ge 1.$$

Then, (A.21) implies that $\mathscr{E}(l) \subseteq \cup_{m=1}^{\infty} \mathscr{E}_m(l)$ and hence $\mathbb{P}\{\mathscr{E}(l)\} \le \sum_{m\ge 1} \mathbb{P}\{\mathscr{E}_m(l)\}$. The following lemma provides an upper bound of $\mathbb{P}\{\mathscr{E}_m(l)\}$ for each $m \ge 1$. Let

$$Z_m = \sup_{\boldsymbol{\Delta} \in \mathbb{D}_m(l)} \left| \frac{1}{n}\sum_{i=1}^{n} \pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) - \mathbb{E}\pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \right|.$$

Lemma A.2.1. Under Conditions (B2) and (B3), it holds

$$\mathbb{P}\left\{ Z_m > \frac{1}{2\gamma}\gamma^m \eta + (16\gamma)^2 l^2 \frac{d_1(z+\log d)}{n} \right\} \leq e^{-n(\gamma^m \eta)^2/(64\gamma^2)},$$

where $d = d_1 + d_2$.

To conclude the proof, we choose $\gamma = \sqrt{2}$. Then it follows from the union bound and Lemma A.2.1 repeatedly that

$$\begin{aligned}\mathbb{P}\{\mathscr{E}(l)\} \leq \sum_{m=1}^{\infty} \mathbb{P}\{\mathscr{E}_m(l)\} \leq \sum_{m=1}^{\infty} e^{-n2^m\eta^2/128} \leq \sum_{m=1}^{\infty} e^{-mn\eta^2/64} = \sum_{m=1}^{\infty} e^{-m(z+\log d)} \\ = \frac{e^{-z}}{d - e^{-z}} \leq e^{-z},\end{aligned}$$

where the third step uses the basic inequality that $2^m \geq 2m$ for any $m \geq 1$, and the last inequality follows from the assumption $d \geq 2$. This completes the proof. □

A.2.3 Proof of Theorem 1.3.3

The proof employs techniques from the proof of Theorem 3 in Klopp (2014) and Corollary 2 in Negahban and Wainwright (2012) as well as the localized analysis (via proof by contradiction) as in the proof of Theorem 1.3.1. Throughout the proof, we write $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}_{\tau,\lambda}$, $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ and $d = d_1 + d_2$.

For a threshold $t > 0$ to be specified, choose r as

$$r = |\{j \in \{1, 2, \ldots, d_2 : \sigma_j(\mathbf{\Theta}^*) \geq t\}|.$$

Using the same arguments as in the proof of Theorem 1.3.1, we have

$$\|\widehat{\mathbf{\Delta}}\|_* \leq 4\sqrt{2\rho t^{-q}}\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} + 4t^{1-q}\rho. \tag{A.22}$$

We now consider two cases. Suppose first that $\|\widehat{\mathbf{\Delta}}\|_\infty^2 > (8d_1d_2)^{-1}\sqrt{n/(z+\log d)}\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}^2$.

By the definition of $\widehat{\mathbf{\Theta}}$ and condition (B1), we have $\|\widehat{\mathbf{\Delta}}\|_\infty \le 2\alpha_0$. It thus follows that

$$\frac{1}{d_1 d_2}\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}^2 \le 32\alpha_0^2\sqrt{\frac{z+\log d}{n}}.$$

Substituting this into (A.22) and taking $t = \{\alpha_0^2 d_1 d_2 \rho^{-1}\sqrt{(z+\log d)/n}\}^{1/(2-q)}$, which minimizes the right-hand side of (A.22), leads to an upper bound for $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_*$ in this case.

Next suppose that $\|\widehat{\mathbf{\Delta}}\|_\infty^2 \le (8d_1d_2)^{-1}\sqrt{n/(z+\log d)}\cdot\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}^2$. In view of Proposition 1.3.3, we choose

$$\tau \asymp \sigma\sqrt{n/\{d_2(z+\log d)\}}, \quad \text{and } \lambda \asymp \sigma\sqrt{(z+\log d)/(d_2 n)}$$

so that $\lambda \ge 2\|\nabla\widehat{L}_\tau(\mathbf{\Theta}^*)\|_2$ with probability at least $1-e^{-z}$, where $\sigma = \max\{\sigma_0, \alpha_0\}$. Using similar arguments as in Theorem 1.3.1, this implies that with the same probability,

$$\langle \nabla\widehat{L}_\tau(\widehat{\mathbf{\Theta}}) - \nabla\widehat{L}_\tau(\mathbf{\Theta}^*), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\rangle \le \frac{3\lambda}{2}\|\widehat{\mathbf{\Delta}}\|_*. \tag{A.23}$$

Moreover, choose $s = C\sqrt{d_1 d_2\rho}(\sqrt{d_1 d_2}\lambda)^{1-q/2}$ for a sufficiently large constant $C > 0$ and assume that $\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} > s$. Then, there exists $\eta \in (0,1)$ such that $\widetilde{\mathbf{\Theta}}_\eta := \mathbf{\Theta}^* + \eta(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)$ satisfies $\|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}^*\|_{\mathrm{F}} = s$. Taking $t = \sqrt{d_1 d_2}\lambda \asymp \sigma\sqrt{d_1(z+\log d)/n}$, we have $(d_1 d_2)^{-1/2}\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} > C\sqrt{\rho}\, t^{1-q/2}$. Substituting this into (A.22) gives

$$\|\widehat{\mathbf{\Delta}}\|_* \le 4\sqrt{2\rho t^{-q}}\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} + 4\sqrt{\rho t^{-q}}\cdot\sqrt{\rho}\, t^{1-q/2} < (4\sqrt{2}+4/C)\sqrt{\rho}\, t^{-q/2}\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}},$$

and hence $\|\widehat{\mathbf{\Delta}}\|_* \le l\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}$ with $l = (4\sqrt{2}+4/C)\sqrt{\rho}\, t^{-q/2}$. This means that $\widetilde{\mathbf{\Delta}}_\eta := \widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}^* \in \mathbb{A}(s,l)$. Applying Lemma F.2 in Fan *et al.* (2018) to the loss function $\widehat{L}_\tau(\cdot)$ and by (A.23),

$$\begin{aligned}\langle \nabla\widehat{L}_\tau(\widetilde{\mathbf{\Theta}}_\eta) - \nabla\widehat{L}_\tau(\mathbf{\Theta}^*), \widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}^*\rangle &\le \eta\langle \nabla\widehat{L}_\tau(\widehat{\mathbf{\Theta}}) - \nabla\widehat{L}_\tau(\mathbf{\Theta}^*), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\rangle &\text{(A.24)}\\ &\le \frac{3}{2}\lambda\eta\|\widehat{\mathbf{\Delta}}\|_* \le \frac{3}{2}\lambda l\|\widetilde{\mathbf{\Delta}}_\eta\|_{\mathrm{F}}. &\text{(A.25)}\end{aligned}$$

To lower bound the left-hand side of (A.24), note from the above choice of (λ, s, l, t) that

$$\begin{aligned}\frac{ns^2}{l^2 d_1^2 d_2(z+\log d)} &= \left(\frac{C}{4\sqrt{2}+4/C}\right)^2 \frac{nd_1d_2\rho t^{2-q}}{\rho t^{-q} d_1^2 d_2(z+\log d)} \\ &= \left(\frac{C}{4\sqrt{2}+4/C}\right)^2 \frac{nt^2}{d_1(z+\log d)} \asymp \sigma^2,\end{aligned}$$

which implies $\tau^2 \gtrsim \max[ns^2/\{l^2 d_1^2 d_2(z+\log d)\}, \sigma_0^2]$ as long as $n \gtrsim d_2(z+\log d)$. Then, it follows from Proposition 1.3.4 that with probability at least $1-2d^{-1}$,

$$\langle \nabla \widehat{L}_\tau(\widetilde{\boldsymbol{\Theta}}_\eta) - \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*), \widetilde{\boldsymbol{\Theta}}_\eta - \boldsymbol{\Theta}^* \rangle \geq \frac{1}{4d_1d_2}\|\widetilde{\boldsymbol{\Delta}}_\eta\|_{\mathrm{F}}^2 - C_1\rho\, t^{-q}\frac{d_1(z+\log d)}{n}\|\widetilde{\boldsymbol{\Delta}}_\eta\|_\infty^2 \qquad \text{(A.26)}$$

for some constant $C_1 > 0$ independent of (n, d_1, d_2). Since $\widetilde{\boldsymbol{\Delta}}_\eta = \eta\widehat{\boldsymbol{\Delta}}$, we also have $\|\widetilde{\boldsymbol{\Delta}}_\eta\|_\infty \leq \|\widehat{\boldsymbol{\Delta}}\|_\infty \leq 2\alpha_0$. Combining this with (A.25) and (A.26), we conclude that with probability at least $1-2e^{-z}$,

$$\frac{1}{4d_1d_2}\|\widetilde{\boldsymbol{\Delta}}_\eta\|_{\mathrm{F}}^2 \leq C_1\rho\, t^{-q}\frac{d_1(z+\log d)}{n}\|\widetilde{\boldsymbol{\Delta}}_\eta\|_\infty^2 + 3\lambda l\sqrt{d_1d_2}\cdot\frac{\|\widetilde{\boldsymbol{\Delta}}_\eta\|_{\mathrm{F}}}{2\sqrt{d_1d_2}}.$$

Set $y = \frac{\|\widetilde{\boldsymbol{\Delta}}_\eta\|_{\mathrm{F}}}{2\sqrt{d_1d_2}}$, and note that $y^2 \leq a + by$ for some $a, b > 0$. It then follows that $y \leq b + \sqrt{a}$, which in turn implies

$$\begin{aligned}\frac{\|\widetilde{\boldsymbol{\Delta}}_\eta\|_{\mathrm{F}}}{\sqrt{d_1d_2}} &\leq 6l\underbrace{\sqrt{d_1d_2}\lambda}_{=t} + 4\sqrt{C_1}\sqrt{\rho}\, t^{-q/2}\alpha_0\sqrt{\frac{d_1(z+\log d)}{n}} \\ &\leq 24(\sqrt{2}+1/C)\sqrt{\rho}\, t^{1-q/2} + 4\sqrt{C_1}\sqrt{\rho}\, t^{-q/2}\alpha_0\sqrt{\frac{d_1(z+\log d)}{n}} \\ &\lesssim \sigma^{1-q/2}\sqrt{\rho}\left\{\frac{d_1(z+\log d)}{n}\right\}^{1/2-q/4}.\end{aligned}$$

A sufficiently large constant C in the definition of s ensures that $\|\widetilde{\boldsymbol{\Delta}}_\eta\|_{\mathrm{F}} < s$. This, however,

contradicts the fact that $\|\widetilde{\boldsymbol{\Delta}}_\eta\|_{\mathrm{F}} = s$ by construction. Therefore, we must have

$$\frac{1}{\sqrt{d_1 d_2}}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}} \leq \frac{1}{\sqrt{d_1 d_2}} s \asymp \sigma^{1-q/2}\sqrt{\rho}\left\{\frac{d_1(z+\log d)}{n}\right\}^{1/2-q/4}$$

with high probability, as claimed.

Combining this Frobenius norm error bound with (A.22) proves the error bound under nuclear norm, which completes the proof. □

A.2.4 Proof of Lemma A.2.1

To bound $Z_m = \sup_{\boldsymbol{\Delta}\in\mathbb{D}_m(l)} |(1/n)\sum_{i=1}^n \pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) - \mathbb{E}\pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i)|$, note that $|\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle| \leq 1$ for any $\boldsymbol{\Delta} \in \mathbb{D}_m(l)$, and hence $0 \leq \pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i) \leq \langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle^2 \leq 1$ by (A.11). Applying Theorem 3.26 in Wainwright (2019), a functional Hoeffding inequality, yields

$$\mathbb{P}\big(Z_m \geq \mathbb{E}Z_m + x\big) \leq e^{-nx^2/4} \text{ for any } x \geq 0. \tag{A.27}$$

To bound the expectation $\mathbb{E}Z_m$, by Rademacher symmetrization we have

$$\mathbb{E}Z_m \leq 2\mathbb{E}\left\{\sup_{\boldsymbol{\Delta}\in\mathbb{D}_m(l)} \frac{1}{n}\sum_{i=1}^n e_i \pi_{\boldsymbol{\Delta}}(\boldsymbol{X}_i, \varepsilon_i)\right\} = 2\mathbb{E}\left\{\sup_{\boldsymbol{\Delta}\in\mathbb{D}_m(l)} \frac{1}{n}\sum_{i=1}^n e_i \varphi_{\frac{\tau}{2s}\|\boldsymbol{\Delta}\|_{\mathrm{F}}}(\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle)\chi_i\right\},$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables. By (A.11),

$$\varphi_{\frac{\tau}{2s}\|\boldsymbol{\Delta}\|_{\mathrm{F}}}(\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle) = \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \cdot \varphi_{\frac{\tau}{2s}}(\langle \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}\rangle).$$

By the definition of $\mathbb{D}_m(l)$ in (A.20), the expectation $\mathbb{E}Z_m$ is further bounded as

$$\begin{aligned}
\mathbb{E}Z_m &\leq 2\mathbb{E}\left\{\sup_{\boldsymbol{\Delta}\in\mathbb{D}_m(l)} \frac{\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2}{n}\sum_{i=1}^n e_i \varphi_{\frac{\tau}{2s}}(\langle \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}\rangle\chi_i\right\} \\
&\leq 2\cdot d_1 d_2 \gamma^m \eta \cdot \mathbb{E}\left\{\sup_{\boldsymbol{\Delta}\in\mathbb{D}_m(l)} \left|\frac{1}{n}\sum_{i=1}^n e_i \varphi_{\frac{\tau}{2s}}(\langle \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}\rangle)\chi_i\right|\right\}.
\end{aligned}$$

Since $\chi_i = \mathbb{1}(|\varepsilon_i| \le \tau/2) \in \{0,1\}$, we can write $\varphi_{\tau/(2s)}(\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle)\chi_i = \varphi_{\tau/(2s)}(\chi_i\langle \boldsymbol{X}_i, \boldsymbol{\Delta}\rangle)$ for any $\boldsymbol{\Delta}$. Also, note that $|\langle \chi_i \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}\rangle| \le (d_1 d_2 \gamma^{m-1}\eta)^{-1/2}$ for any $\boldsymbol{\Delta} \in \mathbb{D}_m(l)$. By the definition of $\varphi_R(\cdot)$, for each sample $(\boldsymbol{X}_i, \varepsilon_i)$ and for any $\boldsymbol{\Delta}, \boldsymbol{\Delta}' \in \mathbb{D}_m(l)$,

$$\begin{aligned}
&\left|\varphi_{\frac{\tau}{2s}}(\chi_i\langle \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}\rangle) - \varphi_{\frac{\tau}{2s}}(\chi_i\langle \boldsymbol{X}_i, \boldsymbol{\Delta}'/\|\boldsymbol{\Delta}'\|_{\mathrm{F}}\right| \\
&\qquad \le \frac{2}{\sqrt{d_1 d_2 \gamma^{m-1}\eta}} |\langle \chi_i \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}\rangle - \langle \chi_i \boldsymbol{X}_i, \boldsymbol{\Delta}'/\|\boldsymbol{\Delta}'\|_{\mathrm{F}}\rangle|.
\end{aligned}$$

Moreover, $\varphi_{\tau/(2s)}(\chi_i\langle \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}) = 0$ whenever $\chi_i\langle \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}\rangle = 0$. Define the subset $\mathbb{T} \in \mathbb{R}^n$ as

$$\mathbb{T} = \{\boldsymbol{t} = (t_1, \ldots, t_n)^{\mathrm{T}} : t_i = \chi_i\langle \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}\rangle,\, i = 1, \ldots, n,\ \ \boldsymbol{\Delta} \in \mathbb{D}_m(l)\},$$

and the contraction $\phi : \mathbb{R} \to \mathbb{R}$ as $\phi(t) = \varphi_{\tau/(2s)}(t)$. Applying Talagrand's contraction principle yields

$$\begin{aligned}
\mathbb{E}Z_m &\le 2 \cdot d_1 d_2 \gamma^m \eta \cdot \mathbb{E}\left\{ \sup_{\boldsymbol{\Delta} \in \mathbb{D}_m(l)} \left| \frac{1}{n}\sum_{i=1}^n e_i \varphi_{\frac{\tau}{2s}}(\langle \chi_i \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}\rangle) \right| \right\} \\
&\le 8 \cdot \sqrt{d_1 d_2 \gamma^{m+1}\eta} \cdot \mathbb{E}\left\{ \sup_{\boldsymbol{\Delta} \in \mathbb{D}_m(l)} \left| \frac{1}{n}\sum_{i=1}^n \langle e_i \chi_i \boldsymbol{X}_i, \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_{\mathrm{F}}\rangle \right| \right\} \\
&\le 8l \cdot \sqrt{d_1 d_2 \gamma^{m+1}\eta} \cdot \mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^n e_i \chi_i \boldsymbol{X}_i \right\|_2,
\end{aligned} \tag{A.28}$$

where the last step follows from the definition of $\mathbb{D}_m(l)$ and the inequality $\langle \boldsymbol{A}, \boldsymbol{B}\rangle \le \|\boldsymbol{A}\|_* \|\boldsymbol{B}\|_2$.

To bound the expected value of the operator norm, we apply the matrix Bernstein inequality. Note that $e_i \chi_i \boldsymbol{X}_i$ is a zero-mean random matrix with $\|e_i \chi_i \boldsymbol{X}_i\|_2 \le 1$. Also, for any $\boldsymbol{u} = (u_1, \ldots, u_{d_1}) \in \mathbb{S}^{d_1 - 1}$, we have

$$\boldsymbol{u}^{\mathrm{T}}(\mathbb{E}\boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}})\boldsymbol{u} = \frac{1}{d_1 d_2}\sum_{j=1}^{d_1}\sum_{k=1}^{d_2} u_j^2 = \frac{1}{d_1}.$$

Taking the supremum over $\boldsymbol{u} \in \mathbb{S}^{d_1-1}$ yields $\|\mathbb{E}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\|_2 = 1/d_1$. Similarly, it can be shown that $\|\mathbb{E}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\|_2 = 1/d_2$. Recall that $d_1 \geq d_2$. Applying the matrix Bernstein inequality (see, e.g. Theorem 6.1.1 in Tropp (2015)), we have

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} e_i\chi_i\boldsymbol{X}_i\right\|_2 \leq \sqrt{\frac{2\log d}{nd_2}} + \frac{1}{3}\frac{\log d}{n} \leq 2\sqrt{\frac{\log d}{nd_2}},$$

where the last inequality holds as long as $n \geq d_2 \log d$. Combining this with (A.28), we obtain $\mathbb{E}Z_m \leq 16l\sqrt{\gamma^{m+1}\eta d_1 \log(d)/n}$. By the elementary inequality that $ab \leq a^2/(4\gamma) + \gamma b^2$ for any $a, b \in \mathbb{R}$, it follows that

$$\mathbb{E}Z_m + \frac{1}{4\gamma}\gamma^m\eta \leq \frac{1}{2\gamma}\gamma^m\eta + (16\gamma)^2 l^2 \frac{d_1\log d}{n} \leq \frac{1}{2\gamma}\gamma^m\eta + (16\gamma)^2 l^2 \frac{d_1(z+\log d)}{n}.$$

This, joint with the concentration inequality (A.27) (taking $x = (4\gamma)^{-1}\gamma^m\eta$), implies

$$\begin{aligned}
&\mathbb{P}\left\{Z_m \geq \frac{1}{2\gamma}\gamma^m\eta + (16\gamma)^2 l^2 \frac{d_1(z+\log d)}{n}\right\} \\
&\leq \mathbb{P}\left(Z_m \geq \mathbb{E}Z_m + \frac{1}{4\gamma}\gamma^m\eta\right) \leq e^{-n(\gamma^m\eta)^2/(64\gamma^2)},
\end{aligned}$$

as claimed. □

A.3 Proofs of multitask regression

A.3.1 Proof of Proposition 1.3.5

By the definition of $\widehat{L}_\tau(\cdot)$ in (1.19) and the chain rule, we have

$$\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_\tau(\|\boldsymbol{\varepsilon}_i\|_2)}{\|\boldsymbol{\varepsilon}_i\|_2}\boldsymbol{x}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\min\{\|\boldsymbol{\varepsilon}_i\|_2, \tau\}}{\|\boldsymbol{\varepsilon}_i\|_2}\boldsymbol{x}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}} \in \mathbb{R}^{d_1\times d_2},$$

where $\psi_\tau(u) = \ell_\tau'(u) = \mathrm{sign}(u)\min(|u|, \tau)$. Similarly to the proof of Proposition 1.3.1, we will bound the spectral norms of $\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*) - \mathbb{E}\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*)$ and $\mathbb{E}\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*)$, respectively.

First, we bound $\|\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*) - \mathbb{E}\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2$ using the matrix Bernstein inequality. Define symmetric random matrices $\boldsymbol{A}_i \in \mathbb{R}^{d\times d}$ $(d = d_1 + d_2)$ as

$$\boldsymbol{A}_i = \begin{bmatrix} \mathbf{0}_{d_1\times d_1} & \boldsymbol{B}_i \\ \boldsymbol{B}_i^{\mathrm{T}} & \mathbf{0}_{d_2\times d_2} \end{bmatrix} \text{ with } \boldsymbol{B}_i = \frac{\psi_\tau(\|\boldsymbol{\varepsilon}_i\|_2)}{\|\boldsymbol{\varepsilon}_i\|_2}\boldsymbol{x}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}} \in \mathbb{R}^{d_1\times d_2}.$$

It remains to bound the spectral norm of $\boldsymbol{S} := \sum_{i=1}^n(\boldsymbol{A}_i - \mathbb{E}\boldsymbol{A}_i)$. We remark that $\|\boldsymbol{S}\|_2 = \max\{\lambda_{\max}(\boldsymbol{S}), |\lambda_{\min}(\boldsymbol{S})|\}$. Thus, we have $\mathbb{P}(\|\boldsymbol{S}\|_2 \geq u) \leq \mathbb{P}\{\lambda_{\max}(\boldsymbol{S}) \geq u\} + \mathbb{P}\{\lambda_{\max}(-\boldsymbol{S}) \geq u\}$ for any $u \geq 0$. Since the maximum eigenvalue can be represented as the supremum of an empirical process, via Rademacher symmetrization the problem boils down to bounding $\|\sum_{i=1}^n e_i\boldsymbol{A}_i\|_2$, where $e_1, \ldots, e_n$ are independent Rademacher random variables that are independent of the observations. See, for example, Example 6.14 in Wainwright (2019).

Next we show that the zero-mean symmetric random matrices $e_i\boldsymbol{A}_i$ satisfy Bernstein's condition for matrices. Note that $\mathbb{E}(e_i\boldsymbol{A}_i)^p = \mathbf{0}$ for any odd integer $p > 2$. For any even integer $p > 2$, i.e. $p = 2m$ $(m \geq 2)$, we have

$$(e_i\boldsymbol{A}_i)^p = \begin{bmatrix} (\boldsymbol{B}_i\boldsymbol{B}_i^{\mathrm{T}})^m & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{B}_i)^m \end{bmatrix},$$

implying that $\|\mathbb{E}(e_i\boldsymbol{A}_i)^p\|_2 \leq \max\{\|\mathbb{E}(\boldsymbol{B}_i\boldsymbol{B}_i^{\mathrm{T}})^m\|_2, \|\mathbb{E}(\boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{B}_i)^m\|_2\}$.

Starting with $\mathbb{E}(\boldsymbol{B}_i\boldsymbol{B}_i^{\mathrm{T}})^m \in \mathbb{R}^{d_1\times d_1}$, it holds for any $\boldsymbol{u} \in \mathbb{S}^{d_1-1}$ that

$$\boldsymbol{u}^{\mathrm{T}}\mathbb{E}\left\{\frac{\psi_\tau^{2m}(\|\boldsymbol{\varepsilon}_i\|_2)}{\|\boldsymbol{\varepsilon}_i\|_2^{2m}}(\boldsymbol{x}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}}\boldsymbol{\varepsilon}_i\boldsymbol{x}_i^{\mathrm{T}})^m\right\}\boldsymbol{u} = \mathbb{E}\big\{\psi_\tau^{2m}(\|\boldsymbol{\varepsilon}_i\|_2)(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{u})^2(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i)^{m-1}\big\}. \tag{A.29}$$

Recall that $|\psi_\tau(\cdot)| \leq \tau$ and by Condition (C3), $\mathbb{E}\{\psi_\tau^2(\|\boldsymbol{\varepsilon}_i\|_2)|\boldsymbol{x}_i\} \leq \mathbb{E}(\|\boldsymbol{\varepsilon}_i\|_2^2|\boldsymbol{x}_i) \leq \sigma_0^2 d_2$. Concerning the moments of $\boldsymbol{x}_i$, under the sub-Gaussian assumption (C2), for each $k \geq 1$ and $\boldsymbol{u} \in \mathbb{S}^{d_1-1}$,

we have

$$\mathbb{E}|\langle \boldsymbol{u}, \boldsymbol{x}_i \rangle / v_0|^{2k} = 2k \int_0^\infty u^{2k-1} \mathbb{P}\big(|\langle \boldsymbol{u}, \boldsymbol{x}_i \rangle| \geq v_0 t\big) \mathrm{d}u$$
$$\leq 4k \int_0^\infty u^{2k-1} e^{-u^2/2} \mathrm{d}u = 4k \int_0^\infty (2u)^{k-1} e^{-u} \mathrm{d}u = 2^{k+1} k \Gamma(k-1) = 2^{k+1} k!.$$

Applying the Cauchy-Schwarz inequality and the above moment bound, the expected value in (A.29) is further bounded by

$$\begin{aligned}
&\tau^{2m-2} \cdot \sigma_0^2 d_2 \cdot \mathbb{E}\big\{ (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u})^2 (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i)^{m-1} \big\} \\
&\leq \tau^{2m-2} \cdot \sigma_0^2 d_2 \cdot \big\{ \mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u})^4 \big\}^{1/2} \big\{ \mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i)^{2m-2} \big\}^{1/2} \\
&\leq \tau^{2m-2} \cdot 4 v_0^2 \sigma_0^2 d_2 \cdot \big\{ \mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i)^{2m-2} \big\}^{1/2}. \qquad \text{(A.30)}
\end{aligned}$$

To bound $\mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i)^q$ for any $q \geq 2$, using the higher-order moment bound again yields

$$\mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i)^q = \mathbb{E}\bigg(\sum_{j=1}^{d_1} x_{ij}^2 \bigg)^q \leq d_1^{q-1} \mathbb{E}\bigg(\sum_{j=1}^{d_1} x_{ij}^{2q} \bigg) \leq d_1^q \cdot 2^{q+1} q! \cdot v_0^{2q}. \qquad \text{(A.31)}$$

Substituting this (with $q = 2m-2$) into (A.30), we obtain

$$\begin{aligned}
\mathbb{E}\big\{ \psi_\tau^{2m}(\|\boldsymbol{\varepsilon}_i\|_2) (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u})^2 (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i)^{m-1} \big\} &\leq \tau^{2m-2} \cdot 4\sqrt{2} v_0^2 \sigma_0^2 d_2 \cdot d_1^{m-1} 2^{m-1} \sqrt{(2m-2)!} \cdot v_0^{2m-2} \\
&\leq (\tau v_0 \sqrt{d_1})^{2m-2} \cdot 4\sqrt{2} v_0^2 \sigma_0^2 d_2 \cdot 2^{m-1} (2m-2)^{m-1} \\
&= (2e\tau v_0 \sqrt{d_1})^{2m-2} \cdot 4\sqrt{2} v_0^2 \sigma_0^2 d_2 \cdot \bigg(\frac{m-1}{e} \bigg)^{m-1} \\
&\leq (e\tau v_0 \sqrt{2d_1})^{2m-2} \cdot 4\sqrt{2} v_0^2 \sigma_0^2 d_2 \cdot 2^{m-1} (m-1)!,
\end{aligned}$$

where the second and third inequalities follow from the fact that $(k/e)^k \leq k! \leq k^k$ for any positive

integer k. Taking the supremum over $\boldsymbol{u} \in \mathbb{S}^{d_1-1}$ gives, for any $m \geq 2$, that

$$\begin{aligned}\|\mathbb{E}(\boldsymbol{B}_i\boldsymbol{B}_i^{\mathrm{T}})^m\|_2 &\leq 2^{m-1}(m-1)!\cdot 4\sqrt{2}\nu_0^2\sigma_0^2 d_2\cdot(e\tau\nu_0\sqrt{2d_1})^{2m-2}\\ &\leq \frac{1}{2}(2m)!\cdot 4\nu_0^2\sigma_0^2 d_2\cdot(e\tau\nu_0\sqrt{2d_1})^{2m-2}.\end{aligned}$$

Moreover, it is easy to see that $\|\mathbb{E}(\boldsymbol{B}_i\boldsymbol{B}_i^{\mathrm{T}})\|_2 \leq 4\nu_0^2\sigma_0^2 d_2$.

Turning to $\mathbb{E}(\boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{B}_i)^m \in \mathbb{R}^{d_2\times d_2}$, note that for any $\boldsymbol{v}\in\mathbb{S}^{d_2-1}$,

$$\begin{aligned}&\boldsymbol{v}^{\mathrm{T}}\mathbb{E}\left\{\frac{\psi_\tau^{2m}(\|\boldsymbol{\varepsilon}_i\|_2)}{\|\boldsymbol{\varepsilon}_i\|_2^{2m}}(\boldsymbol{\varepsilon}_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}})^m\right\}\boldsymbol{v} = \mathbb{E}\left\{\frac{\psi_\tau^{2m}(\|\boldsymbol{\varepsilon}_i\|_2)}{\|\boldsymbol{\varepsilon}_i\|_2^{2}}(\boldsymbol{\varepsilon}_i^{\mathrm{T}}\boldsymbol{v})^2\|\boldsymbol{x}_i\|_2^{2m}\right\}\\ &\leq \tau^{2m-2}\mathbb{E}\left\{\frac{\psi_\tau^{2}(\|\boldsymbol{\varepsilon}_i\|_2)}{\|\boldsymbol{\varepsilon}_i\|_2^{2}}(\boldsymbol{\varepsilon}_i^{\mathrm{T}}\boldsymbol{v})^2\|\boldsymbol{x}_i\|_2^{2m}\right\}\\ &\leq \tau^{2m-2}\mathbb{E}\big\{(\boldsymbol{\varepsilon}_i^{\mathrm{T}}\boldsymbol{v})^2\|\boldsymbol{x}_i\|_2^{2m}\big\}\\ &\leq \tau^{2m-2}\sigma_0^2\cdot\mathbb{E}\|\boldsymbol{x}_i\|_2^{2m}.\end{aligned}$$

Taking $q = m \geq 2$ in (A.31) yields $\mathbb{E}\|\boldsymbol{x}_i\|_2^{2m} \leq \nu_0^{2m}d_1^m 2^{m+1}m! \leq 2\nu_0^{2m}d_1^m(2m)!$. Substituting this into the above bound, and taking the supremum over $\boldsymbol{v}\in\mathbb{S}^{d_2-1}$, we conclude that

$$\|\mathbb{E}(\boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{B}_i)^m\|_2 \leq \frac{1}{2}(2m)!\cdot 4\nu_0^2\sigma_0^2 d_1\cdot(\tau\nu_0\sqrt{d_1})^{2m-2},\ \ m\geq 2.$$

In particular, $\|\mathbb{E}(\boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{B}_i)\|_2 \leq \sigma_0^2 d_1$.

Combining the above bounds on $\|\mathbb{E}(\boldsymbol{B}_i\boldsymbol{B}_i^{\mathrm{T}})^m\|_2$ and $\|\mathbb{E}(\boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{B}_i)^m\|_2$, we obtain

$$\|\mathbb{E}(e_i\boldsymbol{A}_i)^p\|_2 \leq \frac{1}{2}p!\cdot 4\nu_0^2\sigma_0^2 d_1\cdot(e\tau\nu_0\sqrt{2d_1})^{p-2},\ \ \text{valid for any even integer } p>2. \tag{A.32}$$

Applying the matrix Bernstein inequality (see, e.g. Theorem 6.2 in Tropp (2012)), we obtain that

for any $x > 0$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} e_i \boldsymbol{A}_i \right\|_2 \lesssim \nu_0 \sqrt{\max(d_1, d_2)} \left(\sigma_0 \sqrt{\frac{x}{n}} + \frac{\tau x}{n} \right)$$

with probability at least $1 - 2de^{-x}$. Taking $x = \log(2d) + z \leq 2\log d + z$ for given $z > 0$, it follows that

$$\|\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*) - \mathbb{E}\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2 \lesssim \nu_0 \sqrt{\max(d_1, d_2)} \left\{ \sigma_0 \sqrt{\frac{z + \log d}{n}} + \frac{\tau(z + \log d)}{n} \right\} \tag{A.33}$$

with probability at least $1 - e^{-z}$.

For the deterministic term $\|\mathbb{E}\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2$, since $\mathbb{E}(\boldsymbol{\varepsilon}_i | \boldsymbol{x}_i) = \boldsymbol{0}$ and $\psi_\tau(u) = \min(u, \tau)$ for $u \geq 0$, we have for any $\boldsymbol{u} \in \mathbb{S}^{d_1 - 1}$ and $\boldsymbol{v} \in \mathbb{S}^{d_2 - 1}$ that

$$\begin{aligned}
\boldsymbol{u}^{\mathrm{T}} \{\mathbb{E}\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\} \boldsymbol{v} &= \mathbb{E}\left\{ \frac{\psi_\tau(\|\boldsymbol{\varepsilon}_i\|_2)}{\|\boldsymbol{\varepsilon}_i\|_2} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u} \cdot \boldsymbol{\varepsilon}_i^{\mathrm{T}} \boldsymbol{v} \right\} \\
&= \mathbb{E}\left\{ \frac{\min(\|\boldsymbol{\varepsilon}_i\|_2, \tau) - \|\boldsymbol{\varepsilon}_i\|_2}{\|\boldsymbol{\varepsilon}_i\|_2} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u} \cdot \boldsymbol{\varepsilon}_i^{\mathrm{T}} \boldsymbol{v} \right\} \\
&= \mathbb{E}\left\{ \frac{\tau - \|\boldsymbol{\varepsilon}_i\|_2}{\|\boldsymbol{\varepsilon}_i\|_2} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u} \cdot \boldsymbol{\varepsilon}_i^{\mathrm{T}} \boldsymbol{v} \cdot \mathbb{1}(\|\boldsymbol{\varepsilon}_i\|_2 > \tau) \right\}.
\end{aligned}$$

Noting that $\mathbb{1}(\|\boldsymbol{\varepsilon}_i\|_2 > \tau) \leq \tau^{-2} \|\boldsymbol{\varepsilon}_i\|_2^2$, this further implies

$$|\boldsymbol{u}^{\mathrm{T}} \{\mathbb{E}\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\} \boldsymbol{v}| \leq \frac{1}{\tau} \mathbb{E}\{|\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u}| \cdot \|\boldsymbol{\varepsilon}_i\|_2 |\boldsymbol{\varepsilon}_i^{\mathrm{T}} \boldsymbol{v}|\}.$$

By Condition (C3) and the Cauchy-Schwarz inequality,

$$\mathbb{E}\big\{ \|\boldsymbol{\varepsilon}_i\|_2 |\boldsymbol{\varepsilon}_i^{\mathrm{T}} \boldsymbol{v}| \big| \boldsymbol{x}_i \big\} \leq \big(\mathbb{E}\|\boldsymbol{\varepsilon}_i\|_2^2 \big)^{1/2} \big\{ \mathbb{E}(\boldsymbol{\varepsilon}_i^{\mathrm{T}} \boldsymbol{v})^2 \big\}^{1/2} \leq \sigma_0^2 \sqrt{d_2},$$

and $\mathbb{E}|\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u}| \leq \sqrt{\mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u})^2} \leq 2\nu_0$ due to (A.8). Putting together the pieces, and taking the supre-

mum over $\boldsymbol{u} \in \mathbb{S}^{d_1-1}$ and $\boldsymbol{v} \in \mathbb{S}^{d_2-1}$, we conclude that

$$\|\mathbb{E}\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2 \le 2\nu_0\sigma_0^2\sqrt{d_2}/\tau. \tag{A.34}$$

Finally, taking $\tau = \sigma\sqrt{n/(z+\log d)}$ for any $\sigma \ge \sigma_0$ in (A.33) and (A.34) proves the claim. $\square$

A.3.2 Proof of Proposition 1.3.6

To begin with, note that for each i, the function $f_i(\cdot) : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}$ defined as $f_i(\boldsymbol{\Theta}) = \ell_\tau(\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2)$ is a convex function since it is a composition of two convex functions. By convexity, $\langle \nabla f_i(\boldsymbol{\Theta}) - \nabla f_i(\boldsymbol{\Theta}^*), \boldsymbol{\Theta} - \boldsymbol{\Theta}^*\rangle \ge 0$ for any i, and hence

$$\begin{aligned}
D(\boldsymbol{\Theta}) &:= \langle \nabla \widehat{L}_\tau(\boldsymbol{\Theta}) - \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*), \boldsymbol{\Theta} - \boldsymbol{\Theta}^*\rangle \\
&= \frac{1}{n}\sum_{i=1}^n \left\langle \frac{\psi_\tau(\|\boldsymbol{\varepsilon}_i\|_2)}{\|\boldsymbol{\varepsilon}_i\|_2}\boldsymbol{x}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}} - \frac{\psi_\tau(\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2)}{\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2}\boldsymbol{x}_i(\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i)^{\mathrm{T}}, \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \right\rangle \\
&\ge \frac{1}{n}\sum_{i=1}^n \left\langle \frac{\psi_\tau(\|\boldsymbol{\varepsilon}_i\|_2)}{\|\boldsymbol{\varepsilon}_i\|_2}\boldsymbol{x}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}} - \frac{\psi_\tau(\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2)}{\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2}\boldsymbol{x}_i(\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i)^{\mathrm{T}}, \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \right\rangle \mathbb{1}_{\mathscr{E}_i},
\end{aligned}$$

where $\psi_\tau(\cdot) = \ell_\tau'(\cdot)$, and $\mathbb{1}_{\mathscr{E}_i}$ is the indicator function of the event

$$\mathscr{E}_i = \big\{\|\boldsymbol{\varepsilon}_i\|_2 \le \tau/2\big\} \cap \big\{\|(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)^{\mathrm{T}}\boldsymbol{x}_i\|_2 \le (2s)^{-1}\tau\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}\big\}.$$

On event $\mathscr{E}_i$, observe that $\psi_\tau(\|\boldsymbol{\varepsilon}_i\|_2) = \|\boldsymbol{\varepsilon}_i\|_2$ and $\psi_\tau(\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2) = \|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2$ for all $\boldsymbol{\Theta} \in \boldsymbol{\Theta}^* + \mathbb{B}(s)$. Consequently,

$$D(\boldsymbol{\Theta}) \ge \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*), \boldsymbol{\Theta} - \boldsymbol{\Theta}^*\rangle \mathbb{1}_{\mathscr{E}_i} \text{ for any } \boldsymbol{\Theta} \in \boldsymbol{\Theta}^* + \mathbb{B}(s).$$

Write $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{d_2}) = \boldsymbol{\Theta} - \boldsymbol{\Theta}^*$ with $\boldsymbol{\delta}_k$ denoting the k-th column of $\boldsymbol{\Delta}$. Under this notation, note that $\|(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)^{\mathrm{T}}\boldsymbol{x}_i\|_2^2 = \sum_{k=1}^{d_2}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}_k)^2 \le \|\boldsymbol{x}_i\|_2^2\sum_{k=1}^{d_2}\|\boldsymbol{\delta}_k\|_2^2 = \|\boldsymbol{x}_i\|_2^2\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2$. Provided

$\tau \geq 2s\max_{1\leq i\leq n}\|\boldsymbol{x}_i\|_2$, it holds

$$\|\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{x}_i\|_2 \leq \frac{\tau}{2s}\|\boldsymbol{\Delta}\|_{\mathrm{F}} \text{ for any } \boldsymbol{\Delta}\in\mathbb{B}(s), \tag{A.35}$$

which in turn implies

$$\begin{aligned} D(\boldsymbol{\Theta}) &\geq \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{x}_i\|_2^2\mathbb{1}\big(\|\boldsymbol{\varepsilon}_i\|_2\leq\tau/2\big) \\ &\geq \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2\cdot\inf_{\boldsymbol{u}\in\mathbb{S}^{d_1-1}}\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{u})^2\mathbb{1}\big(\|\boldsymbol{\varepsilon}_i\|_2\leq\tau/2\big) \\ &= \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2\cdot\lambda_{\min}\bigg(\frac{1}{n}\sum_{i=1}^{n}\chi_i\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\bigg), \end{aligned}$$

where $\chi_i=\mathbb{1}(\|\boldsymbol{\varepsilon}_i\|_2\leq\tau/2)\in\{0,1\}$. In what follows, we provide lower and upper bounds on $\lambda_{\min}(n^{-1}\sum_{i=1}^{n}\chi_i\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}})$ and $\max_{1\leq i\leq n}\|\boldsymbol{x}_i\|_2$ with high probability, respectively.

Write $\boldsymbol{\Sigma}_\tau=\mathbb{E}(\chi_i\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}})$. By Condition (C3) and Markov's inequality,

$$1\geq\mathbb{P}\big(\|\boldsymbol{\varepsilon}_i\|_2\leq\tau/2|\boldsymbol{x}_i\big)\geq 1-\bigg(\frac{2}{\tau}\bigg)^2\mathrm{tr}\big(\mathbb{E}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}}|\boldsymbol{x}_i)\big)\geq 1-\frac{4\sigma_0^2d_2}{\tau^2}\geq\frac{3}{4}$$

as long as $\tau\geq 4\sigma_0\sqrt{d_2}$, thus implying $\lambda_{\min}(\boldsymbol{\Sigma}_\tau)\geq 3c_l/4$. The sub-Gaussianity of $\boldsymbol{x}_i$ (see Condition (C2)) ensures that $\|\langle\boldsymbol{u},\chi_i\boldsymbol{x}_i\rangle\|_{\psi_2}\lesssim\nu_0\|\boldsymbol{u}\|_2$ for any $\boldsymbol{u}\in\mathbb{R}^{d_1}$, where $\|\cdot\|_{\psi_2}$ denotes the ψ_2 Orlicz norm or the sub-Gaussian norm. Following the proof of Theorem 1 by Zhivotovskiy (2024), it can be similarly shown that for any $z>0$,

$$\bigg\|\frac{1}{n}\sum_{i=1}^{n}\chi_i\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}-\boldsymbol{\Sigma}_\tau\bigg\|_2\lesssim\nu_0^2\sqrt{\frac{d_1+z}{n}}$$

with probability at least $1-e^{-z}$ whenever $n\gtrsim d_1+z$. Thus, it follows from the above analysis that with probability at least $1-e^{-z}$,

$$\lambda_{\min}\bigg(\frac{1}{n}\sum_{i=1}^{n}\chi_i\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\bigg)\geq\frac{c_l}{2}$$

as long as $\tau \geq 4\sigma_0\sqrt{d_2}$ and $n \gtrsim \nu_0^4 c_l^{-2}(d_1+z)$.

For $\max_{1\leq i\leq n}\|\boldsymbol{x}_i\|_2$, applying first Theorem 2.1 in Hsu, Kakade and Zhang (2012) to each $\|\boldsymbol{x}_i\|_2$, and then taking the union bound over $i=1,\ldots,n$, we see that for any $x>0$,

$$\mathbb{P}\left\{\max_{1\leq i\leq n}\|\boldsymbol{x}_i\|_2^2 > \nu_0^2(d_1+2\sqrt{d_1 x}+2x)\right\} \leq ne^{-x}.$$

With $x=z+\log n$, this further implies that $\max_{1\leq i\leq n}\|\boldsymbol{x}_i\|_2 \leq \nu_0\sqrt{2d_1+3z+3\log n}$ with probability at least $1-ne^{-z-\log n}=1-e^{-z}$. We thus let $\tau \geq 2\nu_0 s\sqrt{2d_1+3z+3\log n}$ so that (A.35) holds with the same probability.

Putting together the pieces, we have shown that with probability at least $1-2e^{-z}$,

$$\langle \nabla\widehat{L}_\tau(\boldsymbol{\Theta})-\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*), \boldsymbol{\Theta}-\boldsymbol{\Theta}^*\rangle \geq \frac{c_l}{2}\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 \text{ for all } \boldsymbol{\Theta}\in\boldsymbol{\Theta}^*+\mathbb{B}(s),$$

provided $\tau \geq \max\{4\sigma_0\sqrt{d_2}, 2\nu_0 s\sqrt{2d_1+3z+3\log n}\}$ and $n \gtrsim \nu_0^4 c_l^{-2}(d_1+z)$. This completes the proof. □

A.3.3 Proof of Theorem 1.3.4

Similarly to the proof of Theorem 1.3.1, we employ the localized analysis via proof by contradiction. Recall that $d=d_1+d_2$. In view of Proposition 1.3.5, we choose $\tau \asymp \sigma_0\sqrt{n/(z+\log d)}$ and $\lambda \asymp \sigma_0\sqrt{d(z+\log d)/n}$ so that $\lambda \geq 2\|\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2$ with probability at least $1-e^{-z}$. Choose

$$s \asymp \frac{\tau}{\nu_0(d+z+\log n)^{1/2}} \asymp \sigma_0\sqrt{\frac{n}{(d+z+\log n)(z+\log d)}}$$

so that (1.20) is satisfied. Moreover, Proposition 1.3.6 implies that with probability at least $1-2e^{-z}$

$$\langle \nabla\widehat{L}_\tau(\boldsymbol{\Theta})-\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*), \boldsymbol{\Theta}-\boldsymbol{\Theta}^*\rangle \geq \frac{c_l}{2}\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 \text{ for all } \boldsymbol{\Theta}\in\boldsymbol{\Theta}^*+\mathbb{B}(s).$$

under the sample complexity $n \gtrsim d(z+\log d)$. Conditioning on the above good events, we see from the proof of Theorem 1.3.1 that the robust matrix estimator $\widehat{\mathbf{\Theta}}_{\tau,\lambda}$ satisfies the error bounds

$$\|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_{\mathrm{F}} \lesssim \sigma_0^{1-q/2}\sqrt{\rho}\left\{\frac{d(z+\log d)}{n}\right\}^{\frac{1}{2}-\frac{q}{4}}$$

and

$$\|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_{*} \lesssim \sigma_0^{1-q}\rho\left\{\frac{d(z+\log d)}{n}\right\}^{\frac{1-q}{2}}$$

as long as

$$\sigma_0\sqrt{\frac{n}{(d+z+\log n)(z+\log d)}} \asymp s \gtrsim \sqrt{\rho}\,\lambda^{1-q/2} \asymp \sigma_0^{1-q/2}\sqrt{\rho}\left\{\frac{d(z+\log d)}{n}\right\}^{1/2-q/4}.$$

This requirement holds under the sample complexity

$$n \gtrsim (\rho/\sigma_0^q)^{2/(4-q)}(d+z+\log n)(z+\log d),$$

thus completing the proof. □

Appendix B

Supplementary Material for Chapter 2

B.1 Extension to other differential private mechanisms

In this section, we briefly introduce the extension of our construction for ε-GDP robust mean estimators to incorporate other DP mechanisms.

Given a pair of positive privacy parameters ε and δ, we first construct an (ε,δ)-DP estimator. The following two lemmas serve as pivotal tools to construct (ε,δ)-DP estimators, which constitute counterparts to Lemma 2.3.1 and Lemma 2.3.2, respectively, in the context of (ε,δ)-DP mechanism. Recall the definition of the sensitivity of a statistic $\boldsymbol{h}\in\mathbb{R}^d$ in (2.20).

Lemma B.1.1. (Gaussian mechanism Dwork and Roth (2014)) Define the Gaussian mechanism that operates on a statistic $\boldsymbol{h}\in\mathbb{R}^d$ as

$$M(\boldsymbol{X})=\boldsymbol{h}(\boldsymbol{X})+\frac{\text{sens}(\boldsymbol{h})\sqrt{2\log(2/\delta)}}{\varepsilon}\boldsymbol{g},$$

where $\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\mathbf{I}_d)$. Then, the Gaussian mechanism M is (ε,δ)-DP.

Lemma B.1.2. (Composition of DP Dwork *et al.* (2006b)) Let $M_1:\mathscr{X}^n\to\mathscr{Y}_1$ be the first mechanism and $M_t:\mathscr{X}^n\times\mathscr{Y}_1\times\cdots\times\mathscr{Y}_{t-1}\to\mathscr{Y}_t$ be the t-th mechanism for $t=2,\ldots,k$. We define the k-fold composed mechanism $M:\mathscr{X}^n\to\mathscr{Y}_1\times\cdots\times\mathscr{Y}_k$ as $M(\boldsymbol{X})=(y_1,y_2,\ldots,y_k)$ where $y_1=M_1(\boldsymbol{X})$ and $y_t=M_t(\boldsymbol{X},y_1,\ldots,y_{t-1})$ for $t=2,\ldots,k$. If M_1 is (ε_1,δ_1)-DP and $M_t(\cdot,y_1,\ldots,y_{t-1})$ is (ε_t,δ_t)-DP for any $y_1\in\mathscr{Y}_1,\ldots,y_{t-1}\in\mathscr{Y}_{t-1}$, then the k-fold composed mechanism M is

$(\sum_{t=1}^{k}\varepsilon_t, \sum_{t=1}^{k}\delta_t)$-DP.

Next, we introduce a robust (ε,δ)-DP estimator. Given initial estimator $\boldsymbol{\mu}_{\mathrm{DP}}^{(0)}$ and predetermined number of iterations T, define

$$\boldsymbol{\mu}_{\mathrm{DP}}^{(t+1)} = \boldsymbol{\mu}_{\mathrm{DP}}^{(t)} + \frac{\eta_0}{n}\sum_{i=1}^{n}\frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{DP}}^{(t)}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{DP}}^{(t)}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{DP}}^{(t)}) + 2\tau T\frac{\eta_0\sqrt{2\log(2/\delta)}}{\varepsilon}\boldsymbol{g}_t \tag{B.1}$$

for $t = 0, 1, \ldots, T-1$, where $\eta_0 > 0$ is the step size, $\{\boldsymbol{g}_t\}_{t=0}^{T-1}$ is a sequence of independent standard d-variate normal random vectors. The final private estimator is denoted by $\boldsymbol{\mu}_{\mathrm{DP}}^{(T)}$. Given $r_0 > 0$ and $\chi \in (0,1)$, recall the event $\mathscr{E}_1(r_0,\chi)$ defined in (2.22). In the following, we present an oracle-type concentration bound for the (ε,δ)-DP estimator $\boldsymbol{\mu}_{\mathrm{DP}}^{(T)}$ around the Huber estimator $\widehat{\boldsymbol{\mu}}$ conditioning on the event $\mathscr{E}_1$. Building upon Lemma B.1.1 and B.1.2, the proof is almost identical to the proof of Theorem 2.3.1, so we omit the proof for brevity.

Proposition B.1.1. For the given step size $\eta_0 \in (0,1]$ and the number of iterations $T \geq 1$, the private estimator $\boldsymbol{\mu}_{\mathrm{DP}}^{(T)}$ obtained from (B.1) is (ε,δ)-DP. Furthermore, assume that the initial estimate satisfies $\|\boldsymbol{\mu}_{\mathrm{DP}}^{(0)} - \boldsymbol{\mu}\|_2 \leq r_0$ for some $r_0 > 0$. Let $\chi \in (0,1), z > 0$, and define

$$\mathrm{r}_{\mathrm{opt}}^2 = (1-\rho)^T r_0^2 \;\text{ and }\; \mathrm{r}_{\mathrm{p}}^2 = \eta_0 T^2\{\eta_0 + (1-\chi)^{-1}\}\left(\frac{d}{\rho} + z\right)\left(\frac{\tau\sqrt{\log(2/\delta)}}{\varepsilon n}\right)^2,$$

where $\rho = (1-\chi)^2\eta_0^2$. Assume that the sample size satisfies

$$n \gtrsim T\tau\frac{(\sqrt{d} + \sqrt{z+\log T})\sqrt{\log(2/\delta)}}{(1-\chi)\varepsilon r_0}.$$

Then, conditioning on the event $\mathscr{E}_1 = \mathscr{E}_1(r_0,\chi)$, $\boldsymbol{\mu}_{\mathrm{DP}}^{(T)}$ satisfies

$$\|\boldsymbol{\mu}_{\mathrm{DP}}^{(T)} - \widehat{\boldsymbol{\mu}}\|_2 \lesssim \mathrm{r}_{\mathrm{opt}} + \mathrm{r}_{\mathrm{p}}$$

with probability (over $\{\boldsymbol{g}_t\}_{t=0}^{T-1}$) at least $1 - 2e^{-z}$.

We remark that the only difference between the deviation bound of the (ε, δ)-DP estimator and ε-GDP is the dependence on T. The deviation bound of ε-GDP scales with $\sqrt{T}$, whereas that of (ε, δ)-GDP scales with T. However, when we choose $T \asymp \log n$ as in Corollary 2.3.1, the non-asymptotic bounds for both estimators become almost the same up to logarithmic terms. Nonetheless, it is important to note that, unlike the ε-GDP estimator, the privacy of (ε, δ)-DP estimator is not tightly characterized.

We next introduce another variant of differential privacy known as zero-concentrated differential privacy (zCDP). To begin with, we recall the definition of Rényi divergence.

Definition B.1.1. For $\alpha > 1$, the Rényi divergence of order α or α-Rényi divergence of a distribution P from a distribution Q is defined to be

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \left(\mathbb{E}_{X \sim Q} \left[\left\{ \frac{P(X)}{Q(X)} \right\}^\alpha \right] \right).$$

Now, we are ready to introduce the definition of zCDP.

Definition B.1.2. (Bun and Steinke (2016)) A randomized algorithm $M : \mathscr{X}^n \to \mathscr{Y}$ is said to be ε-zero-concentrated differential private (ε-zCDP) for $\varepsilon > 0$ if for any neighboring datasets $\boldsymbol{X}, \boldsymbol{X}' \in \mathscr{X}^n$, and any $\alpha > 1$, the α-Rényi divergence between $M(\boldsymbol{X})$ and $M(\boldsymbol{X}')$ satisfies

$$D_\alpha(M(\boldsymbol{X})\|M(\boldsymbol{X}')) \leq \varepsilon\alpha.$$

The following two lemmas provide essential tools for the construction of zCDP estimators.

Lemma B.1.3. (Gaussian mechanism (Bun and Steinke, 2016)) Define the Gaussian mechanism that operates on a statistic $\boldsymbol{h} \in \mathbb{R}^d$ as

$$M(\boldsymbol{X}) = \boldsymbol{h}(\boldsymbol{X}) + \frac{\mathrm{sens}(h)}{\sqrt{2\varepsilon}} \boldsymbol{g},$$

where $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Then, the mechanism M is ε-zCDP.

Lemma B.1.4. (Composition of zCDP Bun and Steinke (2016)) Let $M_1 : \mathscr{X}^n \to \mathscr{Y}_1$ be the first mechanism and $M_t : \mathscr{X}^n \times \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_{t-1} \to \mathscr{Y}_t$ be the t-th mechanism for $t = 2, \ldots, k$. Define the k-fold composed mechanism $M : \mathscr{X}^n \to \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_k$ as $M(\boldsymbol{X}) = (y_1, y_2, \ldots, y_k)$ where $y_1 = M_1(\boldsymbol{X})$ and $y_t = M_t(\boldsymbol{X}, y_1, \ldots, y_{t-1})$ for $t = 2, \ldots, k$. If M_1 is ε_1-zCDP and $M_t(\cdot, y_1, \ldots, y_{t-1})$ is ε_t-DP for any $y_1 \in \mathscr{Y}_1, \ldots, y_{t-1} \in \mathscr{Y}_{t-1}$, then the k-fold composed mechanism M is $(\sum_{t=1}^k \varepsilon_t)$-zCDP.

We are now prepared to outline the procedure for constructing a robust ε-zCDP. Given an initial estimator $\boldsymbol{\mu}_{\text{zCDP}}^{(0)}$ and predetermined number of iterations T, define

$$\boldsymbol{\mu}_{\text{zCDP}}^{(t+1)} = \boldsymbol{\mu}_{\text{zCDP}}^{(t)} + \frac{\eta_0}{n} \sum_{i=1}^{n} \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}_{\text{zCDP}}^{(t)}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}_{\text{zCDP}}^{(t)}\|_2} (\boldsymbol{x}_i - \boldsymbol{\mu}_{\text{zCDP}}^{(t)}) + 2\tau T \frac{\eta_0}{\sqrt{2\varepsilon}} \boldsymbol{g}_t \tag{B.2}$$

for $t = 0, 1, \ldots, T-1$, where $\eta_0 > 0$ is the step size, $\{\boldsymbol{g}_t\}_{t=0}^{T-1}$ is a sequence of independent standard d-variate normal random vectors and τ is the robustification parameter. The final private estimator is denoted by $\boldsymbol{\mu}_{\text{zCDP}}^{(T)}$. The following proposition gives a deviation bound for the ε-zCDP estimator $\boldsymbol{\mu}_{\text{zCDP}}^{(T)}$ around $\widehat{\boldsymbol{\mu}}$ conditioning on $\mathscr{E}_1$. Combined with Lemma B.1.3 and Lemma B.1.4, the proof of this proposition closely follows the same argument as the proof of Theorem 2.3.1, so is omitted for brevity.

Proposition B.1.2. For the given step size $\eta_0 \in (0, 1]$ and the number of iterations $T \geq 1$, the estimator $\boldsymbol{\mu}_{\text{zCDP}}^{(T)}$ obtained from (B.2) is ε-zCDP. Furthermore, assume that the initial estimate satisfies $\|\boldsymbol{\mu}_{\text{zCDP}}^{(0)} - \boldsymbol{\mu}\|_2 \leq r_0$ for some $r_0 > 0$. Let $\chi \in (0, 1), z > 0$, and define

$$\mathrm{r}_{\text{opt}}^2 = (1-\rho)^T r_0^2 \text{ and } \mathrm{r}_{\text{p}}^2 = \eta_0 T^2 \{\eta_0 + (1-\chi)^{-1}\} \left(\frac{d}{\rho} + z\right) \left(\frac{\tau}{\sqrt{\varepsilon} n}\right)^2,$$

where $\rho = (1-\chi)^2\eta_0^2$. Assume that the sample size satisfies

$$n \gtrsim T\tau \frac{\sqrt{d}+\sqrt{z+\log T}}{(1-\chi)\sqrt{\varepsilon}r_0}.$$

Then, conditioning on the event $\mathscr{E}_1 = \mathscr{E}_1(r_0,\chi)$, $\boldsymbol{\mu}_{\mathrm{zCDP}}^{(T)}$ satisfies

$$\|\boldsymbol{\mu}_{\mathrm{zCDP}}^{(T)} - \widehat{\boldsymbol{\mu}}\|_2 \lesssim \mathrm{r}_{\mathrm{opt}} + \mathrm{r}_{\mathrm{p}}$$

with probability (over $\{\boldsymbol{g}_t\}_{t=0}^{T-1}$) at least $1-2e^{-z}$.

It should be noted that the resulting ε-zCDP estimator, in contrast to the (ε,δ)-DP constructed by (B.1), has a tight privacy characterization. However, the concept of zCDP no longer encompasses hypothesis testing interpretations.

B.2 Details of implementation

For the implementation of ε-GDP robust mean estimators and CIs, we first need to choose an appropriate robustification parameter τ. Motivated by the bound (2.28) in Corollary 2.3.1, we take

$$\tau \asymp m_2^{1/2}\left\{\frac{\varepsilon n}{\sqrt{(d+\log n)\log n}}\right\}^{1/2}, \tag{B.3}$$

where m_2 is defined as $m_2 = \mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2$. Motivated by Liu et al. (2023), we use the histogram learner algorithm (Karwa and Vadhan, 2018), summarized in Algorithm 3, to estimate m_2. In detail, for a prespecified number of partitions M, we first partition the set $\{\|\boldsymbol{x}_{2i-1}-\boldsymbol{x}_{2i}\|_2/2\}_{i=1}^{\lfloor n/2\rfloor}$ into M batches, compute the median for each partition, and use the private histogram learner algorithm with geometrically increasing bin sizes to get a private estimator. We describe the algorithm in Algorithm 4. Here, we use the median for a robust estimation of $m_2 = \mathbb{E}\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2/2$.

Algorithm 3. Private histogram learner (HL) algorithm

Input: privacy parameters (ε, δ), dataset $\{w_m\}_{m=1}^M \in \Omega^M$ for some domain Ω, collection of disjoint bins $\{B_k\}_{k=1}^K$ defined on Ω with $K \in \mathbb{N} \cup \{\infty\}$

for $k = 1, 2, \ldots, K$ **do**
  $\widehat{p}_k = \sum_{m=1}^M \mathbb{I}(w_m \in B_k)/M$
  if $\widehat{p}_k = 0$ **then**
    $\widetilde{p}_k = 0$
  else
    Generate a Laplace random variable Z_k with mean 0 and scale $2/(\varepsilon M)$;
    $\widetilde{p}_k = \widehat{p}_k + Z_k$;
    if $\widehat{p}_k + Z_k < 2\log(2/\delta)/(\varepsilon M) + (1/M)$ **then**
      $\widetilde{p}_k = 0$;
    end if
  end if
end for

Output: $\{\widetilde{p}_k\}_{k=1}^K = \mathrm{HL}(\{w_m\}_{m=1}^M, (\varepsilon, \delta))$

By combining Lemma 2.3 in Karwa and Vadhan (2017) and Corollary 1 in Dong, Roth and Su (2022), the specific choice of δ implies that Algorithm 4 gives an ε-GDP estimator of m_2 for a given $\varepsilon > 0$. Based on this, we propose a heuristic data-driven approach to construct an ε-GDP robust estimator of $\boldsymbol{\mu}$.

We initialize the initial estimate $\boldsymbol{\mu}^{(0)} = \mathbf{0} \in \mathbb{R}^d$, the step size $\eta_0 = 1$, and set the number of iterations $T = \lfloor \log n \rfloor$. We first run Algorithm 4 with $M = \lfloor \sqrt{n}/2 \rfloor$ and the privacy parameter $\varepsilon/\sqrt{T+1}$ to get an $(\varepsilon/\sqrt{T+1})$-GDP estimator denoted as $\widehat{m}_2$ for m_2. Subsequently, at iteration $t = 1, 2, \ldots, T-1$, we compute

$$\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} + \frac{\eta_0}{n} \sum_{i=1}^n \frac{\psi_{\widehat{\tau}}(\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}\|_2} (\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}) + 2(T+1)^{1/2} \widehat{\tau} \frac{\eta_0}{\varepsilon n} \boldsymbol{g}_t.$$

Here, $\{\boldsymbol{g}_t\}_{t=0}^{T-1}$ is a sequence of independent standard d-variate normal random vectors, and $\widehat{\tau}$ is defined as in (B.3) with m_2 replaced by $\widehat{m}_2$. Following a similar line of arguments in the proof of Proposition 2.3.1, it can be demonstrated that the final private estimator $\boldsymbol{\mu}^{(T)}$ is ε-GDP.

To construct a GDP CI for $\boldsymbol{\mu}$, we follow the construction outlined in (2.33). However, for computational efficiency, we replace $\widehat{\Sigma}_{\xi,\varepsilon}$ with the following perturbed plug-in covariance

Algorithm 4. Private and robust estimator for m_2

Input: dataset $\{\boldsymbol{x}_i\}_{i=1}^n$, privacy parameter ε, the number of partitions M

1: Partition $\{\|\boldsymbol{x}_{2i-1}-\boldsymbol{x}_{2i}\|_2/2\}_{i=1}^{\lfloor n/2\rfloor}$ into M partitions of equal size;
2: For $1\le m\le M$, compute w_m to be the median value of m-th partition;
3: Partition $[0,\infty)$ into geometrically increasing intervals

$$[0,\infty)=\sum_{k=-\infty}^{\infty}[2^k,2^{k+1})\cup[0,0];$$

4: Run (ε,δ)-DP private histogram leaner $\mathrm{HL}(\{w_m\}_{m=1}^M,(\varepsilon,\delta))$ with $\delta=\Phi(-1+\varepsilon/2)-e^{\varepsilon}\Phi(-1-\varepsilon/2)$;
5: Let $k'\in\mathrm{argmax}_{-\infty\le k\le\infty}\widetilde{p}_k$;

Output: $2^{k'}$

estimator:

$$\widetilde{\Sigma}_{\xi,\varepsilon}:=\operatorname*{argmin}_{\boldsymbol{H}\succeq\zeta\boldsymbol{I}}\left\|\boldsymbol{H}-\left\{\frac{1}{n}\sum_{i=1}^n\frac{\psi_\xi(\|\boldsymbol{x}_i-\boldsymbol{\mu}^{(T)}\|_2^2)}{\|\boldsymbol{x}_i-\boldsymbol{\mu}^{(T)}\|_2^2}(\boldsymbol{x}_i-\boldsymbol{\mu}^{(T)})(\boldsymbol{x}_i-\boldsymbol{\mu}^{(T)})^{\mathrm{T}}+\frac{2\xi}{\varepsilon n}\boldsymbol{E}\right\}\right\|_2,$$

where $\{\boldsymbol{H}:\boldsymbol{H}\succeq\zeta\boldsymbol{I}\}$ is a cone of positive definite matrices, whose minimal eigenvalues are not smaller than a prespecified positive number ζ. Here, $\boldsymbol{E}\in\mathbb{R}^{d\times d}$ is a symmetric random matrix where upper-triangular and diagonal entries are i.i.d. $\mathcal{N}(0,1)$. The parameter $\xi>0$ represents the robustification parameter, and we use $\widehat{\xi}=10\widehat{m}_2\sqrt{n/\log(nd)}$ with $\widehat{m}_2$ computed during the estimation of $\boldsymbol{\mu}^{(T)}$. We remark that the mechanism $(\widehat{m}_2,\boldsymbol{\mu}^{(T)})$ is ε-GDP. Combining this with Lemma 2.3.2, we have that $(\boldsymbol{\mu}^{(T)},\widetilde{\Sigma}_{\widehat{\xi},\varepsilon})$ is a $(\sqrt{2}\varepsilon)$-GDP mechanism. Therefore, the $100(1-\alpha)\%$ (approximate) confidence interval

$$\left[\langle\boldsymbol{u},\boldsymbol{\mu}^{(T)}\rangle-z_{\alpha/2}\frac{(\boldsymbol{u}^{\mathrm{T}}\widetilde{\Sigma}_{\widehat{\xi},\varepsilon}\boldsymbol{u})^{1/2}}{\sqrt{n}},\langle\boldsymbol{u},\boldsymbol{\mu}^{(T)}\rangle+z_{\alpha/2}\frac{(\boldsymbol{u}^{\mathrm{T}}\widetilde{\Sigma}_{\widehat{\xi},\varepsilon}\boldsymbol{u})^{1/2}}{\sqrt{n}}\right]$$

is $(\sqrt{2}\varepsilon)$-GDP.

B.3 Proofs in Section 2.2

B.3.1 Supporting lemmas

We first provide several technical lemmas regarding the gradient and Hessian of the empirical loss $\widehat{\mathscr{L}_\tau}(\cdot)$, given by

$$\begin{aligned}
\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_\tau(\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2)}{\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2}(\boldsymbol{x}_i-\boldsymbol{\theta}) \\
\text{and}\ \ \nabla^2\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{I}_d-\frac{(\boldsymbol{x}_i-\boldsymbol{\theta})(\boldsymbol{x}_i-\boldsymbol{\theta})^{\mathrm{T}}}{\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2^2}\right\}\frac{\psi(\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2/\tau)}{\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2/\tau} \\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\frac{(\boldsymbol{x}_i-\boldsymbol{\theta})(\boldsymbol{x}_i-\boldsymbol{\theta})^{\mathrm{T}}}{\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2^2}\psi'(\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2/\tau).
\end{aligned}$$

The concavity of $\psi(\cdot)$ implies $\psi(u)\geq u\psi'(u)$ for any $u\geq 0$, from which it follows that for any $\boldsymbol{u}\in\mathbb{S}^{d-1}$,

$$\frac{1}{n}\sum_{i=1}^{n}\psi'(\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2/\tau)\leq\boldsymbol{u}^{\mathrm{T}}\nabla^2\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta})\boldsymbol{u}\leq\frac{1}{n}\sum_{i=1}^{n}\frac{\tau\psi(\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2/\tau)}{\|\boldsymbol{x}_i-\boldsymbol{\theta}\|_2}. \tag{B.4}$$

We denote the population loss $\mathscr{L}_\tau(\boldsymbol{\theta})=\mathbb{E}\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta})$. For any $r>0$, define the local ball around the true mean vector $\boldsymbol{\mu}=\mathbb{E}(\boldsymbol{x})$ as $\Theta(r)=\{\boldsymbol{\theta}\in\mathbb{R}^d:\|\boldsymbol{\theta}-\boldsymbol{\mu}\|_2\leq r\}$. The following lemma demonstrates the local strong convexity of the empirical loss function $\widehat{\mathscr{L}_\tau}(\cdot)$.

Lemma B.3.1. Let $r>0$ and $\tau=\gamma+r$ with $\gamma>0$ and $z>0$. Then, with probability at least $1-e^{-z}$,

$$1-\mathbb{P}\big(\|\boldsymbol{x}-\boldsymbol{\mu}\|_2>\gamma\big)-\sqrt{\frac{z}{2n}}\leq\boldsymbol{u}^{\mathrm{T}}\nabla^2\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta})\boldsymbol{u}\leq 1$$

holds uniformly over $\boldsymbol{\theta}\in\Theta(r)$ and $\boldsymbol{u}\in\mathbb{S}^{d-1}$.

Proof of Lemma B.3.1. For the Huber loss, $\psi(u)=\mathrm{sign}(u)\min(|u|,1)$ is 1-Lipschitz continuous and differentiable except at ±1. By (B.4), the sample Hessian $\nabla^2\widehat{\mathscr{L}_\tau}(\cdot)$ satisfies for any $\boldsymbol{\theta}\in\mathbb{R}^d$

and $\boldsymbol{u} \in \mathbb{S}^{d-1}$ that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big(\|\boldsymbol{x}_i - \boldsymbol{\theta}\|_2 \le \tau\big) \le \boldsymbol{u}^{\mathrm{T}} \nabla^2 \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) \boldsymbol{u} \le 1.$$

For all $\boldsymbol{\theta}$ in the local region $\Theta(r)$, $\{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2 \le \gamma\} \subseteq \{\|\boldsymbol{x}_i - \boldsymbol{\theta}\|_2 \le \tau\}$ and hence

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big(\|\boldsymbol{x}_i - \boldsymbol{\theta}\|_2 \le \tau\big) \ge \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2 \le \gamma\big) \\
&= 1 - \mathbb{P}\big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2 > \gamma\big) - \frac{1}{n}\sum_{i=1}^{n} \big\{\mathbb{1}\big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2 > \gamma\big) - \mathbb{P}\big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2 > \gamma\big)\big\}.
\end{aligned}$$

For the last term, applying Hoeffding's inequality yields that, with probability at least $1 - e^{-z}$,

$$\frac{1}{n}\sum_{i=1}^{n} \big\{\mathbb{1}\big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2 > \gamma\big) - \mathbb{P}\big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2 > \gamma\big)\big\} \le \sqrt{\frac{z}{2n}}.$$

Putting together the pieces proves the claimed bound. □

We next establish an upper bound of the sample gradient $\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu})$.

Lemma B.3.2. For any $\tau > 0$, the sample gradient $\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu})$ satisfies the bound

$$\|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu})\|_2 \le 2\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \sqrt{\frac{2\|\Sigma\|_2 z}{n}} + \frac{4\tau z}{3n} + b_\tau$$

with probability at least $1 - e^{-z}$ for any $z \ge 0$, where $b_\tau \le \tau^{-1}\sqrt{\bar{\lambda}\,\mathrm{tr}(\Sigma)}$ is defined in (2.3).

Proof of Lemma B.3.2. To begin with, note that

$$\|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu})\|_2 = \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^{n} \underbrace{\tau\psi\big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2/\tau\big)\langle \boldsymbol{u}, \boldsymbol{x}_i - \boldsymbol{\mu}\rangle / \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}_{=: f_{\boldsymbol{u}}(\boldsymbol{x}_i)},$$

where the function $f_{\boldsymbol{u}}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is such that $|f_{\boldsymbol{u}}(\cdot)| \le \tau$ and $\mathbb{E} f_{\boldsymbol{u}}^2(\boldsymbol{x}_i) \le \mathbb{E}\langle \boldsymbol{u}, \boldsymbol{x}_i - \boldsymbol{\mu}\rangle^2 \le \|\Sigma\|_2$.

Moreover, using the fact that $\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{x}_i - \boldsymbol{\mu}\rangle = 0$ we obtain

$$\begin{aligned}\mathbb{E}\{f_{\boldsymbol{u}}(\boldsymbol{x}_i)\} &= \mathbb{E}\langle \boldsymbol{u}, \boldsymbol{x}_i - \boldsymbol{\mu}\rangle \frac{\min(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2, \tau) - \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2} \\ &= \mathbb{E}\langle \boldsymbol{u}, \boldsymbol{x}_i - \boldsymbol{\mu}\rangle \frac{\tau - \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2} \mathbb{1}\big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2 > \tau\big),\end{aligned}$$

which in turn implies

$$\begin{aligned}|\mathbb{E} f_{\boldsymbol{u}}(\boldsymbol{x}_i)| \le \tau^{-1}\mathbb{E}\big(|\langle \boldsymbol{x}_i - \boldsymbol{\mu}, \boldsymbol{u}\rangle| \cdot \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2\big) &\le \tau^{-1}\sqrt{\mathbb{E}\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^2 \cdot \mathbb{E}\langle \boldsymbol{x}_i - \boldsymbol{\mu}, \boldsymbol{u}\rangle^2} \\ &= \frac{\sqrt{\operatorname{tr}(\Sigma)\|\Sigma\|_2}}{\tau}. \qquad \text{(B.5)}\end{aligned}$$

For the supremum $\Delta := \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} (1/n)\sum_{i=1}^n (1-\mathbb{E}) f_{\boldsymbol{u}}(\boldsymbol{x}_i)$, it follows from Talagrand's inequality (see, e.g., Theorem 7.3 in Bousquet (2003)) that, with probability at least $1 - e^{-z}$,

$$\Delta \le 2\mathbb{E}\Delta + \sqrt{\frac{2\|\Sigma\|_2 z}{n}} + \frac{4\tau z}{3n}.$$

For $\mathbb{E}\Delta$, by the Cauchy-Schwarz inequality we have

$$\begin{aligned}&\mathbb{E}\left\{\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E}) f_{\boldsymbol{u}}(\boldsymbol{x}_i)\right\} \\ &\quad = \mathbb{E}\|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}) - \mathbb{E}\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu})\|_2 \le \sqrt{\mathbb{E}\|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}) - \nabla\mathscr{L}_\tau(\boldsymbol{\mu})\|_2^2} \\ &\quad \le \left\{\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\psi_\tau^2\big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2\big)\right\}^{1/2} = \left\{\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\min\big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^2, \tau^2\big)\right\}^{1/2} \le \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}.\end{aligned}$$

Note that $\|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu})\|_2 \le \Delta + b_\tau$, where

$$b_\tau = \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \mathbb{E} f_{\boldsymbol{u}}(\boldsymbol{x}_i) = \left\|\mathbb{E}\left\{\frac{(\boldsymbol{x} - \boldsymbol{\mu})\psi_\tau(\|\boldsymbol{x} - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x} - \boldsymbol{\mu}\|_2}\right\}\right\|_2 \le \tau^{-1}\sqrt{\|\Sigma\|_2 \operatorname{tr}(\Sigma)}$$

due to (B.5). Combining this with the concentration bound and the bound of $\mathbb{E}\Delta$ proves the

claimed result. □

B.3.2 Proof of Lemma 2.2.1

By the definition of b_τ, we have

$$b_\tau = \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \mathbb{E}\left\{ \frac{\psi_\tau(\|\boldsymbol{x}-\boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2}\langle \boldsymbol{u}, \boldsymbol{x}-\boldsymbol{\mu}\rangle \right\}.$$

For any $\boldsymbol{u}\in\mathbb{S}^{d-1}$ fixed, following the same argument as in the proof of Lemma B.3.2 we obtain

$$\mathbb{E}\left\{ \frac{\psi_\tau(\|\boldsymbol{x}-\boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2}\langle \boldsymbol{x}-\boldsymbol{\mu}, \boldsymbol{u}\rangle \right\} = \mathbb{E}\langle \boldsymbol{u}, \boldsymbol{x}-\boldsymbol{\mu}\rangle \frac{\tau - \|\boldsymbol{x}-\boldsymbol{\mu}\|_2}{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2}\mathbb{1}(\|\boldsymbol{x}-\boldsymbol{\mu}\|_2 > \tau).$$

Since $m_q = \mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^q < \infty$, this further implies

$$\begin{aligned}
\left| \mathbb{E}\left\{ \frac{\psi_\tau(\|\boldsymbol{x}-\boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2}\langle \boldsymbol{x}-\boldsymbol{\mu}, \boldsymbol{u}\rangle \right\} \right| &\le \mathbb{E}|\langle \boldsymbol{u}, \boldsymbol{x}-\boldsymbol{\mu}\rangle| \mathbb{1}(\|\boldsymbol{x}-\boldsymbol{\mu}\|_2 > \tau) \\
&\le \frac{1}{\tau^{q-1}}\mathbb{E}|\langle \boldsymbol{u}, \boldsymbol{x}-\boldsymbol{\mu}\rangle| \cdot \|\boldsymbol{x}-\boldsymbol{\mu}\|_2^{q-1} \\
&\le \nu_q^{1/q} \frac{(\boldsymbol{u}^{\mathrm{T}}\Sigma\boldsymbol{u})^{1/2} m_q^{(q-1)/q}}{\tau^{q-1}},
\end{aligned}$$

where the last inequality follows from Hölder's inequality. Since $\boldsymbol{u}^{\mathrm{T}}\Sigma\boldsymbol{u} \le \bar{\lambda}$ uniformly over all $\boldsymbol{u}\in\mathbb{S}^{d-1}$, this proves the first inequality of Lemma 2.2.1. The second inequality of the claim easily follows from the fact that $|\langle \boldsymbol{u}, \boldsymbol{x}-\boldsymbol{\mu}\rangle| \le \|\boldsymbol{x}-\boldsymbol{\mu}\|_2$, which completes the proof.

□

B.3.3 Proof of Theorem 2.2.3

Throughout the proof, let (n,τ) satisfy $n \gtrsim \mathrm{r}(\Sigma) + \log n$ and $\tau \gtrsim m_q^{1/q}(n/\log n)^{1/(2q)}$. Since $m_q^{1/q} \ge \mathrm{tr}(\Sigma)^{1/2}$, we have $\tau \gtrsim \mathrm{tr}(\Sigma)^{1/2}$ and $m_q/\tau^q \lesssim \sqrt{(\log n)/n}$. Therefore, applying

(2.2) and (2.8) with $z = \log n$ yields that with probability at least $1 - 3n^{-1}$,

$$\|\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\|_2 \lesssim \sqrt{\frac{\mathrm{tr}(\Sigma) + \bar{\lambda}\log n}{n}} + b_\tau + \tau\frac{\log n}{n}$$

and

$$\begin{aligned}
&\left|\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle - \frac{1}{n}\sum_{i=1}^n \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}\langle \boldsymbol{u}, \boldsymbol{x}_i - \boldsymbol{\mu}\rangle\right| \\
&\qquad \lesssim \|\boldsymbol{u}\|_2 \left\{ \frac{\sqrt{\mathrm{tr}(\Sigma)\log n} + \bar{\lambda}^{1/2}\log n}{n} + \left(b_\tau + \tau\frac{\log n}{n}\right)\sqrt{\frac{\log n}{n}} \right\}
\end{aligned}$$

uniformly over all $\boldsymbol{u} \in \mathbb{R}^d$. For each $\boldsymbol{u} \in \mathbb{R}^d$, define independent random variables

$$S_{i,\boldsymbol{u}} = \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}\langle \boldsymbol{u}, \boldsymbol{x}_i - \boldsymbol{\mu}\rangle, \quad i = 1, \ldots, n.$$

The definition of $S_{i,\boldsymbol{u}}$ is similar to that of $f_{\boldsymbol{u}}(\boldsymbol{x}_i)$ in the proof of Lemma B.3.2, except that we allow any $\boldsymbol{u} \in \mathbb{R}^d$ here while $\boldsymbol{u} \in \mathbb{S}^{d-1}$ in $f_{\boldsymbol{u}}(\boldsymbol{x}_i)$. From the proof of Lemma 2.2.1, we bound the mean of $S_{i,\boldsymbol{u}}$ as

$$|\mathbb{E}S_{i,\boldsymbol{u}}| \le \nu_q^{1/q} m_q^{1-1/q} \frac{\|\boldsymbol{u}\|_\Sigma}{\tau^{q-1}} \le \frac{m_q}{\tau^{q-1}}\|\boldsymbol{u}\|_2. \tag{B.6}$$

With the above notation, we have

$$\begin{aligned}
&\left|\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle - \frac{1}{\sqrt{n}}\sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}})\right| \le \|\boldsymbol{u}\|_2 \cdot R_{n,\tau} \\
&\text{with } \; R_{n,\tau} \asymp \frac{\sqrt{\mathrm{tr}(\Sigma)\log n} + \bar{\lambda}^{1/2}\log n}{\sqrt{n}} + \frac{\tau(\log n)^{3/2}}{n} + \frac{m_q\sqrt{n}}{\tau^{q-1}}
\end{aligned} \tag{B.7}$$

with probability at least $1 - 3n^{-1}$ as long as $n \gtrsim \mathrm{r}(\Sigma) + \log n$.

To establish the Gaussian approximation for the centered partial sum $\sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}})$,

we need to control the second and third moments of $S_{i,\boldsymbol{u}}$. For the second moment, note that

$$\mathbb{E}(S_{i,\boldsymbol{u}}^2) = \|\boldsymbol{u}\|_\Sigma^2 - \underbrace{\mathbb{E}\Big\{ \frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^2 - \tau^2}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^2} \boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{u} \cdot \mathbb{1}(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2 \geq \tau) \Big\}}_{=:\mathrm{I}}.$$

Analogous to (B.6), we have

$$0 \leq \mathrm{I} \leq \frac{1}{\tau^{q-2}} \mathbb{E}\{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^{q-2} \cdot \boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{u}\} \leq \nu_q^{2/q} m_q^{1-2/q} \frac{\|\boldsymbol{u}\|_\Sigma^2}{\tau^{q-2}},$$

from which it follows that

$$1 - \nu_q^{2/q} \frac{m_q^{1-2/q}}{\tau^{q-2}} \leq \frac{\mathbb{E}(S_{i,\boldsymbol{u}}^2)}{\|\boldsymbol{u}\|_\Sigma^2} \leq 1 \;\text{ and }\; (\mathbb{E}S_{i,\boldsymbol{u}})^2 \leq \nu_q^{2/q} m_q^{2-2/q} \frac{\|\boldsymbol{u}\|_\Sigma^2}{\tau^{2q-2}}.$$

Provided $\tau \gtrsim \nu_q^{\frac{2}{q(q-2)}} m_q^{1/q}$, this implies

$$\delta_\tau := \left| \frac{\mathrm{var}(S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_\Sigma^2} - 1 \right| \leq \nu_q^{2/q} m_q^{1-2/q} \tau^{2-q}(1 + m_q \tau^{-q}) \leq \frac{1}{2}. \tag{B.8}$$

With the above preparations, we are ready to establish the Gaussian approximation for $\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu} \rangle$. Define two centered Gaussian random variables $Z_1 \sim \mathcal{N}(0, \mathrm{var}(S_{1,\boldsymbol{u}}))$ and $Z_2 \sim \mathcal{N}(0, \|\boldsymbol{u}\|_\Sigma^2)$. By Lemma A.7 in Spokoiny and Zhilova (2015),

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(Z_1 \leq x) - \mathbb{P}(Z_2 \leq x)| \leq \frac{\delta_\tau}{2}.$$

It then follows from the Berry-Esseen inequality (see, e.g., Shevtsova (2014)) that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(S_{i,\boldsymbol{u}}-\mathbb{E}S_{i,\boldsymbol{u}})\le x\right\}-\mathbb{P}(Z_2\le x)\right| \\
&\le \sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(S_{i,\boldsymbol{u}}-\mathbb{E}S_{i,\boldsymbol{u}})\le x\right\}-\mathbb{P}(Z_1\le x)\right|+\sup_{x\in\mathbb{R}}|\mathbb{P}(Z_1\le x)-\mathbb{P}(Z_2\le x)| \\
&\le \frac{\mathbb{E}|S_{1,\boldsymbol{u}}-\mathbb{E}S_{1,\boldsymbol{u}}|^3}{2\mathrm{var}(S_{1,\boldsymbol{u}})^{3/2}\sqrt{n}}+\frac{\delta_\tau}{2}. \qquad \text{(B.9)}
\end{aligned}$$

This, together with the Bahadur representation (B.7), implies that for any $x\in\mathbb{R}$,

$$\begin{aligned}
\mathbb{P}(\sqrt{n}\langle\boldsymbol{u},\widehat{\boldsymbol{\mu}}_\tau-\boldsymbol{\mu}\rangle\le x) &\le \mathbb{P}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(S_{i,\boldsymbol{u}}-\mathbb{E}S_{i,\boldsymbol{u}})\le x+R_{n,\tau}\|\boldsymbol{u}\|_2\right\}+\frac{3}{n} \\
&\le \mathbb{P}(Z_2\le x+R_{n,\tau}\|\boldsymbol{u}\|_2)+\frac{\delta_\tau}{2}+\frac{\mathbb{E}|S_{1,\boldsymbol{u}}-\mathbb{E}S_{1,\boldsymbol{u}}|^3}{2\mathrm{var}(S_{1,\boldsymbol{u}})^{3/2}\sqrt{n}}+\frac{3}{n} \\
&\le \mathbb{P}(Z_2\le x)+\frac{R_{n,\tau}\|\boldsymbol{u}\|_2}{\sqrt{2\pi}\|\boldsymbol{u}\|_\Sigma}+\frac{\delta_\tau}{2}+\frac{\mathbb{E}|S_{1,\boldsymbol{u}}-\mathbb{E}S_{1,\boldsymbol{u}}|^3}{2\mathrm{var}(S_{1,\boldsymbol{u}})^{3/2}\sqrt{n}}+\frac{3}{n}.
\end{aligned}$$

For the third moment, we note that $\mathbb{E}|S_{1,\boldsymbol{u}}-\mathbb{E}S_{1,\boldsymbol{u}}|^3\le 4\mathbb{E}|S_{1,\boldsymbol{u}}|^3+4|\mathbb{E}S_{1,\boldsymbol{u}}|^3\le 8\mathbb{E}|S_{1,\boldsymbol{u}}|^3$ and $\mathbb{E}|S_{1,\boldsymbol{u}}|^3\le\mathbb{E}|\langle\boldsymbol{x}-\boldsymbol{\mu},\boldsymbol{u}\rangle|^3\le\nu_3\|\boldsymbol{u}\|_\Sigma^3$. Therefore, we have that

$$\frac{\mathbb{E}|S_{1,\boldsymbol{u}}-\mathbb{E}S_{1,\boldsymbol{u}}|^3}{2\mathrm{var}(S_{1,\boldsymbol{u}})^{3/2}\sqrt{n}}\lesssim\nu_3 n^{-1/2}. \qquad \text{(B.10)}$$

To bound the above key quantities $R_{n,\tau}\|\boldsymbol{u}\|_2/\|\boldsymbol{u}\|_\Sigma$ and δ_τ, we combine the definition of $R_{n,\tau}$ in (B.7) with (B.8) after taking $\tau\asymp m_q^{1/q}(n/\log n)^\gamma$ with $\gamma\in[1/(q-1),1/2]$ to get

$$\frac{R_{n,\tau}\|\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_\Sigma}\lesssim\frac{\sqrt{\mathrm{tr}(\Sigma)\log n}+\bar{\lambda}^{1/2}\log n}{\sqrt{\underline{\lambda}n}}+\frac{m_q^{1/q}}{\underline{\lambda}^{1/2}}\frac{\log n}{\sqrt{n}}\ \text{ and }\ \delta_\tau\lesssim\nu_q^{2/q}\left(\frac{\log n}{n}\right)^{(q-2)/(q-1)}. \qquad \text{(B.11)}$$

Putting together the pieces and note that $m_q^{1/q} \geq \mathrm{tr}(\Sigma)^{1/2}$, we obtain

$$\mathbb{P}(\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle \leq x) - \mathbb{P}(Z_2 \leq x) \lesssim \frac{m_q^{1/q}}{\underline{\lambda}^{1/2}} \frac{\log n}{\sqrt{n}} + \nu_q^{2/q} \left(\frac{\log n}{n}\right)^{(q-2)/(q-1)} + \frac{\nu_3}{\sqrt{n}}$$

for all $x \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^d$ as long as $n \gtrsim \mathrm{r}(\Sigma) + \log n$. A reversed inequality can be obtained via the same argument. Combining the two sides of the inequalities proves the claim. □

B.3.4 Proof of Theorem 2.2.4

Let $\tau \asymp m_4^{1/4}(n/\log n)^\gamma$ with $\gamma \in [1/3, 1/2]$ and $\xi \asymp v_0\sqrt{n/\log(nd)}$. Without loss of generality, we assume

$$n \gtrsim \nu_4 \left(\frac{\bar{\lambda}}{\underline{\lambda}}\right)^2 \mathrm{r}(\Sigma)\log(nd); \tag{B.12}$$

otherwise, the right-hand side of (2.14) is greater than 1 so that (2.14) holds trivially. As in the proof of Theorem 2.2.3, define

$$S_{i,\boldsymbol{u}} = \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}\langle \boldsymbol{u}, \boldsymbol{x}_i - \boldsymbol{\mu}\rangle$$

for each $\boldsymbol{u} \in \mathbb{R}^d$. By (B.7) and (B.11) with $q = 4$,

$$\begin{aligned}\left|\frac{\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle - n^{-1/2}\sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_\Sigma}\right| &\lesssim \frac{\sqrt{\mathrm{tr}(\Sigma)\log n} + \bar{\lambda}^{1/2}\log n}{\sqrt{\underline{\lambda} n}} + \frac{m_4^{1/4}}{\underline{\lambda}^{1/2}}\frac{\log n}{\sqrt{n}} \\ &\lesssim \frac{m_4^{1/4}}{\underline{\lambda}^{1/2}}\frac{\log n}{\sqrt{n}}\end{aligned} \tag{B.13}$$

with probability at least $1-3n^{-1}$ uniformly over all $\boldsymbol{u}\in\mathbb{R}^d$. On the other hand, using the bound (B.8) and the Berry-Esseen inequality (B.9), we get

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{n^{-1/2}\sum_{i=1}^n(S_{i,\boldsymbol{u}}-\mathbb{E}S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_\Sigma}\le x\right\}-\Phi(x)\right|\le\frac{\delta_\tau}{2}+\frac{\mathbb{E}|S_{1,\boldsymbol{u}}-\mathbb{E}S_{1,\boldsymbol{u}}|^3}{2\mathrm{var}(S_{1,\boldsymbol{u}})^{3/2}\sqrt{n}}$$
$$\lesssim \nu_4^{1/2}\left(\frac{\log n}{n}\right)^{(q-2)/(q-1)}+\frac{\nu_3}{\sqrt{n}}, \qquad \text{(B.14)}$$

for all $\boldsymbol{u}\in\mathbb{R}^d$. By Hölder's inequality, $\mathbb{E}|\langle\boldsymbol{x}-\boldsymbol{\mu},\boldsymbol{u}\rangle|^3\le(\mathbb{E}\langle\boldsymbol{x}-\boldsymbol{\mu},\boldsymbol{u}\rangle^2)^{1/2}(\mathbb{E}\langle\boldsymbol{x}-\boldsymbol{\mu},\boldsymbol{u}\rangle^4)^{1/2}$ for any $\boldsymbol{u}\in\mathbb{S}^{d-1}$, from which it follows that $\nu_3\le\nu_4^{1/2}$. Therefore, (B.14) is further bounded as

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{n^{-1/2}\sum_{i=1}^n(S_{i,\boldsymbol{u}}-\mathbb{E}S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_\Sigma}\le x\right\}-\Phi(x)\right|\lesssim\nu_4^{1/2}\sqrt{\frac{\log n}{n}}. \qquad \text{(B.15)}$$

For simplicity, we write $\widehat{\Sigma}=\widehat{\Sigma}_\xi$ and $\|\boldsymbol{u}\|_{\widehat{\Sigma}}^2=\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_\xi\boldsymbol{u}$. Note that v_0^2 can be written as

$$\frac{1}{2}\|\mathbb{E}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\}^2+\mathrm{tr}(\Sigma)\Sigma+2\Sigma^2\|_2,$$

so $v_0^2\le(\nu_4+1)\bar{\lambda}\mathrm{tr}(\Sigma)/2+\bar{\lambda}^2\le2\nu_4\bar{\lambda}\mathrm{tr}(\Sigma)$ according to Lemma 4.1 in Minsker and Wei (2020). It then follows from Proposition 2.2.1 that with probability at least $1-n^{-1}$,

$$\left|\|\boldsymbol{u}\|_{\widehat{\Sigma}}^2-\|\boldsymbol{u}\|_\Sigma^2\right|\lesssim\nu_4^{1/2}\|\boldsymbol{u}\|_2^2\sqrt{\frac{\bar{\lambda}\,\mathrm{tr}(\Sigma)\log(nd)}{n}}$$

for all $\boldsymbol{u}\in\mathbb{R}^d$. This further implies

$$\left|\frac{\|\boldsymbol{u}\|_{\widehat{\Sigma}}^2}{\|\boldsymbol{u}\|_\Sigma^2}-1\right|\lesssim\nu_4^{1/2}\frac{\|\boldsymbol{u}\|_2^2}{\|\boldsymbol{u}\|_\Sigma^2}\bar{\lambda}^{1/2}\sqrt{\frac{\mathrm{tr}(\Sigma)\log(nd)}{n}}\le\nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}}.$$

Using the elementary inequality $|x^{-1}-1|\le2|x^2-1|$ for any $x\ge1/2$, we obtain that with

probability at least $1-n^{-1}$,

$$\left|\frac{\|\boldsymbol{u}\|_{\Sigma}}{\|\boldsymbol{u}\|_{\widehat{\Sigma}}}-1\right| \lesssim \nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}}, \tag{B.16}$$

and hence $1/2 \le \|\boldsymbol{u}\|_{\widehat{\Sigma}}/\|\boldsymbol{u}\|_{\Sigma} \le 3/2$ under the same size condition (B.12). Combining this with (B.13) yields

$$\begin{aligned}
&\left|\frac{\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle}{\|\boldsymbol{u}\|_{\widehat{\Sigma}}} - \frac{n^{-1/2}\sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_{\Sigma}}\right| \\
&\le \left|\frac{\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle}{\|\boldsymbol{u}\|_{\widehat{\Sigma}}} - \frac{n^{-1/2}\sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_{\widehat{\Sigma}}}\right| + \left|\frac{n^{-1/2}\sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_{\Sigma}}\left(\frac{\|\boldsymbol{u}\|_{\Sigma}}{\|\boldsymbol{u}\|_{\widehat{\Sigma}}}-1\right)\right| \\
&\lesssim \frac{m_4^{1/4}}{\underline{\lambda}^{1/2}}\frac{\log n}{\sqrt{n}} + \nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}} \cdot \left|\frac{n^{-1/2}\sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_{\Sigma}}\right|.
\end{aligned}$$

By the Gaussian tail inequality, $\mathbb{P}\{|Z| \ge \sqrt{2\log(2n)}\} \le n^{-1}$, where $Z \sim \mathscr{N}(0,1)$. This together with (B.15) implies

$$\mathbb{P}\left\{\left|\frac{n^{-1/2}\sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_{\Sigma}}\right| \ge \sqrt{2\log(2n)}\right\} \lesssim \nu_4^{1/2}\left(\frac{\log n}{n}\right)^{1/2}.$$

Combining the pieces yields that with probability at least $1 - C_1\nu_4^{1/2}(n/\log n)^{-1/2} - 4n^{-1}$,

$$\begin{aligned}
&\left|\frac{\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle}{\|\boldsymbol{u}\|_{\widehat{\Sigma}}} - \frac{n^{-1/2}\sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_{\Sigma}}\right| \\
&\lesssim \frac{m_4^{1/4}}{\underline{\lambda}^{1/2}}\frac{\log n}{\sqrt{n}} + \nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}}\sqrt{\log n} \lesssim \nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\sqrt{\frac{\mathrm{r}(\Sigma)\log(n)\log(nd)}{n}}
\end{aligned}$$

uniformly over all $\boldsymbol{u}$, where the last inequality follows from the fact that $m_4 \le \kappa_4 \mathrm{tr}(\Sigma)^2$ (see (2.5)).

Finally, we conclude that for any $\boldsymbol{u} \in \mathbb{R}^d$ and $x \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}(\sqrt{n}\langle \boldsymbol{u}/\|\boldsymbol{u}\|_{\widehat{\Sigma}}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle \le x) \\
&\le \mathbb{P}\left\{ \frac{n^{-1/2}\sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}})}{\|\boldsymbol{u}\|_\Sigma} \le x + C_2 \nu_4^{1/2} \frac{\bar{\lambda}}{\underline{\lambda}} \sqrt{\frac{\mathrm{r}(\Sigma)\log(n)\log(nd)}{n}} \right\} + C_1 \nu_4^{1/2} \sqrt{\frac{\log n}{n}} + \frac{4}{n} \\
&\le \Phi(x) + C_3 \nu_4^{1/2} \frac{\bar{\lambda}}{\underline{\lambda}} \sqrt{\frac{\mathrm{r}(\Sigma)\log(n)\log(nd)}{n}}.
\end{aligned}$$

A similar argument leads to a series of reverse inequalities, thus proving the claimed bound. □

B.3.5 Proof of Theorem 2.2.5

Let $\tau \asymp m_4^{1/4}(n/\log n)^\gamma$ with $\gamma \in [1/3, 1/2]$ and $\xi \asymp v_0\sqrt{n/\log(nd)}$. Define the random vectors

$$\boldsymbol{S}_i = \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}) = (S_{i1}, \ldots, S_{id})^{\mathrm{T}}, \;\; i = 1, \ldots, n,$$

and write $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_\tau = (\widehat{\mu}_1, \ldots, \widehat{\mu}_d)^{\mathrm{T}}$. Recall from Lemma 2.2.1 that $\|\mathbb{E}\boldsymbol{S}_i\|_2 = b_\tau \le m_4\tau^{-3}$. Combining this with (2.8) ($z = \log n$) and the fact that $\mathrm{tr}(\Sigma) \le m_4^{1/2}$ yields

$$\left\| \sqrt{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - \frac{1}{\sqrt{n}}\sum_{i=1}^n (\boldsymbol{S}_i - \mathbb{E}\boldsymbol{S}_i) \right\|_2 \lesssim \frac{\sqrt{\mathrm{tr}(\Sigma)\log n} + \bar{\lambda}^{1/2}\log n}{\sqrt{n}} + m_4^{1/4}\frac{\log n}{\sqrt{n}} \lesssim m_4^{1/4}\frac{\log n}{\sqrt{n}}$$

with probability at least $1 - 3n^{-1}$. Consequently, we have

$$\max_{1\le k\le d} \left| \sqrt{n}(\widehat{\mu}_k - \mu_k) - \frac{1}{\sqrt{n}}\sum_{i=1}^n (S_{ik} - \mathbb{E}S_{ik}) \right| \le \left\| \sqrt{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - \frac{1}{\sqrt{n}}\sum_{i=1}^n (\boldsymbol{S}_i - \mathbb{E}\boldsymbol{S}_i) \right\|_2 \lesssim m_4^{1/4}\frac{\log n}{\sqrt{n}},$$

which in turn implies

$$\max_{1\le k\le d} \left| \frac{\sqrt{n}(\widehat{\mu}_k - \mu_k) - n^{-1/2}\sum_{i=1}^n (S_{ik} - \mathbb{E}S_{ik})}{\sqrt{\sigma_{kk}}} \right| \lesssim \frac{m_4^{1/4}}{\underline{\lambda}^{1/2}}\frac{\log n}{\sqrt{n}}. \tag{B.17}$$

Next we establish the Gaussian approximation for

$$\max_{1\le k\le d}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{S_{ik}-\mathbb{E}S_{ik}}{\sqrt{\sigma_{kk}}}\right|.$$

Let $\boldsymbol{G}=(G_1,\ldots,G_d)^{\mathrm{T}}$ be a zero-mean Gaussian vector whose covariance matrix, $\mathrm{cov}(\boldsymbol{G})$, is the correlation matrix of Σ, and define the Gaussian approximation error

$$\rho_n=\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\max_{1\le k\le d}\left|\frac{n^{-1/2}\sum_{i=1}^{n}(S_{ik}-\mathbb{E}S_{ik})}{\sqrt{\sigma_{kk}}}\right|\le x\right\}-\mathbb{P}(\|\boldsymbol{G}\|_\infty\le x)\right|.$$

Denote by $\mathbf{A}=(A_{kl})$ the covariance matrix of $(S_{i1}/\sqrt{\sigma_{11}},\ldots,S_{id}/\sqrt{\sigma_{dd}})^{\mathrm{T}}$, and let σ_*^2 be the smallest eigenvalue of $\mathbf{B}=(B_{kl}):=\mathrm{cov}(\boldsymbol{G})$. Note that, for any $1\le k,l\le d$,

$$A_{kl}=(\mathbb{E}S_{ik}S_{il}-\mathbb{E}S_{ik}\mathbb{E}S_{il})/\sqrt{\sigma_{kk}\sigma_{ll}}\ \text{ and }\ B_{kl}=\mathbb{E}(x_{ik}-\mu_k)(x_{il}-\mu_l)/\sqrt{\sigma_{kk}\sigma_{ll}}.$$

Moreover, define

$$\Delta_0=\frac{\log(d)}{\sigma_*^2}\|\mathbf{A}-\mathbf{B}\|_\infty,\quad \Delta_1=\frac{\log^2(d)}{n^2\sigma_*^4}\max_{1\le k\le d}\sum_{i=1}^{n}\frac{\mathbb{E}(S_{ik}-\mathbb{E}S_{ik})^4}{\sigma_{kk}^2},$$
$$\mathscr{M}=\left[\mathbb{E}\left\{\max_{1\le k\le d}\max_{1\le i\le n}\frac{(S_{ik}-\mathbb{E}S_{ik})^4}{\sigma_{kk}^2}\right\}\right]^{1/4}\ \text{ and }\ M=\max_{1\le i\le n}\left[\mathbb{E}\left\{\max_{1\le k\le d}\frac{(S_{ik}-\mathbb{E}S_{ik})^4}{\sigma_{kk}^2}\right\}\right]^{1/4}.$$

It thus follows from Theorem 2.2 in Chernozhukov, Chetverikov and Koike (2023) that

$$\rho_n\lesssim\log(n)\left\{\Delta_0+\sqrt{\Delta_1\log d}+\frac{(\mathscr{M}\log d)^2}{n\sigma_*^2}\right\}+(M/\sigma_*)^2\log(d)\sqrt{\frac{\log(n)\log(nd)}{n}}.\tag{B.18}$$

Note that $\sigma_*^2\ge\underline{\lambda}/\max_{1\le k\le d}\sigma_{kk}\ge\underline{\lambda}/\bar{\lambda}$, $\mathbb{E}(S_{1k}-\mathbb{E}S_{1k})^4\le 8\mathbb{E}(S_{1k}^4)+8(\mathbb{E}S_{1k})^4\le 16\mathbb{E}(S_{1k}^4)$ and $\mathbb{E}(S_{1k}^4)\le\mathbb{E}(x_{1k}-\mu_k)^4\le\kappa_4\sigma_{kk}^2$. Therefore,

$$\Delta_1\le 16\kappa_4(\bar{\lambda}/\underline{\lambda})^2\frac{\log^2(d)}{n}.$$

To bound $\mathscr{M}$, we have

$$\mathscr{M}^4 = \mathbb{E}\left\{ \max_{1\le k\le d} \max_{1\le i\le n} \frac{(S_{ik} - \mathbb{E}S_{ik})^4}{\sigma_{kk}^2} \right\} \le \sum_{k=1}^{d} \sum_{i=1}^{n} \frac{\mathbb{E}(S_{ik} - \mathbb{E}S_{ik})^4}{\sigma_{kk}^2} \le 16\kappa_4 nd.$$

Similarly, it can be shown that $M^4 \le 16\kappa_4 d$. It remains to bound Δ_0, or equivalently,

$$\|\mathbf{A} - \mathbf{B}\|_\infty = \max_{1\le k,l\le d} |\mathbb{E}(x_{1k} - \mu_k)(x_{1l} - \mu_l) - (\mathbb{E}S_{1k}S_{1l} - \mathbb{E}S_{1k}\mathbb{E}S_{1l})| / \sqrt{\sigma_{kk}\sigma_{ll}}.$$

Recall that $\tau \asymp m_4^{1/4}(n/\log n)^\gamma$ with $\gamma \in [1/4, 1/2]$. By Hölder's inequality,

$$\begin{aligned}
&|\mathbb{E}(x_{1k} - \mu_k)(x_{1l} - \mu_l) - \mathbb{E}S_{1k}S_{1l}| \\
&= \left| \mathbb{E}\left\{ \frac{\|\boldsymbol{x}_1 - \boldsymbol{\mu}\|_2^2 - \tau^2}{\|\boldsymbol{x}_1 - \boldsymbol{\mu}\|_2^2} (x_{1k} - \mu_k)(x_{1l} - \mu_l) \mathbb{1}(\|\boldsymbol{x}_1 - \boldsymbol{\mu}\|_2 \ge \tau) \right\} \right| \\
&\le \frac{1}{\tau^2} \mathbb{E}\|\boldsymbol{x}_1 - \boldsymbol{\mu}\|_2^2 |(x_{1k} - \mu_k)(x_{1l} - \mu_l)| \\
&\le \frac{1}{\tau^2} \left(\mathbb{E}\|\boldsymbol{x}_1 - \boldsymbol{\mu}\|_2^4\right)^{1/2} \left\{\mathbb{E}(x_{1k} - \mu_k)^4\right\}^{1/4} \left\{\mathbb{E}(x_{1l} - \mu_l)^4\right\}^{1/4} \\
&\lesssim \kappa_4^{1/2} \sqrt{\sigma_{kk}\sigma_{ll}} \left(\frac{\log n}{n}\right)^{2\gamma}.
\end{aligned}$$

On the other hand, it follows from (B.6) with slight modification that

$$|\mathbb{E}S_{1k}| \lesssim \kappa_4^{1/4} \sqrt{\sigma_{kk}} \left(\frac{\log n}{n}\right)^{3\gamma}.$$

Combining the above two inequalities, we obtain

$$\|\mathbf{A} - \mathbf{B}\|_\infty \lesssim \kappa_4^{1/2} \left(\frac{\log n}{n}\right)^{2\gamma} + \kappa_4^{1/2} \left(\frac{\log n}{n}\right)^{6\gamma} \lesssim \kappa_4^{1/2} \left(\frac{\log n}{n}\right)^{2\gamma},$$

which further implies

$$\Delta_0 \lesssim \kappa_4^{1/2} \frac{\bar{\lambda}}{\underline{\lambda}} \log(d) \left(\frac{\log n}{n}\right)^{2\gamma} \le \kappa_4^{1/2} \frac{\bar{\lambda}}{\underline{\lambda}} \log(d) \sqrt{\frac{\log n}{n}}.$$

Putting together the pieces, we conclude that

$$\begin{aligned}\rho_n &\lesssim \log(n)\left\{\kappa_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log(d)\sqrt{\frac{\log n}{n}}+\kappa_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\frac{\log^{3/2}(d)}{\sqrt{n}}+\kappa_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log^2(d)\sqrt{\frac{d}{n}}\right\}\\ &\qquad\qquad +\kappa_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log(d)\sqrt{\log(n)\log(nd)}\sqrt{\frac{d}{n}}\\ &\lesssim \kappa_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log(d)\log(n)\left\{\sqrt{\frac{\log n}{n}}+\log(d)\sqrt{\frac{d}{n}}\right\}\lesssim \kappa_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log^2(d)\log(n)\sqrt{\frac{d}{n}}. \qquad \text{(B.19)}\end{aligned}$$

Thus far, the obtained bounds (B.17) and (B.19) only involve the true variances. To account for the impact of variance estimation error, let $\widehat{\sigma}_{kk}$ be the k-th diagonal element of $\widehat{\Sigma}=\widehat{\Sigma}_{\xi}$, and recall from (B.16) that

$$\max_{1\le k\le d}\left|\frac{\sqrt{\sigma_{kk}}}{\sqrt{\widehat{\sigma}_{kk}}}-1\right|\lesssim \nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}}$$

with probability at least $1-n^{-1}$. Without loss of generality, we assume sample size condition (B.12) holds; otherwise, the right-hand side of (2.16) is greater than 1 so that the claim of Theorem 2.2.5 holds trivially. Consequently, $1/2\le\sqrt{\sigma_{kk}/\widehat{\sigma}_{kk}}\le 3/2$ for all $1\le k\le d$. Combining this with (B.17) yields

$$\begin{aligned}&\max_{1\le k\le d}\left|\frac{\sqrt{n}(\widehat{\mu}_k-\mu_k)}{\sqrt{\widehat{\sigma}_{kk}}}-\frac{n^{-1/2}\sum_{i=1}^n(S_{ik}-\mathbb{E}S_{ik})}{\sqrt{\sigma_{kk}}}\right|\\ &\le \max_{1\le k\le d}\left|\frac{\sqrt{n}(\widehat{\mu}_k-\mu_k)-n^{-1/2}\sum_{i=1}^n(S_{ik}-\mathbb{E}S_{ik})}{\sqrt{\widehat{\sigma}_{kk}}}\right|\\ &\qquad+\max_{1\le k\le d}\left|\frac{n^{-1/2}\sum_{i=1}^n(S_{ik}-\mathbb{E}S_{ik})}{\sqrt{\sigma_{kk}}}\left(\frac{\sqrt{\sigma_{kk}}}{\sqrt{\widehat{\sigma}_{kk}}}-1\right)\right|\\ &\lesssim \kappa_4^{1/4}\frac{\mathrm{tr}(\Sigma)^{1/2}}{\underline{\lambda}^{1/2}}\frac{\log n}{\sqrt{n}}+\nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}}\max_{1\le k\le d}\left|\frac{n^{-1/2}\sum_{i=1}^n(S_{ik}-\mathbb{E}S_{ik})}{\sqrt{\sigma_{kk}}}\right| \qquad \text{(B.20)}\end{aligned}$$

with probability at least $1-4n^{-1}$, where we used the property $m_4\le\kappa_4\mathrm{tr}(\Sigma)^2$ in the last step. By

the Gaussian tail inequality and a union bound argument,

$$\mathbb{P}\left\{\max_{1\le k\le d}|G_k|\ge\sqrt{2\log(2nd)}\right\}\le\frac{1}{n}.$$

This together with the definition of ρ_n implies

$$\begin{aligned}&\mathbb{P}\left\{\max_{1\le k\le d}\left|\frac{n^{-1/2}\sum_{i=1}^n(S_{ik}-\mathbb{E}S_{ik})}{\sqrt{\sigma_{kk}}}\right|\ge\sqrt{2\log(2nd)}\right\}\\&\le\mathbb{P}\left\{\max_{1\le k\le d}|G_k|\ge\sqrt{2\log(2nd)}\right\}+\rho_n\le\frac{1}{n}+\rho_n.\end{aligned}$$

Combining this with (B.20) yields

$$\max_{1\le k\le d}\left|\frac{\sqrt{n}(\widehat{\mu}_k-\mu_k)}{\sqrt{\widehat{\sigma}_{kk}}}-\frac{n^{-1/2}\sum_{i=1}^n(S_{ik}-\mathbb{E}S_{ik})}{\sqrt{\sigma_{kk}}}\right|\lesssim\nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log(nd)\sqrt{\frac{\mathrm{r}(\Sigma)}{n}}$$

with probability at least $1-\rho_n-5n^{-1}$.

With the above preparations, we are ready to prove the final claim. Let $\Psi(\cdot)$ be the distribution function of $\|\boldsymbol{G}\|_\infty$. For any $x\ge0$,

$$\begin{aligned}&\mathbb{P}\left\{\max_{1\le k\le d}\left|\frac{\sqrt{n}(\widehat{\mu}_k-\mu_k)}{\sqrt{\widehat{\sigma}_{kk}}}\right|\le x\right\}\\&\le\mathbb{P}\left\{\max_{1\le k\le d}\left|\frac{n^{-1/2}\sum_{i=1}^n(S_{ik}-\mathbb{E}S_{ik})}{\sqrt{\sigma_{kk}}}\right|\le x+C_4\nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log(nd)\sqrt{\frac{\mathrm{r}(\Sigma)}{n}}\right\}+\rho_n+\frac{5}{n}\\&\le\mathbb{P}\left\{\|\boldsymbol{G}\|_\infty\le x+C_4\nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log(nd)\sqrt{\frac{\mathrm{r}(\Sigma)}{n}}\right\}+2\rho_n+\frac{5}{n}\\&=\Psi(x)+\Psi\left(x+C_4\nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log(nd)\sqrt{\frac{\mathrm{r}(\Sigma)}{n}}\right)-\Psi(x)+2\rho_n+\frac{5}{n}.\end{aligned}$$

By Nazarov's inequality (Nazarov, 2003),

$$\sup_{x\ge0}\left|\Psi\left(x+C_4\nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log(nd)\sqrt{\frac{\mathrm{r}(\Sigma)}{n}}\right)-\Psi(x)\right|\lesssim\nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\log(nd)\sqrt{\frac{\mathrm{r}(\Sigma)\log(ed)}{n}}.$$

Substituting this into the earlier bound, we obtain that

$$\begin{aligned}
&\mathbb{P}\left\{ \max_{1\le k\le d} \left| \frac{\sqrt{n}(\widehat{\mu}_k - \mu_k)}{\sqrt{\widehat{\sigma}_{kk}}} \right| \le x \right\} - \Psi(x) \\
&\lesssim \nu_4^{1/2} \frac{\bar{\lambda}}{\underline{\lambda}} \log(nd) \sqrt{\frac{\mathrm{r}(\Sigma)\log(ed)}{n}} + \kappa_4^{1/2} \frac{\bar{\lambda}}{\underline{\lambda}} \log^2(d)\log(n)\sqrt{\frac{d}{n}} \\
&\lesssim \nu_4^{1/2} \frac{\bar{\lambda}}{\underline{\lambda}} \log^2(d)\log(n)\sqrt{\frac{d}{n}}
\end{aligned}$$

for all $x \ge 0$. A lower bound can be similarly obtained. Since all the bounds are independent of x, taking the supremum over $x \ge 0$ proves the claim. □

B.3.6 Proof of Proposition 2.2.2

Recall that $R = \mathrm{corr}(\Sigma)$ and $\widehat{R} = \mathrm{corr}(\widehat{\Sigma})$. We first claim that

$$\max_{1\le k,l\le d} |\widehat{R}_{kl} - R_{kl}| \le \|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - \mathrm{I}_d\|_2 (2 + \|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}\|_2). \tag{B.21}$$

To show this, note that

$$\begin{aligned}
\max_{1\le k,l\le d} |\widehat{R}_{kl} - R_{kl}| &= \max_{1\le k,l\le d} \left| \frac{\widehat{\Sigma}_{kl}}{\sqrt{\widehat{\Sigma}_{kk}\widehat{\Sigma}_{ll}}} - \frac{\Sigma_{kl}}{\sqrt{\Sigma_{kk}\Sigma_{ll}}} \right| \\
&\le \max_{1\le k,l\le d} \left| \frac{\widehat{\Sigma}_{kl}}{\sqrt{\widehat{\Sigma}_{kk}\widehat{\Sigma}_{ll}}} - \frac{\widehat{\Sigma}_{kl}}{\sqrt{\Sigma_{kk}\Sigma_{ll}}} \right| + \max_{1\le k,l\le d} \left| \frac{\widehat{\Sigma}_{kl}}{\sqrt{\Sigma_{kk}\Sigma_{ll}}} - \frac{\Sigma_{kl}}{\sqrt{\Sigma_{kk}\Sigma_{ll}}} \right|.
\end{aligned} \tag{B.22}$$

Denoting $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$ to be the canonical basis of $\mathbb{R}^d$, we have

$$\begin{aligned}
\max_{1\le k,l\le d} \left| \frac{\widehat{\Sigma}_{kl}}{\sqrt{\Sigma_{kk}\Sigma_{ll}}} - \frac{\Sigma_{kl}}{\sqrt{\Sigma_{kk}\Sigma_{ll}}} \right| &= \max_{1\le k,d\le d} \left| \frac{\boldsymbol{e}_k^{\mathrm{T}}\widehat{\Sigma}\boldsymbol{e}_l}{\sqrt{\boldsymbol{e}_k^{\mathrm{T}}\Sigma\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}\Sigma\boldsymbol{e}_l}} - \frac{\boldsymbol{e}_k^{\mathrm{T}}\Sigma\boldsymbol{e}_l}{\sqrt{\boldsymbol{e}_k^{\mathrm{T}}\Sigma\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}\Sigma\boldsymbol{e}_l}} \right| \\
&= \max_{1\le k,l\le d} \left| \frac{\boldsymbol{a}_k^{\mathrm{T}}\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}\boldsymbol{a}_l}{\|\boldsymbol{a}_k\|_2\|\boldsymbol{a}_l\|_2} - \frac{\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{a}_l}{\|\boldsymbol{a}_k\|_2\|\boldsymbol{a}_l\|_2} \right| \\
&\le \|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - \mathrm{I}_d\|_2,
\end{aligned} \tag{B.23}$$

where $\boldsymbol{a}_k = \Sigma^{1/2}\boldsymbol{e}_k$ for $1 \le k \le d$.

Turning to the first term in the right-hand side of (B.22), it follows that

$$\begin{aligned}
\max_{1\le k,l\le d}\left|\frac{\widehat{\Sigma}_{kl}}{\sqrt{\widehat{\Sigma}_{kk}\widehat{\Sigma}_{ll}}}-\frac{\widehat{\Sigma}_{kl}}{\sqrt{\Sigma_{kk}\Sigma_{ll}}}\right| &\le \max_{1\le k,l\le d}\left|\frac{\widehat{\Sigma}_{kl}}{\sqrt{\widehat{\Sigma}_{kk}\widehat{\Sigma}_{ll}}}\right|\cdot\left|1-\sqrt{\frac{\widehat{\Sigma}_{kk}\widehat{\Sigma}_{ll}}{\Sigma_{kk}\Sigma_{ll}}}\right| \\
&\le \max_{1\le k,l\le d}\left|1-\sqrt{\frac{\widehat{\Sigma}_{kk}\widehat{\Sigma}_{ll}}{\Sigma_{kk}\Sigma_{ll}}}\right| \\
&\le \max_{1\le k,l\le d}\left|1-\frac{\widehat{\Sigma}_{kk}\widehat{\Sigma}_{ll}}{\Sigma_{kk}\Sigma_{ll}}\right| \\
&= \max_{1\le k,l\le d}\left|1-\frac{\widehat{\Sigma}_{kk}}{\Sigma_{kk}}-\left(\frac{\widehat{\Sigma}_{ll}}{\Sigma_{ll}}-1\right)\frac{\widehat{\Sigma}_{kk}}{\Sigma_{kk}}\right|,
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality and the last inequality can be derived from the elementary inequality $|1-\sqrt{x}| \le |1-x|$ for any $x \ge 0$. Together, by the definition of the operator norm of a matrix, the earlier bound can be further bounded as

$$\begin{aligned}
&\max_{1\le k,l\le d}\left|\frac{\widehat{\Sigma}_{kl}}{\sqrt{\widehat{\Sigma}_{kk}\widehat{\Sigma}_{ll}}}-\frac{\widehat{\Sigma}_{kl}}{\sqrt{\Sigma_{kk}\Sigma_{ll}}}\right| \\
&\le \max_{1\le k\le d}\left|1-\frac{\boldsymbol{a}_k^{\mathrm{T}}\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}\boldsymbol{a}_k}{\|\boldsymbol{a}_k\|_2^2}\right|+\max_{1\le l\le d}\left|1-\frac{\boldsymbol{a}_l^{\mathrm{T}}\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}\boldsymbol{a}_l}{\|\boldsymbol{a}_l\|_2^2}\right|\cdot\max_{1\le k\le d}\left|\frac{\boldsymbol{a}_k^{\mathrm{T}}\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}\boldsymbol{a}_k}{\|\boldsymbol{a}_k\|_2^2}\right| \\
&\le \|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}-\mathrm{I}_d\|_2+\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}-\mathrm{I}_d\|_2\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}\|_2.
\end{aligned}$$

Combining this with (B.22) and (B.23) proves the claimed bound (B.21).

Now, we are ready to prove Proposition 2.2.2. Without loss of generality, we assume that the sample size n satisfies (B.12); otherwise, the right-hand side of (2.18) is greater than 1 so that (2.18) holds trivially. By Theorem 2.2.4, we have $\nu_0^2 \le 2\nu_4\bar{\lambda}\mathrm{tr}(\Sigma)$. Combining this with Proposition 2.2.1, the robust covariance estimator $\widehat{\Sigma} = \widehat{\Sigma}_\xi$ defined in (2.10) with

$\xi \asymp v_0\sqrt{n/\log(nd)}$ satisfies

$$\|\widehat{\Sigma}-\Sigma\|_2 \lesssim v_4^{1/2}\bar{\lambda}\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}}$$

with probability at least $1-2n^{-1}$. Thus, with the same probability, it follows that

$$\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}-\mathrm{I}_d\|_2 \lesssim v_4^{1/2}(\bar{\lambda}/\underline{\lambda})\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}}.$$

This further implies that $\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}\|_2 \lesssim 1$ under the sample size condition (B.12). Combining these two bounds with (B.21), we have

$$\max_{1\le k,l\le d}|\widehat{R}_{kl}-R_{kl}| \lesssim v_4^{1/2}(\bar{\lambda}/\underline{\lambda})\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}}.$$

Together, this bound and Theorem 1.1 of Fang and Koike (2021) prove the claim.

□

B.4 Proofs in Section 2.3

B.4.1 Supporting lemmas

To establish the statistical properties of the noisy gradient descent iterates $\boldsymbol{\mu}^{(t)}$, the landscape of the loss function plays an important role. The following lemma shows that the empirical loss function $\widehat{\mathscr{L}_\tau}(\cdot):\mathbb{R}^d\to\mathbb{R}^+$ is locally strongly convex and satisfies a local smoothness condition. Recall that $\widehat{\boldsymbol{\mu}}=\widehat{\boldsymbol{\mu}}_\tau=\mathrm{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d}\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta})$ is the non-private Huber estimator, satisfying $\nabla\widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}})=\mathbf{0}$.

Lemma B.4.1. Conditioned on the event $\mathscr{E}_1=\mathscr{E}_1(r_0,\chi)$ defined in (2.22), we have

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_2)-\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_1)-\langle\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_1),\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\rangle \ge \frac{1-\chi}{2}\|\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\|_2^2 \ \text{ for all } \ \boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\Theta(r_0), \tag{B.24}$$

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) \geq \frac{1-\chi}{4} r_0 \cdot \|\boldsymbol{\theta} - \widehat{\boldsymbol{\mu}}\|_2 \ \text{ for all } \ \boldsymbol{\theta} \in \Theta(r_0)^{\mathrm{c}} \tag{B.25}$$

and

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_2) - \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_1) - \langle \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \rangle \leq \frac{1}{2} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2^2 \ \text{ for all } \ \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d. \tag{B.26}$$

Proof of Lemma B.4.1. For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta(r_0)$, by Taylor's theorem in several variables we have

$$\begin{aligned}
&\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_2) - \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_1) - \langle \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \rangle \\
&= \frac{1}{2} (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)^{\mathrm{T}} \nabla^2 \widehat{\mathscr{L}_\tau}\big(\boldsymbol{\theta}_1 + u(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)\big) (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \ \text{ for some } u \in (0,1),
\end{aligned}$$

from which it follows that

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_2) - \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_1) - \langle \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \rangle \geq \frac{1}{2} \min_{\boldsymbol{h} \in \mathbb{B}^d(r_0)} \lambda_{\min}\big(\nabla^2 \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu} + \boldsymbol{h})\big) \cdot \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2^2. \tag{B.27}$$

Conditioned on $\mathscr{E}_1$, this proves (B.24).

Given $\widehat{\boldsymbol{\mu}} \in \Theta(r_0/2)$, set $\boldsymbol{\delta} = \boldsymbol{\theta} - \widehat{\boldsymbol{\mu}}$. It follows from the first-order Taylor's theorem that

$$\widehat{R}_\tau(\boldsymbol{\delta}) := \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) - \langle \underbrace{\nabla \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}})}_{=\mathbf{0}}, \boldsymbol{\theta} - \widehat{\boldsymbol{\mu}} \rangle = \int_0^1 \langle \nabla \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}} + u\boldsymbol{\delta}) - \nabla \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}), \boldsymbol{\delta} \rangle \mathrm{d}u.$$

For any $v \in (0,1)$, by the convexity lemma—Lemma C.1 in Sun, Zhou and Fan (2020),

$$\langle \nabla \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}} + u\boldsymbol{\delta}) - \nabla \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}), u\boldsymbol{\delta} \rangle \geq \frac{1}{v} \langle \nabla \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}} + uv\boldsymbol{\delta}) - \nabla \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}), uv\boldsymbol{\delta} \rangle, \ \ u > 0,$$

which in turn implies $\widehat{R}_\tau(\boldsymbol{\delta}) \geq v^{-1} \widehat{R}_\tau(v\boldsymbol{\delta})$. Hence, for any $\boldsymbol{\theta} \in \Theta(r_0)^{\mathrm{c}}$ so that $\|\boldsymbol{\delta}\|_2 = \|\boldsymbol{\theta} - \widehat{\boldsymbol{\mu}}\|_2 >$

$r_0/2$, taking $v = r_0/(2\|\boldsymbol{\theta} - \widehat{\boldsymbol{\mu}}\|_2) \in (0,1)$ and $\boldsymbol{\delta}_0 = v \cdot (\boldsymbol{\theta} - \widehat{\boldsymbol{\mu}}) \in \partial \mathbb{B}^d(r_0/2)$ we obtain that

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) \geq 2r_0^{-1}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\mu}}\|_2 \cdot \widehat{R}_\tau(\boldsymbol{\delta}_0) = 2r_0^{-1}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\mu}}\|_2 \big\{\widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}} + \boldsymbol{\delta}_0) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}})\big\}.$$

Moreover, applying the bound (B.27) at $\boldsymbol{\theta}_1 = \widehat{\boldsymbol{\mu}}$ and $\boldsymbol{\theta}_2 = \widehat{\boldsymbol{\mu}} + \boldsymbol{\delta}_0 \in \Theta(r_0)$ gives

$$\widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}} + \boldsymbol{\delta}_0) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) \geq \frac{1}{2} \min_{\boldsymbol{h} \in \mathbb{B}^d(r_0)} \lambda_{\min}\big(\nabla^2 \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu} + \boldsymbol{h})\big) \cdot \|\boldsymbol{\delta}_0\|_2^2.$$

Together, the last two displays imply

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) \geq \frac{1}{2} \min_{\boldsymbol{h} \in \mathbb{B}^d(r_0)} \lambda_{\min}\big(\nabla^2 \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu} + \boldsymbol{h})\big) \cdot \frac{r_0}{2} \cdot \|\boldsymbol{\theta} - \widehat{\boldsymbol{\mu}}\|_2,$$

verifying the second bound (B.25).

Finally, (B.26) is a direct consequence of (B.4) and Taylor's theorem. □

The following lemma provides upper bounds for the i.i.d. standard normal random vectors $\{\boldsymbol{g}_t\}_{t=0}^{T-1}$ in the noisy gradient descent algorithm. In particular, inequality (B.29) is a slightly improved version of the tail bound in Lemma 11 of Cai, Wang and Zhang (2021).

Lemma B.4.2. Let $g_0, g_1, \ldots, g_{T-1} \in \mathbb{R}^d$ ($T \geq 1$) be independent standard multivariate normal random vectors. Then, for any $z \geq 0$,

$$\mathbb{P}\left\{\max_{0 \leq t \leq T-1} \|\boldsymbol{g}_t\|_2 \geq d^{1/2} + \sqrt{2(\log T + z)}\right\} \leq e^{-z}. \tag{B.28}$$

Moreover, for any $\rho \in (0,1)$, we have with probability at least $1 - e^{-z}$ that

$$\sum_{t=0}^{T-1} \rho^t \|\boldsymbol{g}_t\|_2^2 \leq \frac{d}{1-\rho} + 2\sqrt{\frac{dz}{1-\rho^2}} + 2z \leq \frac{2d}{1-\rho} + \left(\frac{1}{1+\rho} + 2\right) z. \tag{B.29}$$

Proof of Lemma B.4.2. For each $t = 0, 1, \ldots, T-1$, we apply the concentration inequality for

Lipschitz functions of standard normal random variables and obtain that

$$\mathbb{P}\big(\|\boldsymbol{g}_t\|_2 \geq d^{1/2} + \sqrt{2z}\big) \leq e^{-z}, \quad \text{valid for any } z \geq 0.$$

Combining this with the union bound (over $t = 0, 1, \ldots, T-1$) yields (B.28).

Note that $\|\boldsymbol{g}_t\|_2^2$ follows the chi-square distribution χ_d^2, which is a special case of the gamma distribution $\Gamma(d/2, 1/2)$. The centered variable, $\|\boldsymbol{g}_t\|_2^2 - d$, is known to be sub-gamma with parameters $v = 2d$ and $c = 2$ (Boucheron, Lugosi and Massart, 2013). Let $Z = \sum_{t=0}^{T-1} \rho^t (\|\boldsymbol{g}_t\|_2^2 - d)$. For each t and $0 < \lambda < 1/c$,

$$\log \mathbb{E} e^{\lambda \rho^t (\|\boldsymbol{g}_t\|_2^2 - d)} \leq \frac{v\lambda^2\rho^{2t}}{2(1 - c\lambda\rho^t)} \leq \frac{v\lambda^2\rho^{2t}}{2(1 - c\lambda)}.$$

By independence,

$$\log \mathbb{E} e^{\lambda Z} = \sum_{t=0}^{T-1} \log \mathbb{E} e^{\lambda \rho^t (\|\boldsymbol{g}_t\|_2^2 - d)} \leq \frac{v\lambda^2 \sum_{t=0}^{T-1} \rho^{2t}}{2(1 - c\lambda)} \leq \frac{v}{1 - \rho^2} \frac{\lambda^2}{2(1 - c\lambda)}.$$

Therefore, the centered variable Z is sub-gamma with parameters $(v/(1-\rho^2), c) = (2d/(1-\rho^2), 2)$. Applying Chernoff's bound to Z (see, e.g., Section 2.4 of Boucheron, Lugosi and Massart (2013)) yields

$$\mathbb{P}\big\{Z > 2(dz)^{1/2}(1-\rho^2)^{-1/2} + 2z\big\} \leq e^{-z} \text{ for any } z > 0.$$

This, combined with the elementary inequality $\sum_{t=0}^{T-1} \rho^t \leq 1/(1-\rho)$, proves (B.29). □

Lemma B.4.3 below provides a useful property that will be needed in the proof of Proposition 2.3.2.

Lemma B.4.3. Let $R_0 = \|\boldsymbol{\mu}^{(0)} - \widehat{\boldsymbol{\mu}}\|_2$ and $\eta_0 \in (0, 1]$. For any $T_0 > 1$ and $z > 0$, let the sample

size satisfy

$$n \geq \frac{8(e-1)}{4-e} \frac{T_0 T^{1/2} B_{T_0}}{\varepsilon} \max\left\{ \frac{\tau\eta_0}{R_0}, \left(\frac{\tau\eta_0}{R_0}\right)^2 \right\}, \tag{B.30}$$

where $B_{T_0} = \sqrt{d} + \sqrt{2(\log T_0 + z)}$. Then, the noisy gradient descent iterates $\boldsymbol{\mu}^{(t)}$ defined in (2.21) satisfy

$$\max_{1 \leq t \leq T_0} \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2 \leq 2R_0$$

with probability (over $\{\boldsymbol{g}_t\}_{t=0}^{T_0-1}$) at least $1 - e^{-z}$.

Proof of Lemma B.4.3. Recall that $\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} - \eta_0 \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) + \eta_0 \boldsymbol{h}_t$, where $\boldsymbol{h}_t = 2T^{1/2}\tau \frac{\boldsymbol{g}_t}{\varepsilon n}$. Moreover, $\|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta})\|_2 \leq \tau$ and $\|\nabla^2 \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta})\|_2 \leq 1$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$. Therefore,

$$\langle \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_1) - \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \geq \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_1) - \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}_2)\|_2^2, \quad \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d.$$

Since $\nabla \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) = \mathbf{0}$, taking $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\boldsymbol{\mu}^{(t)}, \widehat{\boldsymbol{\mu}})$ for any t gives

$$\langle \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}} \rangle \geq \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2.$$

Then, for any fixed step size $0 < \eta_0 \leq 1$, we have

$$\begin{aligned}
\|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2^2 &= \|\boldsymbol{\mu}^{(t)} - \eta_0 \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) + \eta_0 \boldsymbol{h}_t - \widehat{\boldsymbol{\mu}}\|_2^2 \\
&= \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 + \eta_0^2 \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \boldsymbol{h}_t\|_2^2 - 2\eta_0 \langle \boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}, \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \boldsymbol{h}_t \rangle \\
&\leq \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 + \eta_0^2 \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2 + \eta_0^2 \|\boldsymbol{h}_t\|_2^2 + 2\eta_0^2 \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2 \|\boldsymbol{h}_t\|_2 \\
&\quad - 2\eta_0 \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2 + 2\eta_0 \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2 \|\boldsymbol{h}_t\|_2 \\
&\leq \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 + \eta_0^2 \|\boldsymbol{h}_t\|_2^2 + 2\eta_0 \|\boldsymbol{h}_t\|_2 \big\{ \eta_0 \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2 + \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2 \big\} \\
&\leq \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 + \eta_0^2 \|\boldsymbol{h}_t\|_2^2 + 2\eta_0 \|\boldsymbol{h}_t\|_2 (\eta_0 \tau + \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2).
\end{aligned} \tag{B.31}$$

For any integer $T_0 \geq 1$ and $z > 0$, Lemma B.4.2 shows that

$$\max_{0 \leq t \leq T_0 - 1} \|\boldsymbol{h}_t\|_2 \leq \frac{2T^{1/2}\tau B_{T_0}}{\varepsilon n} =: e_{\text{priv}}$$

with probability at least $1 - e^{-z}$, where $B_{T_0} = d^{1/2} + \sqrt{2(\log T_0 + z)}$. Throughout the rest of the proof, we assume this inequality holds.

For some $\eta \in (0,1)$ to be determined, it follows from (B.31) that

$$\|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2^2 \leq (1+\eta)\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 + (1+1/\eta)\eta_0^2 e_{\text{priv}}^2 + 2\eta_0^2 \tau e_{\text{priv}},$$

for $t = 0, 1, \ldots, T_0 - 1$. This recursive bound further implies

$$\begin{aligned}
\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 &\leq (1+\eta)^t \|\boldsymbol{\mu}^{(0)} - \widehat{\boldsymbol{\mu}}\|_2^2 + \{(1+1/\eta)\eta_0^2 e_{\text{priv}}^2 + 2\eta_0^2 \tau e_{\text{priv}}\} \sum_{k=0}^{t-1} (1+\eta)^k \\
&\leq (1+\eta)^t \|\boldsymbol{\mu}^{(0)} - \widehat{\boldsymbol{\mu}}\|_2^2 + \frac{(1+\eta)^t - 1}{\eta} \{(1+1/\eta)e_{\text{priv}} + 2\tau\}\eta_0^2 e_{\text{priv}}.
\end{aligned}$$

Note that under the sample size requirement (B.30), we have

$$(e-1)(T_0+1)T_0\eta_0^2 e_{\text{priv}}^2 \leq 2(e-1)T_0^2\eta_0^2 e_{\text{priv}}^2 \leq \frac{4-e}{2} R_0^2,$$

and $2\tau(e-1)T_0\eta_0^2 e_{\text{priv}} \leq (4-e)R_0^2/2$. Thus, provided $T_0 \geq 2$, we take $\eta = 1/T_0 \in (0,1)$ and obtain

$$\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 \leq eR_0^2 + (e-1)\big\{(T_0+1)e_{\text{priv}} + 2\tau\big\}T_0\eta_0^2 e_{\text{priv}} \leq eR_0^2 + (4-e)R_0^2 = 4R_0^2$$

for all $t = 1, \ldots, T_0$, as claimed. □

Finally, the following lemma is a direct consequence of Corollary 4.4.8 in Vershynin (2018).

Lemma B.4.4. Let $\mathbf{E}$ be a $d \times d$ symmetric random matrix whose entries E_{ij} on and above the diagonal are independent $\mathcal{N}(0,1)$. Then, for any $z > 0$ we have $\|\mathbf{E}\|_2 \lesssim \sqrt{d} + z$ with probability at least $1 - 4e^{-z^2}$.

B.4.2 Proof of Theorem 2.3.1

Recall the event $\mathscr{E}_1 = \mathscr{E}_1(r_0, \chi)$ given in (2.22). To control the random perturbations in noisy gradient descent, for some $\rho \in (0,1)$ to be determined, define

$$\mathscr{E}_2 = \mathscr{E}_2(z) = \left\{ \max_{0 \le t \le T-1} \|\boldsymbol{g}_t\|_2 \le B_T \right\} \bigcap \left\{ \sum_{t=0}^{T-1} (1-\rho)^t \|\boldsymbol{g}_{T-1-t}\|_2^2 \le 2\rho^{-1} d + 3z \right\}, \quad \text{(B.32)}$$

where $B_T = B_T(z) := d^{1/2} + \sqrt{2(\log T + z)}$. It thus follows from Lemma B.4.2 that $\mathbb{P}\{\mathscr{E}_2(z)\} \ge 1 - 2e^{-z}$. In the following, we prove the result by conditioning on the event $\mathscr{E}_1 \cap \mathscr{E}_2$. Starting from an initial value $\boldsymbol{\mu}^{(0)} \in \Theta(r_0)$, the following proposition shows that all successive iterates will stay in the ball $\Theta(r_0)$.

Proposition B.4.1. Under the conditions of Theorem 2.3.1, and conditioning on $\mathscr{E}_1$, all the iterates $\boldsymbol{\mu}^{(t)}$ $(t = 1, \ldots, T)$ stay in the local ball $\Theta(r_0)$.

Next, we establish a contraction property for the noisy gradient descent iterates. Define

$$\widetilde{\boldsymbol{\mu}}^{(t+1)} = \boldsymbol{\mu}^{(t)} - \eta_0 \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) \quad \text{and} \quad \boldsymbol{h}_t = 2T^{1/2} \frac{\tau}{\varepsilon n} \boldsymbol{g}_t, \ \ t = 0, 1, \ldots, T-1,$$

and note that $\boldsymbol{\mu}^{(t+1)} = \widetilde{\boldsymbol{\mu}}^{(t+1)} + \eta_0 \boldsymbol{h}_t$. Under the sample size requirement (2.24), Proposition B.4.1 ensures that $\boldsymbol{\mu}^{(t)} \in \Theta(r_0)$ for all $t = 0, 1, \ldots, T$. Similarly, it can be shown that the non-private gradient descent iterates $\widetilde{\boldsymbol{\mu}}^{(t)}$ also stay in the ball $\Theta(r_0)$ for $t = 1, \ldots, T$. For simplicity, set

$$\boldsymbol{\delta}^{(t)} = \boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}, \quad \widetilde{\boldsymbol{\delta}}^{(t)} = \widetilde{\boldsymbol{\mu}}^{(t)} - \widehat{\boldsymbol{\mu}} \quad \text{for } \ t = 0, 1, \ldots, T.$$

For any $\eta \in (0,1)$, at each iteration we bound $\|\boldsymbol{\delta}^{(t+1)}\|_2 = \|\widetilde{\boldsymbol{\delta}}^{(t+1)} + \eta_0 \boldsymbol{h}_t\|_2$ as

$$\begin{aligned}
\|\boldsymbol{\delta}^{(t+1)}\|_2^2 &= \|\widetilde{\boldsymbol{\delta}}^{(t+1)}\|_2^2 + \eta_0^2\|\boldsymbol{h}_t\|_2^2 + 2\eta_0\langle\widetilde{\boldsymbol{\delta}}^{(t+1)}, \boldsymbol{h}_t\rangle \\
&\leq (1+\eta)\|\widetilde{\boldsymbol{\delta}}^{(t+1)}\|_2^2 + (1+\eta^{-1})\eta_0^2\|\boldsymbol{h}_t\|_2^2 \\
&= (1+\eta)\|\boldsymbol{\delta}^{(t)}\|_2^2 + (1+\eta^{-1})\eta_0^2\|\boldsymbol{h}_t\|_2^2 \\
&\quad + 2\eta_0(1+\eta)\Big\{\underbrace{\frac{\eta_0}{2}\|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2 - \langle\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}, \nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\rangle}_{\Pi}\Big\}.
\end{aligned} \tag{B.33}$$

To bound Π, we use the local strong convexity and smoothness properties of $\widehat{\mathscr{L}_\tau}(\cdot)$. From (B.24) and (B.26) we see that

$$\widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) - \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \langle\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(t)}\rangle \geq \frac{1-\chi}{2}\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2$$

and

$$\widehat{\mathscr{L}_\tau}(\widetilde{\boldsymbol{\mu}}^{(t+1)}) - \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) \leq \langle\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \widetilde{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}^{(t)}\rangle + \frac{1}{2}\|\widetilde{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}^{(t)}\|_2^2.$$

Together, these upper and lower bounds imply

$$\begin{aligned}
0 &\leq \widehat{\mathscr{L}_\tau}(\widetilde{\boldsymbol{\mu}}^{(t+1)}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) \\
&\leq \langle\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \widetilde{\boldsymbol{\mu}}^{(t+1)} - \widehat{\boldsymbol{\mu}}\rangle + \frac{1}{2}\|\widetilde{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}^{(t)}\|_2^2 - \frac{1-\chi}{2}\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 \\
&= \langle\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \widetilde{\boldsymbol{\mu}}^{(t+1)} - \widehat{\boldsymbol{\mu}}\rangle + \frac{\eta_0^2}{2}\|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2 - \frac{1-\chi}{2}\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 \\
&= \langle\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\rangle - \eta_0(1-\eta_0/2)\|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2 - \frac{1-\chi}{2}\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 \\
&\leq \langle\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\rangle - \frac{\eta_0}{2}\|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2 - \frac{1-\chi}{2}\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2.
\end{aligned}$$

Substituting this into (B.33) gives

$$\begin{aligned}\|\boldsymbol{\delta}^{(t+1)}\|_2^2 &\le \big\{1+\eta-\eta_0(1-\chi)(1+\eta)\big\}\|\boldsymbol{\delta}^{(t)}\|_2^2+(1+\eta^{-1})\eta_0^2\|\boldsymbol{h}_t\|_2^2\\ &= (1+\eta)\big\{1-(1-\chi)\eta_0\big\}\|\boldsymbol{\delta}^{(t)}\|_2^2+(1+\eta^{-1})\eta_0^2\|\boldsymbol{h}_t\|_2^2.\end{aligned}$$

Taking $\rho=\eta^2$ and $\eta=(1-\chi)\eta_0$, we conclude that

$$\|\boldsymbol{\delta}^{(t+1)}\|_2^2 \le (1-\rho)\|\boldsymbol{\delta}^{(t)}\|_2^2+(1+\eta^{-1})\eta_0^2\|\boldsymbol{h}_t\|_2^2,\ \ t=0,1,\ldots,T-1.$$

This recursive bound further implies

$$\|\boldsymbol{\mu}^{(T)}-\widehat{\boldsymbol{\mu}}\|_2^2 \le (1-\rho)^T\|\boldsymbol{\delta}^{(0)}\|_2^2+4\eta_0^2(1+\eta^{-1})T\frac{\tau^2}{(\varepsilon n)^2}\sum_{t=0}^{T-1}(1-\rho)^t\|\boldsymbol{g}_{T-1-t}\|_2^2. \tag{B.34}$$

Recall that $\sum_{t=0}^{T-1}(1-\rho)^t\|\boldsymbol{g}_{T-1-t}\|_2^2 \le 2\rho^{-1}d+3z$ on $\mathscr{E}_2(z)$. Conditioned on $\mathscr{E}_1$, the final iterate $\boldsymbol{\mu}^{(T)}$ satisfies the bound

$$\|\boldsymbol{\mu}^{(T)}-\widehat{\boldsymbol{\mu}}\|_2^2 \le (1-\rho)^T r_0^2+4\eta_0(\eta_0+(1-\chi)^{-1})(2\rho^{-1}d+3z)T\frac{\tau^2}{(\varepsilon n)^2}$$

with probability (over $\{\boldsymbol{g}_t\}_{t=0}^{T-1}$) at least $1-2e^{-z}$. This concludes the proof. □

B.4.3 Proof of Theorem 2.3.2

Recall that $m_q=\mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^q \ge \mathrm{tr}(\Sigma)^{q/2}$ and $\varepsilon \le 1$. Throughout the proof, we assume that (n,τ) satisfies $\tau \gtrsim m_q^{1/q}$ so that $m_q/\tau^q \lesssim 1$. For each $\boldsymbol{u}\in\mathbb{R}^d$, define independent random variables

$$S_{i,\boldsymbol{u}}=\frac{\psi_\tau(\|\boldsymbol{x}_i-\boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i-\boldsymbol{\mu}\|_2}\langle \boldsymbol{u},\boldsymbol{x}_i-\boldsymbol{\mu}\rangle,\ \ \text{for}\ \ i=1,\ldots,n.$$

From the proof of Theorem 2.2.3, we have $|\mathbb{E}S_{i,\boldsymbol{u}}| \le \|\boldsymbol{u}\|_2 \cdot b_\tau \le \|\boldsymbol{u}\|_2 \cdot m_q \tau^{1-q}$ for all $\boldsymbol{u} \in \mathbb{R}^d$. Given $0 < \varepsilon \le 1$, it follows from Corollary 2.3.2 with $z = \log n$ that

$$
\begin{aligned}
&\left| \sqrt{n} \langle \boldsymbol{u}/\|\boldsymbol{u}\|_\Sigma, \boldsymbol{\mu}^{(T)} - \boldsymbol{\mu} \rangle - \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}}}{\|\boldsymbol{u}\|_\Sigma} \right| \\
&\le \frac{\|\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_\Sigma} \left\| \sqrt{n}(\boldsymbol{\mu}^{(T)} - \boldsymbol{\mu}) - \frac{1}{\sqrt{n}} \sum_{i=1}^n (S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}}) \right\|_2 \\
&\lesssim \frac{\|\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_\Sigma} \left\{ \left(\sqrt{\mathrm{tr}(\Sigma) + \bar{\lambda} \log n} + \frac{\tau \log n}{\sqrt{n}} + \frac{m_q \sqrt{n}}{\tau^{q-1}} \right) \left(\frac{m_q}{\tau^q} + \sqrt{\frac{\log n}{n}} \right) \right. \\
&\qquad \left. + \frac{m_q \sqrt{n}}{\tau^{q-1}} + (d + \log n)^{1/2} (\log n)^{1/2} \frac{\tau}{\varepsilon \sqrt{n}} \right\} \\
&\lesssim \frac{\|\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_\Sigma} \underbrace{\left\{ \sqrt{\mathrm{tr}(\Sigma) + \bar{\lambda} \log n} \left(\frac{m_q}{\tau^q} + \sqrt{\frac{\log n}{n}} \right) + \frac{m_q \sqrt{n}}{\tau^{q-1}} + (d + \log n)^{1/2} (\log n)^{1/2} \frac{\tau}{\varepsilon \sqrt{n}} \right\}}_{R'_{n,\tau}} \\
&\le \frac{R'_{n,\tau}}{\underline{\lambda}^{1/2}}
\end{aligned}
$$

with probability at least $1 - 8n^{-1}$.

Combining this inequality with (B.8), (B.9) and in (B.10), we obtain that for any $x \in \mathbb{R}$,

$$
\begin{aligned}
&\mathbb{P}(\sqrt{n} \langle \boldsymbol{u}/\|\boldsymbol{u}\|_\Sigma, \boldsymbol{\mu}^{(T)} - \boldsymbol{\mu} \rangle \le x) \\
&\le \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{S_{i,\boldsymbol{u}} - \mathbb{E}S_{i,\boldsymbol{u}}}{\|\boldsymbol{u}\|_\Sigma} \le x + C_8 \frac{R'_{n,\tau}}{\underline{\lambda}^{1/2}} \right) + \frac{8}{n} \\
&\le \Phi(x) + \frac{C_8}{\sqrt{2\pi}} \frac{R'_{n,\tau}}{\underline{\lambda}^{1/2}} + C_9 \left(\frac{\nu_3}{\sqrt{n}} + \nu_q^{2/q} m_q^{1-2/q} \tau^{2-q} \right).
\end{aligned}
$$

A reversed inequality can be similarly obtained. Choosing $\tau \asymp m_q^{1/q} \{ \frac{\varepsilon n}{\sqrt{(d+\log n) \log n}} \}^{1/q}$, we obtain that

$$
\begin{aligned}
R'_{n,\tau} &\lesssim \sqrt{\mathrm{tr}(\Sigma) + \bar{\lambda} \log n} \left\{ \frac{\sqrt{(d + \log n) \log n}}{\varepsilon n} + \sqrt{\frac{\log n}{n}} \right\} + m_q^{1/q} \sqrt{n} \left\{ \frac{\sqrt{(d + \log n) \log n}}{\varepsilon n} \right\}^{1 - \frac{1}{q}} \\
&\lesssim m_q^{1/q} \left\{ \frac{\sqrt{(d + \log n) \log n}}{\varepsilon} \right\}^{1 - 1/q} n^{-(1/2 - 1/q)},
\end{aligned}
$$

where the last step follows from the inequality $m_q^{1/q} \geq \sqrt{\mathrm{tr}(\Sigma)}$ and $\tau \gtrsim m_q^{1/q} \gtrsim \sqrt{\mathrm{tr}(\Sigma)/n}$. Moreover, it follows that

$$v_q^{2/q} m_q^{1-2/q} \tau^{2-q} \lesssim v_q^{2/q} \left\{ \frac{\sqrt{(d+\log n)\log n}}{\varepsilon n} \right\}^{1-2/q}.$$

Combining these bounds and the two sides of the inequalities proves the claim. Note that our assumption $n \gtrsim \sqrt{(d+\log n)\log n}/\varepsilon$ guarantees the requirement $\tau \gtrsim m_q^{1/q}$ in the beginning of the proof. □

B.4.4 Proof of Proposition 2.3.2

To begin with, for any $T_0 \geq 2$ and $z > 0$, define the event

$$\mathscr{E}_3 = \mathscr{E}_3(z) = \left\{ \max_{0 \leq t \leq T_0 - 1} \|\boldsymbol{g}_t\|_2 \leq B_{T_0} \right\}$$

where $B_{T_0} := \sqrt{d} + \sqrt{2(\log T_0 + z)}$. Lemma B.4.2 ensures that $\mathbb{P}\{\mathscr{E}_3(z)\} \geq 1 - e^{-z}$. Moreover, we assume that the sample size n satisfies

$$n \geq \frac{4T^{1/2} B_{T_0}}{\varepsilon} \max \left[\frac{\tau\{2R_0 + (T_0+1)(2\tau + 1/2)\}}{\Delta}, \frac{2(e-1)}{4-e} T_0 \max \left\{ \frac{\tau\eta_0}{R_0}, \left(\frac{\tau\eta_0}{R_0} \right)^2 \right\} \right] \tag{B.35}$$

Recall that $\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} - \eta_0 \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) + \eta_0 \boldsymbol{h}_t$, where $\boldsymbol{h}_t = 2T^{1/2}\tau \boldsymbol{g}_t/(\varepsilon n)$ and $\eta_0 \in (0,1]$. In the following, we prove the result by conditioning on the event $\mathscr{E}_3$. By the definition of $\boldsymbol{h}_t$,

$$\max_{0 \leq t \leq T_0 - 1} \|\boldsymbol{h}_t\|_2 \leq e_{\mathrm{priv}} := \frac{2\tau T^{1/2} B_{T_0}}{\varepsilon n}. \tag{B.36}$$

From the smoothness property (B.26), we see that for each $t = 0, 1, \ldots, T_0 - 1$,

$$\begin{aligned}
\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t+1)}) &\leq \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) + \langle \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^{(t)} \rangle + \frac{1}{2} \|\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^{(t)}\|_2^2 \\
&= \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \eta_0 \langle \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \boldsymbol{h}_t \rangle + \frac{\eta_0^2}{2} \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \boldsymbol{h}_t\|_2^2 \\
&\leq \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \eta_0 (1 - \eta_0/2) \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2 \\
&\quad + \eta_0 (1 + \eta_0) \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2 \|\boldsymbol{h}_t\|_2 + \frac{\eta_0^2}{2} \|\boldsymbol{h}_t\|_2^2.
\end{aligned}$$

Using the bound (B.36), and the facts that $\|\nabla \widehat{\mathscr{L}_\tau}(\cdot)\|_2 \leq \tau$ and $\eta_0 \in (0, 1]$, we obtain

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t+1)}) \leq \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \frac{\eta_0}{2} \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2 + (2\tau + e_{\text{priv}}/2) e_{\text{priv}} \tag{B.37}$$

for all $t \leq T_0 - 1$. Moreover, the convexity of $\widehat{\mathscr{L}_\tau}$ implies

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) \leq \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) + \langle \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}} \rangle,$$

and hence

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t+1)}) \leq \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) + \underbrace{\langle \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}} \rangle - \frac{\eta_0}{2} \|\nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2}_{\Pi_t} + (2\tau + e_{\text{priv}}/2) e_{\text{priv}}.$$

To bound Π_t, note that

$$\begin{aligned}
\Pi_t &= \frac{1}{2\eta_0}\Big(2\eta_0\langle\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\rangle - \eta_0^2\|\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2\Big) \\
&= \frac{1}{2\eta_0}\Big(\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}} - \eta_0\nabla\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)})\|_2^2\Big) \\
&= \frac{1}{2\eta_0}\Big(\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}} - \eta_0\boldsymbol{h}_t\|_2^2\Big) \\
&= \frac{1}{2\eta_0}\Big(\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \eta_0^2\|\boldsymbol{h}_t\|_2^2 + 2\eta_0\langle\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}, \boldsymbol{h}_t\rangle\Big) \\
&\leq \frac{1}{2\eta_0}\Big(\|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \eta_0^2\|\boldsymbol{h}_t\|_2^2\Big) + \|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2\|\boldsymbol{h}_t\|_2,
\end{aligned}$$

where the last step is due the Cauchy-Schwarz inequality. Summing over $t = 0, \ldots, T_0 - 1$ gives

$$\begin{aligned}
&\sum_{t=0}^{T_0-1}\big\{\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}})\big\} \\
&\leq \frac{1}{2\eta_0}\big\{\|\boldsymbol{\mu}^{(0)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \|\boldsymbol{\mu}^{(T_0)} - \widehat{\boldsymbol{\mu}}\|_2^2\big\} + \sum_{t=0}^{T_0-1}\|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2\|\boldsymbol{h}_t\|_2 + T_0(2\tau + e_{\text{priv}}/2)e_{\text{priv}}.
\end{aligned} \tag{B.38}$$

On the other hand, (B.37) implies

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(T_0)}) \leq \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) + T_0(2\tau + e_{\text{priv}}/2)e_{\text{priv}} \text{ for all } t \leq T_0.$$

Therefore,

$$\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(T_0)}) \leq \frac{1}{T_0}\sum_{t=1}^{T_0}\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) + T_0(2\tau + e_{\text{priv}}/2)e_{\text{priv}}.$$

Combining this with (B.38) and Lemma B.4.3, we obtain

$$\begin{aligned}\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(T_0)}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) &\le \frac{1}{T_0}\sum_{t=1}^{T_0}\{\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}})\} + T_0(2\tau + e_{\text{priv}}/2)e_{\text{priv}} \\ &\le \frac{1}{2\eta_0 T_0}R_0^2 + 2e_{\text{priv}}R_0 + (T_0+1)(2\tau+e_{\text{priv}}/2)e_{\text{priv}} \\ &\le \Delta,\end{aligned}$$

provided that $T_0 \ge R_0^2/(\eta_0\Delta)$ and

$$n \ge \frac{4\{2R_0 + (T_0+1)(2\tau+1/2)\}B_{T_0}T^{1/2}\tau}{\Delta\varepsilon}.$$

Here we use the fact that $e_{\text{priv}} < 1$ under the sample size condition (B.35).

It remains to prove the second claim (2.25). Recall that conditioning on $\mathscr{E}_1$, $\widehat{\boldsymbol{\mu}} \in \Theta(r_0/2)$ and $\widehat{\mathscr{L}_\tau}(\cdot)$ satisfies (B.24). Define

$$\widehat{\Delta}_1 = \inf_{\{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\widehat{\boldsymbol{\mu}}\|_2 > r_0/2\}} \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}).$$

The definition implies that any $\boldsymbol{\theta}$ such that $\widehat{\mathscr{L}_\tau}(\boldsymbol{\theta}) < \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) + \widehat{\Delta}_1$ must satisfy $\|\boldsymbol{\theta} - \widehat{\boldsymbol{\mu}}\|_2 \le r_0/2$. By the convexity of $\widehat{\mathscr{L}_\tau}(\cdot)$, the infimum is achieved at some point $\widetilde{\boldsymbol{\mu}}$ such that $\|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}\|_2 = r_0/2$. Also, $\widetilde{\boldsymbol{\mu}} \in \Theta(r_0)$ due to the triangle inequality. Using (B.24) and conditioning on the event $\mathscr{E}_1$, we get

$$\widehat{\Delta}_1 = \widehat{\mathscr{L}_\tau}(\widetilde{\boldsymbol{\mu}}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) \ge \frac{1-\chi}{2}\|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}\|_2^2 = \frac{(1-\chi)r_0^2}{8}.$$

Taking $\Delta = (1-\chi)r_0^2/8$, we see that $\|\boldsymbol{\mu}^{(T_0)} - \widehat{\boldsymbol{\mu}}\|_2 \le r_0/2$ conditioned the event $\mathscr{E}_1 \cap \mathscr{E}_3$. By the triangle inequality, $\|\boldsymbol{\mu}^{(T_0)} - \boldsymbol{\mu}\|_2 \le r_0$, as claimed. □

B.4.5 Proof of Proposition 2.3.3

Recall that $\widehat{\boldsymbol{\mu}}$ is the non-private robust estimator defined in (2.1). Leveraging Theorem 2.2.1 and Lemma 2.2.1, we have

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \lesssim \bar{\lambda}^{1/2}\sqrt{\frac{\mathrm{r}(\Sigma)+z}{n}} + \frac{\tau z}{n} + \bar{\lambda}^{1/2}\frac{\sqrt{\mathrm{tr}(\Sigma)}}{\tau}$$

with probability at least $1-2e^{-z}$. Thus, the event $\{\widehat{\boldsymbol{\mu}} \in \Theta(\tau/4)\}$ holds with the same probability, provided that $\tau \gtrsim \sqrt{\mathrm{tr}(\Sigma)}$ and $n \gtrsim \mathrm{r}(\Sigma)+z$. Turning to the second part of the event $\mathscr{E}_1$, combining Lemma B.3.1 with Markov's inequality yields that with probability at least $1-e^{-z}$,

$$\inf_{\boldsymbol{\theta}\in\Theta(\tau/2)} \lambda_{\min}(\nabla^2 \widehat{\mathscr{L}_\tau}(\boldsymbol{\theta})) \geq 1 - \frac{4\mathrm{tr}(\Sigma)}{\tau^2} - \sqrt{\frac{z}{2n}} = 1-\chi.$$

Remark that χ is strictly less than 1 as long as $\tau \gtrsim \sqrt{\mathrm{tr}(\Sigma)}$ and $n \gtrsim z$. This proves the claim.

□

B.4.6 Proof of Proposition 2.3.5

By definition,

$$\left\|\widehat{\Sigma}_{\xi,\varepsilon} - \left(\widehat{\Sigma}_\xi + \frac{4\xi}{\varepsilon n}\mathbf{E}\right)\right\|_2 \leq \left\|\widehat{\Sigma}_\xi - \left(\widehat{\Sigma}_\xi + \frac{4\xi}{\varepsilon n}\mathbf{E}\right)\right\|_2,$$

which further implies

$$\|\widehat{\Sigma}_{\xi,\varepsilon} - \widehat{\Sigma}_\xi\|_2 \leq \left\|\widehat{\Sigma}_{\xi,\varepsilon} - \left(\widehat{\Sigma}_\xi + \frac{4\xi}{\varepsilon n}\mathbf{E}\right)\right\|_2 + \frac{4\xi}{\varepsilon n}\|\mathbf{E}\|_2 \leq \frac{8\xi}{\varepsilon n}\|\mathbf{E}\|_2.$$

Applying Lemma B.4.4 with $t = \sqrt{\log(4n)}$ we see that $\|\mathbf{E}\|_2 \lesssim \sqrt{d} + \sqrt{\log(4n)}$ with probability at least $1-n^{-1}$. Combining this bound with Proposition 2.2.1 proves the claimed result.

□

B.4.7 Proof of Proposition B.4.1

Note that $\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} - \eta_0 \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) + 2\eta_0 T^{1/2} \tau (\varepsilon n)^{-1} \boldsymbol{g}_t$ for $t = 0, 1, \ldots, T-1$, and $\boldsymbol{\mu}^{(0)} \in \Theta(r_0)$. Now assume that $\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2 \leq r_0$ for some $t \geq 0$. Proceeding via proof by contradiction, suppose $\|\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}\|_2 > r_0$. From Lemma B.4.1 we see that, conditioning on $\mathscr{E}_1$,

$$\frac{1-\chi}{4} r_0 \cdot \|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2 \leq \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t+1)}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}).$$

For the right-hand side, we have

$$\begin{aligned}
&\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t+1)}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) = \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t+1)}) - \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) + \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}) - \widehat{\mathscr{L}_\tau}(\widehat{\boldsymbol{\mu}}) \\
&\overset{\text{(i)}}{\leq} \langle \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^{(t)} \rangle + \frac{1}{2} \|\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^{(t)}\|_2^2 - \langle \nabla \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t)}), \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(t)} \rangle \\
&= \frac{1}{\eta_0} \langle \boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(t+1)}, \boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}} \rangle + \frac{1}{2} \|\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^{(t)}\|_2^2 + 2T^{1/2} \frac{\tau}{\varepsilon n} \langle \boldsymbol{g}_t, \boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}} \rangle \\
&= \frac{1}{2\eta_0} \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \frac{1}{2\eta_0} \|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \frac{1}{2\eta_0} \|\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^{(t)}\|_2^2 \\
&\qquad\qquad + \frac{1}{2} \|\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^{(t)}\|_2^2 + 2T^{1/2} \frac{\tau}{\varepsilon n} \langle \boldsymbol{g}_t, \boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}} \rangle \\
&\overset{\text{(ii)}}{\leq} \frac{1}{2\eta_0} \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \frac{1}{2\eta_0} \|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2^2 + 2T^{1/2} \frac{\tau}{\varepsilon n} \|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2 \cdot \|\boldsymbol{g}_t\|_2 \\
&\overset{\text{(iii)}}{\leq} \frac{1}{2\eta_0} \|\boldsymbol{\mu}^{(t)} - \widehat{\boldsymbol{\mu}}\|_2^2 - \frac{1}{2\eta_0} \|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2^2 + 2T^{1/2} B_T \frac{\tau}{\varepsilon n} \|\boldsymbol{\mu}^{(t+1)} - \widehat{\boldsymbol{\mu}}\|_2
\end{aligned}$$

where inequality (i) follows from the smoothness property (B.26), inequality (ii) holds if $\eta_0 \leq 1$, and inequality (iii) uses conditioning on $\mathscr{E}_2$. Provided $2T^{1/2} B_T \tau (\varepsilon n)^{-1} \leq \frac{1-\chi}{4} r_0$, combining the above lower and upper bounds on $\widehat{\mathscr{L}_\tau}(\boldsymbol{\mu}^{(t+1)}) - \widehat{\mathscr{L}_\tau}(\boldsymbol{\mu})$ yields

$$\|\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}\|_2 \leq \|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2 \leq r_0,$$

which leads to a contradiction. Therefore, starting from an initial value $\boldsymbol{\mu}^{(0)} \in \Theta(r_0)$, and conditioning on the event $\mathscr{E}_1 \cap \mathscr{E}_2$ with suitably chosen parameters, we must have $\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2 \leq r_0$

for all $t = 1, \ldots, T$. $\square$

Appendix C

Supplementary Material for Chapter 3

C.1 Proofs of Main Theorems

C.1.1 Supporting technical lemmas

We first introduce some basic notations which will be used throughout. Recall that quantile regression residuals are defined as $\varepsilon_i = Y_i - f_0(X_i)$ for $1 \le i \le n$ and $\varepsilon = Y - f_0(X)$. For any $f : [0,1]^d \to \mathbb{R}$, we define

$$Z(f)(X,\varepsilon) = \{Y - f(X)\}\mathbb{1}\{Y \le f(X)\} + \alpha f(X),$$

and denote $Z_i(f) := Z(f)(X_i,\varepsilon_i)$ for $1 \le i \le n$. Furthermore, we write

$$\omega_i = Z(f_0)(X_i,\varepsilon_i) - \alpha g_0(X_i). \tag{C.1}$$

Then, for any $\tau > 0$, we can express the empirical joint Huber loss (3.8) as

$$\widehat{\mathscr{R}}_\tau(f,g) = \frac{1}{n}\sum_{i=1}^{n} \ell_\tau(Z_i(f) - \alpha g(X_i))$$

for real-valued functions f, g on $[0,1]^d$. Also, note that

$$\omega_i = \varepsilon_i \mathbb{1}(\varepsilon_i \le 0) + \alpha f_0(X_i) - \alpha g_0(X_i) = \varepsilon_{i,-} - \mathbb{E}(\varepsilon_{i,-}|X_i),$$

where $\varepsilon_{i,-} = \varepsilon_i \mathbb{1}(\varepsilon_i \le 0)$. Throughout the proof, we assume that $\max(\|f_0\|_\infty, \|g_0\|_\infty) \le M_0$. For ease of notations, we write $\sum_{i=1}^n (W_i - \mathbb{E}W_i) = \sum_{i=1}^n (1-\mathbb{E})W_i$ for any sequence of random variables $\{W_i\}_{i=1}^n$.

Recall that $\mathscr{R}_\tau(\cdot,\cdot)$ represents the population joint loss function, which is the expectation of the empirical joint Huber loss function,

$$\mathscr{R}_\tau(f,g) = \mathbb{E}\widehat{\mathscr{R}}_\tau(f,g) = \mathbb{E}\ell_\tau(Z_i(f) - \alpha g(X_i))$$

for any fixed functions f and g. The following two lemmas establish both lower and upper bounds for the excess Huber risk under heavy-tailed noises and light-tailed noises, respectively.

Lemma C.1.1. Assume Condition 1 with $p \ge 2$ holds and let $\tau \ge c_4 = 2\max\{4M_0, (2\nu_p)^{1/p}\}$. Then, for any $f, g : [0,1]^d \to [-M_0, M_0]$, we have

$$\mathscr{R}_\tau(f,g) - \mathscr{R}_\tau(f,g_0) \ge \frac{\alpha^2}{4}\|g-g_0\|_2^2 - \alpha\|g-g_0\|_2\left\{\frac{\bar{p}}{2}\|f-f_0\|_4^2 + \frac{\nu_p}{(\tau/2)^{p-1}}\right\},$$

and

$$\mathscr{R}_\tau(f,g) - \mathscr{R}_\tau(f,g_0) \le \frac{\alpha^2}{2}\|g-g_0\|_2^2 + \alpha\|g-g_0\|_2\left\{\frac{\bar{p}}{2}\|f-f_0\|_4^2 + \frac{\nu_p}{(\tau/2)^{p-1}}\right\}.$$

Lemma C.1.2. Assume Condition 2 holds for some $\sigma_0 > 0$ and let

$$\tau \ge c_7 = 2\max\{4M_0, \sigma_0(\log 4)^{1/2}\}.$$

For any functions $f, g : [0,1]^d \to [-M_0, M_0]$, we have

$$\mathscr{R}_\tau(f,g) - \mathscr{R}_\tau(f,g_0) \ge \frac{\alpha^2}{4}\|g-g_0\|_2^2 - \alpha\|g-g_0\|_2\left(\frac{\bar{p}}{2}\|f-f_0\|_4^2 + c_{14}e^{-\tau^2/(2\sigma_0^2)}\right)$$

and

$$\mathscr{R}_\tau(f,g)-\mathscr{R}_\tau(f,g_0)\le\frac{\alpha^2}{2}\|g-g_0\|_2^2+\alpha\|g-g_0\|_2\left(\frac{\bar{p}}{2}\|f-f_0\|_4^2+c_{14}e^{-\tau^2/(2\sigma_0^2)}\right),$$

where $c_{14}=4M_0+2\sigma_0$.

For the truncated neural network function class $\mathscr{G}_n=\mathscr{F}_{\mathrm{DNN}}(d,L,N,M_0)$, define

$$\mathscr{G}_n(\eta)=\{g\in\mathscr{G}_n:\|g-g_0\|_2\le\eta\},\ \ \eta>0.$$

Moreover, for any function pair (f,g), we denote the difference of Huber losses as

$$h_{f,g}(X,\varepsilon)=\ell_\tau(Z(f)(X,\varepsilon)-\alpha g(X))-\ell_\tau(Z(f)(X,\varepsilon)-\alpha g_0(X)). \tag{C.2}$$

In order to obtain the convergence rate of the ES estimator $\widehat{g}_n$ given a generic QR estimate $\widehat{f}_n\in\mathscr{F}_n$, it is necessary to derive concentration inequalities for the supremum of local empirical processes that are of the form

$$\sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^n\left\{h_{f,g}(X_i,\varepsilon_i)-\mathbb{E}h_{f,g}(X_i,\varepsilon_i)\right\}\right|$$

for some $\eta>0$. To this end, by the fundamental theorem of calculus and the triangle inequality,

the supremum is upper bounded by a sum of three suprema, namely,

$$
\begin{aligned}
&\sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})h_{f,g}(X_i,\varepsilon_i)\right| \\
&=\sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\left\{\int_0^{\alpha\Delta_g(X_i)}\psi_\tau(\omega_i+Z_i(f)-Z_i(f_0)+t)\mathrm{d}t\right\}\right| \\
&\le\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\left\{\int_0^{\alpha\Delta_g(X_i)}\psi_\tau(\omega_i)\mathrm{d}t\right\}\right| \qquad \text{(C.3)}\\
&\quad+\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\left[\int_0^{\alpha\Delta_g(X_i)}\{\psi_\tau(\omega_i+t)-\psi_\tau(\omega_i)\}\mathrm{d}t\right]\right| \\
&\quad+\sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\left[\int_0^{\alpha\Delta_g(X_i)}\left\{\psi_\tau(\omega_i+Z_i(f)-Z_i(f_0)+t)-\psi_\tau(\omega_i+t)\right\}\mathrm{d}t\right]\right|,
\end{aligned}
$$

where $\Delta_g(X)=g_0(X)-g(X)$. The following three lemmas give concentration inequalities for the above three suprema. Recall that

$$
V_{n,\tau,\nu_p}=LN\sqrt{\frac{\log(LN)\log(n^2\tau\nu_p^{-1/p})}{n}}\ \text{ and }\ V_n=LN\sqrt{\frac{\log(LN)\log n}{n}}.
$$

Lemma C.1.3. Assume $\mathbb{E}(|\omega_i|^p|X_i)\le\nu_p<\infty$ almost surely (over X_i) for $p\ge 2$. Then, there exists a universal constant $c_{15}>0$ such that, for any $\eta\ge\max(\sqrt{\tau}V_{n,\tau,\nu_p},1/n), 0\le x\le n\eta^2/\tau$ and $\tau\ge\nu_p^{1/p}$,

$$
\mathbb{P}\left\{\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\Delta_g(X_i)\psi_\tau(\omega_i)\right|\ge c_{15}\cdot\eta(\nu_p^{1/p}+\sqrt{\tau})\left(V_{n,\tau,\nu_p}+\sqrt{\frac{x}{n}}\right)\right\}\le e^{-x}.
$$

Lemma C.1.4. There exists a universal constant $c_{16}>0$ such that for any $\tau>0,\eta\ge V_n$ and $0\le x\le n\eta^2$,

$$
\begin{aligned}
&\mathbb{P}\left[\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\left[\int_0^{\alpha\Delta_g(X_i)}\{\psi_\tau(\omega_i+t)-\psi_\tau(\omega_i)\}\mathrm{d}t\right]\right|\ge c_{16}\cdot\alpha^2\eta\left(V_n+\sqrt{\frac{x}{n}}\right)\right] \\
&\le e^{-x}.
\end{aligned}
$$

Lemma C.1.5. Write

$$W_n := \sqrt{\frac{\{\mathrm{Pdim}(\mathscr{F}_n) + (LN)^2 \log(LN)\} \log n}{n}}.$$

There exists a universal constant $c_{17} > 0$ such that for any $\tau > 0, \eta \geq W_n$ and $0 \leq x \leq n\eta^2$,

$$\mathbb{P}\left\{ \sup_{f \in \mathscr{F}_n} \sup_{g \in \mathscr{G}_n(\eta)} \left| \frac{1}{n} \sum_{i=1}^{n} (1 - \mathbb{E}) \left[\int_0^{\alpha \Delta_g(X_i)} \left\{ \psi_\tau(\omega_i + Z_i(f) - Z_i(f_0) + t) - \psi_\tau(\omega_i + t) \right\} \mathrm{d}t \right] \right| \right.$$
$$\left. \geq c_{17} \cdot \alpha \eta \left(W_n + \sqrt{\frac{x}{n}} \right) \right\} \leq e^{-x}.$$

Assuming that the random variables ω_i defined in (C.1) are sub-Gaussian, we can derive a more refined tail inequality for the supremum of local empirical processes.

Lemma C.1.6. Assume that ω_i satisfies

$$\mathbb{E}\left(e^{\omega_i^2/\sigma_0^2} | X_i\right) \leq 2 \quad \text{almost surely (over } X_i) \tag{C.4}$$

for some $\sigma_0 > 0$. Then, there exists a universal constant $c_{18} > 0$ such that for any $\eta \geq V_n$ and $0 \leq x \leq n\eta^2$, the following bound

$$\sup_{g \in \mathscr{G}_n(\eta)} \left| \frac{1}{n} \sum_{i=1}^{n} (1 - \mathbb{E}) \left\{ \int_0^{\alpha \Delta_g(X_i)} \psi_\tau(\omega_i) \mathrm{d}t \right\} \right| \leq c_{18} \cdot \alpha \sigma_0 \eta \left\{ V_n + e^{-\tau^2/(2\sigma_0^2)} + \sqrt{\frac{x}{n}} \right\}$$

holds with probability at least $1 - 3e^{-x}$.

To establish a convergence rate for the deep quantile regression estimator, we also require lower and upper bounds on the excess quantile risk, similar to the analysis of ES estimators. Recalling the definition of $\widehat{\mathscr{Q}}_\alpha(f)$ in (3.9), we define the population check loss function as

$$\mathscr{Q}_\alpha(f) = \mathbb{E}\widehat{\mathscr{Q}}_\alpha(f) = \mathbb{E}\rho_\alpha(Y_i - f(X_i))$$

for any $f : [0,1]^d \to \mathbb{R}$.

Lemma C.1.7. Assume Condition 3 holds. For any function $f : [0,1]^d \to [-M_0, M_0]$, the population check loss function satisfies

$$c_{19}\|f - f_0\|_2^2 \le \mathscr{Q}_\alpha(f) - \mathscr{Q}_\alpha(f_0) \le c_{20}\|f - f_0\|_2^2,$$

where $c_{19} = \min\{\underline{p}/(8M_0), \underline{p}^2/(32M_0 l_0)\}$ and $c_{20} = \bar{p}/2$.

Next, we write $\mathscr{F}_n = \mathscr{F}_{\mathrm{DNN}}(d, L, N, M_0)$, and for any $\delta > 0$,

$$\mathscr{F}_n(\delta) = \{f \in \mathscr{F}_n : \|f - f_0\|_2 \le \delta\}.$$

The next lemma characterizes the tail probabilities of the empirical quantile process.

Lemma C.1.8. There exists a universal constant $c_{21} > 0$ such that for any $\delta \ge V_n$ and $0 \le x \le n\delta^2$,

$$\mathbb{P}\left[\sup_{f \in \mathscr{F}_n(\delta)} \left|\frac{1}{n}\sum_{i=1}^n (1 - \mathbb{E})\big\{\rho_\alpha(Y_i - f(X_i)) - \rho_\alpha(Y_i - f_0(X_i))\big\}\right| \ge c_{21} \cdot \delta\left(V_n + \sqrt{\frac{x}{n}}\right)\right] \le e^{-x}.$$

C.1.2 Proof of Proposition 3.3.1

Following the proof of Lemma C.1.1, we can readily derive that provided $\tau \ge c_4$,

$$\mathscr{R}_\tau(f_0, g) - \mathscr{R}_\tau(f_0, g_0) \ge \frac{\alpha^2}{4}\|g - g_0\|_2^2 - \alpha\|g - g_0\|_2 \frac{v_p}{(\tau/2)^{p-1}}$$

for any real-valued function f with $\|f\|_\infty \le M_0$. Taking $g = g_{0,\tau}$, the claim follows immediately from the fact that $\mathscr{R}_\tau(f_0, g_{0,\tau}) - \mathscr{R}_\tau(f_0, g_0) \le 0$. □

C.1.3 Proof of Theorem 3.4.1

To begin with, denote for any $u \geq 1$ fixed that

$$\eta_* = c_5 \bigg\{ \eta_{\mathrm{s}} + \eta_{\mathrm{b}} + \eta_{\mathrm{a}} + \delta_{\mathrm{s}} + \eta_{\mathrm{opt}} + \delta_4^2 + (\nu_p^{1/p} + \sqrt{\tau}) \sqrt{\frac{u}{n}} \bigg\},$$

where c_5 is given by

$$c_5 = \max(\sqrt{24 \cdot 7}\bar{p}, \sqrt{28 \cdot 24}, 192c_{15}, 192c_{16}, 192c_{17}) \geq 1. \tag{C.5}$$

Here, c_{15}, c_{16} and c_{17} are defined in Lemma C.1.3, Lemma C.1.4 and Lemma C.1.5, respectively. For integers $j = 1, 2, \ldots$, define donut-shaped sets

$$\mathscr{D}_{n,j} := \mathscr{G}_n(2^j \eta_*/\alpha) \setminus \mathscr{G}_n(2^{j-1}\eta_*/\alpha) = \big\{ g \in \mathscr{G}_n : 2^{j-1}\eta_* < \alpha \|g - g_0\|_2 \leq 2^j \eta_* \big\}.$$

Recall the local function class $\mathscr{F}_0(\delta) = \{f \in \mathscr{F}_n : \|f - f_0\|_4 \leq \delta\}$ for any $\delta > 0$. Write

$$\begin{aligned} &\mathbb{P}\bigg\{ \exists f \in \mathscr{F}_0(\delta_4) \text{ such that } \sup_{g \in \mathscr{T}_n(\eta_{\mathrm{opt}}; f)} \alpha \|g - g_0\|_2 \geq \eta_* \bigg\} \\ &\qquad \leq \sum_{j=1}^{\infty} \mathbb{P}\big\{ \exists f \in \mathscr{F}_0(\delta_4) \text{ such that } \exists g \in \mathscr{T}_n(\delta_{\mathrm{opt}}; f) \cap \mathscr{D}_{n,j} \big\}. \end{aligned} \tag{C.6}$$

Then, it suffices to bound each probability on the right-hand side of (C.6). Conditioning on the event $\{f \in \mathscr{F}_0(\delta_4)\}$, Lemma C.1.1 implies that every $g \in \mathscr{D}_{n,j}$ satisfies

$$\begin{aligned} \frac{2^{2j-2}}{4} \eta_*^2 &\leq \mathscr{R}_\tau(f, g) - \mathscr{R}_\tau(f, g_0) + 2^j \eta_* \bigg(\frac{\bar{p}}{2} \delta_4^2 + \eta_{\mathrm{b}} \bigg) \\ &\leq \mathscr{R}_\tau(f, g) - \mathscr{R}_\tau(f, g_0) + 6\bar{p}^2 \delta_4^4 + 24\eta_{\mathrm{b}}^2 + \frac{2^{2j}}{48} \eta_*^2, \end{aligned} \tag{C.7}$$

where the last inequality follows from the basic inequalities $ab \leq 12a^2 + b^2/48$ and $(a+b)^2 \leq 2(a^2 + b^2)$ for any $a, b \in \mathbb{R}$.

We will now establish an upper bound for $\mathscr{R}_\tau(f,g)-\mathscr{R}_\tau(f,g_0)$, which appears on the right-hand side of inequality (C.7). The definition of η_{a} in (3.2) allows us to choose $g_n \in \mathscr{G}_n$ such that $\|g_n - g_0\|_2 \leq 2\eta_{\mathrm{a}}$. When we condition on the event $\{f \in \mathscr{F}_0(\delta_4), g \in \mathscr{T}_n(\eta_{\mathrm{opt}}; f) \cap \mathscr{D}_{n,j}\}$, it follows from the definition of $\mathscr{T}_n(\eta_{\mathrm{opt}}; f)$ that

$$\begin{aligned}
&\mathscr{R}_\tau(f,g)-\mathscr{R}_\tau(f,g_0)\\
&\leq \mathscr{R}_\tau(f,g)-\widehat{\mathscr{R}}_\tau(f,g)+\widehat{\mathscr{R}}_\tau(f,g_n)-\mathscr{R}_\tau(f,g_n)+\mathscr{R}_\tau(f,g_n)-\mathscr{R}_\tau(f,g_0)+\eta_{\mathrm{opt}}^2.
\end{aligned}$$

The upper bound in Lemma C.1.1 with $g = g_n$ implies

$$\begin{aligned}
\mathscr{R}_\tau(f,g_n)-\mathscr{R}_\tau(f,g_0) &\leq 2\alpha^2\eta_{\mathrm{a}}^2+\alpha\cdot\eta_{\mathrm{a}}\left(\frac{\bar{p}}{2}\delta_4^2+\eta_{\mathrm{b}}\right)\\
&\leq \frac{17}{8}\eta_{\mathrm{a}}^2+\bar{p}^2\delta_4^4+4\eta_{\mathrm{b}}^2,
\end{aligned}$$

which, combined with the earlier inequality, further yields

$$\begin{aligned}
\mathscr{R}_\tau(f,g)-\mathscr{R}_\tau(f,g_0) \leq{}& \mathscr{R}_\tau(f,g)-\widehat{\mathscr{R}}_\tau(f,g)+\widehat{\mathscr{R}}_\tau(f,g_n)-\mathscr{R}_\tau(f,g_n)\\
&+3\eta_{\mathrm{a}}^2+\bar{p}^2\delta_4^4+4\eta_{\mathrm{b}}^2+\eta_{\mathrm{opt}}^2. \qquad \text{(C.8)}
\end{aligned}$$

For any $f \in \mathscr{F}_n$ and $g \in \mathscr{G}_n$, recall the definition of $h_{f,g}(X,\varepsilon)$ in (C.2). Moreover, define

$$\Delta_n(f,g) = \frac{1}{n}\sum_{i=1}^{n}\left\{h_{f,g}(X_i,\varepsilon_i)-\mathbb{E}h_{f,g}(X_i,\varepsilon_i)\right\}, \qquad \text{(C.9)}$$

such that

$$\mathscr{R}_\tau(f,g)-\widehat{\mathscr{R}}_\tau(f,g)+\widehat{\mathscr{R}}_\tau(f,g_n)-\mathscr{R}_\tau(f,g_n) = \Delta_n(f,g_n)-\Delta_n(f,g).$$

Combining this with the bounds (C.7) and (C.8) yields

$$
\begin{aligned}
&\mathbb{P}\big\{\exists f \in \mathscr{F}_0(\delta_4) \text{ such that } \exists g \in \mathscr{T}_n(\eta_{\text{opt}}; f) \cap \mathscr{D}_{n,j}\big\} \\
&\le \mathbb{P}\Bigg\{ \exists f \in \mathscr{F}_0(\delta_4) \text{ and } \exists g \in \mathscr{D}_{n,j} \text{ such that} \\
&\qquad\qquad \Delta_n(f, g_n) - \Delta_n(f, g) \ge \frac{2^{2j}}{24}\eta_*^2 - 3\eta_{\text{a}}^2 - 7\bar{p}^2\delta_4^4 - 28\eta_{\text{b}}^2 - \eta_{\text{opt}}^2 \Bigg\} \\
&\overset{\text{(i)}}{\le} \mathbb{P}\Bigg\{ \exists f \in \mathscr{F}_0(\delta_4) \text{ and } \exists g \in \mathscr{D}_{n,j} \text{ such that } \Delta_n(f, g_n) - \Delta_n(f, g) \ge \frac{2^{2j}}{32}\eta_*^2 \Bigg\} \\
&\overset{\text{(ii)}}{\le} \mathbb{P}\Bigg\{ \sup_{f \in \mathscr{F}_n} \sup_{g \in \mathscr{G}_n(2^j\eta_*/\alpha)} |\Delta_n(f, g)| \ge \frac{1}{64} 2^{2j}\eta_*^2 \Bigg\},
\end{aligned} \tag{C.10}
$$

where the second inequality (i) follows from the definition of c_5 in (C.5) that

$$
\begin{aligned}
3\eta_{\text{a}}^2 + 7\bar{p}^2\delta_4^4 + 28\eta_{\text{b}}^2 + \eta_{\text{opt}}^2 &\le \frac{1}{24}c_5^2(\eta_{\text{a}}^2 + \delta_4^4 + \eta_{\text{b}}^2 + \eta_{\text{opt}}^2) \\
&\le \frac{2^{2j}}{96}\eta_*^2 \quad \text{for } j \ge 1,
\end{aligned}
$$

and the last inequality (ii) follows from the choice of g_n, which satisfies

$$
\|g_n - g_0\|_2 \le 2\eta_{\text{a}} \le 2^j\eta_*/\alpha
$$

for any $j \ge 1$.

So, the key task is to derive a concentration inequality for the supremum of the empirical

process $\{\Delta_n(f,g) : f \in \mathscr{F}_n, g \in \mathscr{G}_n(2^j\eta_*/\alpha)\}$. From the bound (C.3), we can see that

$$
\begin{aligned}
&\mathbb{P}\left\{ \sup_{f\in\mathscr{F}_n} \sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)} |\Delta_n(f,g)| \geq \frac{1}{64}2^{2j}\eta_*^2 \right\} \\
&\leq \mathbb{P}\left[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)} \left| \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\left\{ \int_0^{\alpha\Delta_g(X_i)} \psi_\tau(\omega_i)\mathrm{d}t \right\} \right| \geq \frac{1}{192}2^{2j}\eta_*^2 \right] \\
&+ \mathbb{P}\left\{ \sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)} \left| \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\left[\int_0^{\alpha\Delta_g(X_i)} \left\{ \psi_\tau(\omega_i+t) - \psi_\tau(\omega_i) \right\}\mathrm{d}t \right] \right| \geq \frac{1}{192}2^{2j}\eta_*^2 \right\} \\
&+ \mathbb{P}\left\{ \sup_{f\in\mathscr{F}_n} \sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)} \left| \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\left[\int_0^{\alpha\Delta_g(X_i)} \left\{ \psi_\tau(\omega_i+Z_i(f)-Z_i(f_0)+t) - \psi_\tau(\omega_i+t) \right\}\mathrm{d}t \right] \right| \right. \\
&\qquad\qquad \left. \geq \frac{1}{192}2^{2j}\eta_*^2 \right\} \\
&=: \mathrm{P}_1 + \mathrm{P}_2 + \mathrm{P}_2. \qquad \text{(C.11)}
\end{aligned}
$$

We proceed to bound the three probabilities $\mathrm{P}_1, \mathrm{P}_2$ and P_3, separately. We choose $\eta = 2^j\eta_*/\alpha$ and $x = 2^{2j}u$ for the given $u \geq 1$ to apply Lemma C.1.3. Note that $\eta \geq \max((\nu_p^{1/p} + \sqrt{\tau})V_{n,\tau,\nu_p}, 1/n)$ and $0 \leq x \leq n\eta^2/\tau$ as $0 < \alpha < 1$ and $c_5 > 1$. Furthermore, $\tau \geq c_4$ implies $\tau/\nu_p^{1/p} \geq 1$. Therefore, applying Lemma C.1.3 gives

$$
\begin{aligned}
&\mathbb{P}\left[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)} \left| \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\left\{ \int_0^{\alpha\Delta_g(X_i)} \psi_\tau(\omega_i)\mathrm{d}t \right\} \right| \geq \frac{c_{15}}{c_5}2^{2j}\eta_*^2 \right] \\
&\leq \mathbb{P}\left[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)} \left| \frac{\alpha}{n}\sum_{i=1}^n (1-\mathbb{E})\left\{ \psi_\tau(\omega_i)\Delta_g(X_i) \right\} \right| \geq c_{15}\cdot\alpha\eta(\nu_p^{1/p}+\sqrt{\tau})\left(V_{n,\tau,\nu_p} + \sqrt{\frac{x}{n}} \right) \right] \\
&\leq e^{-x} = e^{-2^{2j}u}.
\end{aligned}
$$

Here, we remark that the choice of c_5 in (C.5) is such that $c_{15}/c_5 \leq 1/192$. Thus, the above probability bound implies

$$\mathrm{P}_1 \leq \exp(-2^{2j}u).$$

Similarly, for $\eta = 2^j\eta_*/\alpha$ and $x = 2^{2j}u$, it follows that $\eta \geq \max(W_n, V_n)$ and $x \leq n\eta^2$. Combin-

ing Lemma C.1.4, Lemma C.1.5 with the choice of c_5 in (C.5) yields

$$\mathrm{P}_2 \le \exp(-2^{2j}u) \quad \text{and} \quad \mathrm{P}_3 \le \exp(-2^{2j}u).$$

Together, the above bounds on $\mathrm{P}_1, \mathrm{P}_2$ and P_3, (C.6), (C.10) and (C.11) imply

$$\begin{aligned}
\mathbb{P}\left\{ \exists f \in \mathscr{F}_0(\delta_4) \text{ such that } \sup_{g \in \mathscr{T}_n(\eta_{\mathrm{opt}}; f)} \|g - g_0\|_2 \ge \eta_* \right\} &\le \sum_{j=1}^{\infty} 3\exp(-2^{2j}u) \\
&\le \sum_{j=1}^{\infty} 3e^{-ju} \\
&= 3(1-e^{-u^2})^{-1}e^{-u} \\
&\le 3(1-e^{-1})^{-1}e^{-u},
\end{aligned}$$

where the last inequality uses the fact that $u \ge 1$. This completes the proof. □

C.1.4 Proof of Theorem 3.4.2

The proof employs the truncation argument as in Kuchibhotla and Patra (2022) and Fan et al. (2022), and the peeling argument as in the proof of Theorem 3.4.1.

For any $u \ge 1$, define

$$\eta_* := c_6 \cdot \sqrt{u}(\eta_{\mathrm{s}} + \eta_{\mathrm{a}} + \delta_{\mathrm{s}} + \delta_4^2 + \eta_{\mathrm{opt}}),$$

where c_6 is given by

$$c_6 = \max\left(\sqrt{24 \cdot 4}\bar{p}, \sqrt{72}, 4 \cdot 192c_{15}, 2 \cdot 192c_{16}, 2 \cdot 192c_{17}\right) \ge 4. \tag{C.12}$$

We note that it is sufficient to consider the case where $u \le n$ and $\nu_p \le n^p$. Otherwise, $\eta_* \gtrsim 1$, so that the deviation bound becomes trivial due to the uniform bounded property of g_0 and $\mathscr{G}_n$. Denote $\mathscr{R}(f,g) = \mathbb{E}\widehat{\mathscr{R}}(f,g)$ for any f, g, where $\widehat{\mathscr{R}}$ is given in (3.8), and define $\mathscr{D}_{n,j} =$

$\mathscr{G}_n(2^j\eta_*/\alpha) \setminus \mathscr{G}_n(2^{j-1}\eta_*/\alpha)$ for any $j \geq 1$. Taking $\tau = \infty$ in Lemma C.1.1, every $g \in \mathscr{D}_{n,j}$ satisfies

$$\begin{aligned}\frac{2^{2j-2}}{4}\eta_*^2 &\leq \mathscr{R}(f,g) - \mathscr{R}(f,g_0) + 2^j\eta_* \cdot \frac{\bar{p}}{2}\delta_4^2 \\ &\leq \mathscr{R}(f,g) - \mathscr{R}(f,g_0) + 3\bar{p}^2\delta_4^4 + \frac{2^{2j}}{48}\eta_*^2, \end{aligned} \tag{C.13}$$

conditioning on the event $\{f \in \mathscr{F}_0(\delta_4)\}$. Choose $g_n \in \mathscr{G}_n$ satisfying $\|g_n - g_0\|_2 \leq 2\eta_{\mathrm{a}}$, which is possible by the definition of η_{a}. Conditioning on the event $\{f \in \mathscr{F}_0(\delta_4)$ and $g \in \mathscr{T}_n(\eta_{\mathrm{opt}}) \cap \mathscr{D}_{n,j}\}$, we have

$$\begin{aligned}&\mathscr{R}(f,g) - \mathscr{R}(f,g_0) \\ &\leq \mathscr{R}(f,g) - \widehat{\mathscr{R}}(f,g) + \widehat{\mathscr{R}}(f,g_n) - \mathscr{R}(f,g_n) + \mathscr{R}(f,g_n) - \mathscr{R}(f,g_0) + \eta_{\mathrm{opt}}^2 \\ &\leq \mathscr{R}(f,g) - \widehat{\mathscr{R}}(f,g) + \widehat{\mathscr{R}}(f,g_n) - \mathscr{R}(f,g_n) + 2\alpha^2\eta_{\mathrm{a}}^2 + \alpha\eta_{\mathrm{a}} \cdot \frac{\bar{p}}{2}\delta_4^2 + \eta_{\mathrm{opt}}^2 \\ &\leq \mathscr{R}(f,g) - \widehat{\mathscr{R}}(f,g) + \widehat{\mathscr{R}}(f,g_n) - \mathscr{R}(f,g_n) + \frac{33}{16}\alpha^2\eta_{\mathrm{a}}^2 + \bar{p}^2\delta_4^4 + \eta_{\mathrm{opt}}^2,\end{aligned}$$

where the second inequality follows from the upper bound in Lemma C.1.1. Combining this bound with (C.13) gives

$$\frac{2^{2j}}{24}\eta_*^2 \leq \mathscr{R}(f,g) - \widehat{\mathscr{R}}(f,g) + \widehat{\mathscr{R}}(f,g_n) - \mathscr{R}(f,g_n) + 4\bar{p}^2\delta_4^2 + 3\eta_{\mathrm{a}}^2 + \eta_{\mathrm{opt}}^2,$$

conditioning on the same event. From the choice of c_6 in (C.12), we have

$$4\bar{p}^2\delta_4^2 + 3\eta_{\mathrm{a}}^2 + \eta_{\mathrm{opt}}^2 \leq \frac{2^{2j}}{96}\eta_*^2.$$

Then, by employing the peeling argument and following a similar line of reasoning that leads

to (C.10), we can obtain

$$\mathbb{P}\Big\{\exists f \in \mathscr{F}_0(\delta_4) \text{ such that } \sup_{g\in\mathscr{T}_n(\eta_{\text{opt}};f)} \alpha\|g-g_0\|_2 \geq \eta_*\Big\}$$
$$\leq \sum_{j=1}^{\infty} \mathbb{P}\Big\{\sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)} |\Delta_n(f,g)| \geq \frac{1}{64}2^{2j}\eta_*^2\Big\}, \tag{C.14}$$

where $\Delta_n(f,g)$ is defined as

$$\Delta_n(f,g) = \frac{1}{2n}\sum_{i=1}^{n}(1-\mathbb{E})[\{Z_i(f)-\alpha g(X_i)\}^2 - \{Z_i(f)-\alpha g_0(X_i)\}^2].$$

The bound (C.3) with $\tau=\infty$ gives

$$\begin{aligned}
&\mathbb{P}\Big\{\sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)} |\Delta_n(f,g)| \geq \frac{1}{64}2^{2j}\eta_*^2\Big\}\\
&\leq \mathbb{P}\Big[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)}\Big|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\Big\{\int_0^{\alpha\Delta_g(X_i)}\omega_i \mathrm{d}t\Big\}\Big| \geq \frac{1}{192}2^{2j}\eta_*^2\Big]\\
&+\mathbb{P}\Big[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)}\Big|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\Big\{\int_0^{\alpha\Delta_g(X_i)} t\mathrm{d}t\Big\}\Big| \geq \frac{1}{192}2^{2j}\eta_*^2\Big]\\
&+\mathbb{P}\Big\{\sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)}\Big|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\Big[\int_0^{\alpha\Delta_g(X_i)}\{Z_i(f)-Z_i(f_0)\}\mathrm{d}t\Big]\Big| \geq \frac{1}{192}2^{2j}\eta_*^2\Big\}\\
&=: \mathrm{P}_1+\mathrm{P}_2+\mathrm{P}_2.
\end{aligned} \tag{C.15}$$

For $\eta = 2^j\eta_*/\alpha$ and $x = 2^{2j}nuV_n^2$, it follows that $\eta \geq V_n$ and $x \leq n\eta^2$. By Lemma C.1.4, we have

$$\begin{aligned}
&\mathbb{P}\Big[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)}\Big|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\Big\{\int_0^{\alpha\Delta_g(X_i)} t\mathrm{d}t\Big\}\Big| \geq \frac{2c_{16}}{c_6}2^{2j}\eta_*^2\Big]\\
&\leq \mathbb{P}\Big[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)}\Big|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\Big\{\int_0^{\alpha\Delta_g(X_i)} t\mathrm{d}t\Big\}\Big| \geq c_{16}\cdot\alpha\eta\left(V_n+\sqrt{\frac{x}{n}}\right)\Big]\\
&\leq \exp(-2^{2j}nuV_n^2).
\end{aligned}$$

By the definition of c_6 in (C.12), we have $2c_{16}/c_6 \le 1/192$ so that $\mathrm{P}_2 \le \exp(-2^{2j} n u V_n^2)$. Similarly, applying Lemma C.1.5 yields $\mathrm{P}_3 \le \exp(-2^{2j} n u V_n^2)$.

We next derive an upper bound of the probability P_1. Remark that

$$\begin{aligned}
&\sup_{g \in \mathscr{G}_n(2^j \eta_*/\alpha)} \left| \frac{1}{n} \sum_{i=1}^n (1-\mathbb{E}) \big\{ \alpha \omega_i \Delta_g(X_i) \big\} \right| \\
&\quad \le \sup_{g \in \mathscr{G}_n(2^j \eta_*/\alpha)} \left| \frac{1}{n} \sum_{i=1}^n (1-\mathbb{E}) \big\{ \alpha \psi_{B_j}(\omega_i) \Delta_g(X_i) \big\} \right| \\
&\quad + \sup_{g \in \mathscr{G}_n(2^j \eta_*/\alpha)} \left| \frac{1}{n} \sum_{i=1}^n (1-\mathbb{E}) \big\{ \alpha \omega_i \mathbb{1}(|\omega_i| > B_j) \Delta_g(X_i) \big\} \right|, \qquad \text{(C.16)}
\end{aligned}$$

where we choose $B_j = u \nu_p^{1/p} V_n^{-2/p}$. Given our assumption that $u \ge 1$ and $V_n \le 1$, it follows that $B_j / \nu_p^{1/p} \ge 1$. Furthermore, we only consider the case $u \le n$ and $\nu_p \le n^p$, implying $B_j / \nu_p^{1/p} \le n^3$. Thus, η_* satisfies

$$\begin{aligned}
\eta_* \ge 4u \big(\nu_p^{1/p} V_n + \nu_p^{1/2p} V_n^{1-1/p} \big) &\ge LN (\nu_p^{1/p} + \sqrt{B_j}) \sqrt{\frac{(NL)^2 \log(n B_j \nu_p^{-1/p})}{n}} \\
&=: (\nu_p^{1/p} + \sqrt{B_j}) V_{n, B_j, \nu_p}.
\end{aligned}$$

Choose $\tau = B_j, \eta = 2^j \eta_* / \alpha$ and $x = 2^{2j} n V_n^2$, which satisfy $\tau \cdot x \le n \eta^2$. From Lemma C.1.3 it follows that

$$\begin{aligned}
&\mathbb{P} \left[\sup_{g \in \mathscr{G}_n(2^j \eta_*/\alpha)} \left| \frac{1}{n} \sum_{i=1}^n (1-\mathbb{E}) \big\{ \alpha \psi_{B_j}(\omega_i) \Delta_g(X_i) \big\} \right| \ge \frac{2c_{15}}{c_6} 2^{2j} \eta_*^2 \right] \\
&\le \mathbb{P} \left[\sup_{g \in \mathscr{G}_n(2^j \eta_*)} \left| \frac{1}{n} \sum_{i=1}^n (1-\mathbb{E}) \big\{ \alpha \psi_{B_j}(\omega_i) \Delta_g(X_i) \big\} \right| \ge \alpha \cdot c_{15} \eta (\nu_p^{1/p} + \sqrt{B_j}) \left(V_{n, B_j, \nu_p} + \sqrt{\frac{x}{n}} \right) \right] \\
&\le e^{-2^{2j} n V_n^2}.
\end{aligned}$$

Thus, from the choice of c_6, the above probability bound implies that

$$\mathbb{P}\left[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\big\{\alpha\psi_{B_j}(\omega_i)\Delta_g(X_i)\big\}\right|\geq\frac{1}{2\cdot 192}2^{2j}\eta_*^2\right]\leq e^{-2^{2j}nV_n^2}. \tag{C.17}$$

Turning to the second term on the right-hand side of (C.16), we apply Markov's inequality to obtain that for any $y>0$,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\big\{\alpha\omega_i\mathbb{1}(|\omega_i|>B_j)\Delta_g(X_i)\big\}\right|>y\right]\\
&\leq\frac{1}{y}\mathbb{E}\left[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\big\{\alpha\omega_i\mathbb{1}(|\omega_i|>B_j)\Delta_g(X_i)\big\}\right|\right]\\
&\leq\frac{4\alpha M_0}{y}\mathbb{E}\{|\omega_i|\mathbb{1}(|\omega_i|>B_j)\},
\end{aligned}$$

where the last inequality follows from the uniform boundedness of $\mathscr{G}_n$ and g_0. Furthermore,

$$\mathbb{E}\{|\omega_i|\mathbb{1}(|\omega_i|>B_j)\}\leq\frac{\mathbb{E}(|\omega_i|^p)}{B_j^{p-1}}\leq\frac{\nu_p}{B_j^{p-1}}.$$

Combining this expectation bound with $y=2^{2j}\eta_*^2/(2\cdot 192)$ in the earlier bound gives

$$\begin{aligned}
\mathbb{P}\left[\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\big\{\alpha\omega_i\mathbb{1}(|\omega_i|>B_j)\Delta_g(X_i)\big\}\right|>\frac{1}{2\cdot 192}2^{2j}\eta_*^2\right]&\leq\frac{8\cdot 192M_0\nu_p}{2^{2j}\eta_*^2B_j^{p-1}}\\
&\lesssim\frac{1}{u^p2^{2j}},
\end{aligned}$$

where the last inequality follows from the choice of B_j, which satisfies

$$B_j^{p-1}\eta_*^2\geq u^{p-1}\nu_p^{1-1/p}V_n^{-2+2/p}u\nu_p^{1/p}V_n^{2-2/p}=u^p\nu_p.$$

Together, the above probability bound, (C.17) and the choice of c_6 yield

$$\mathrm{P}_1 \lesssim \exp(-2^{2j} n V_n^2) + \frac{1}{u^p 2^{2j}}.$$

Finally, it follows from (C.14), (C.15) and the upper bounds on $\mathrm{P}_1, \mathrm{P}_2$ and P_3 that

$$\begin{aligned}
&\mathbb{P}\left\{ \exists f \in \mathscr{F}_0(\delta_4) \text{ such that } \sup_{g \in \mathscr{T}_n(\eta_{\mathrm{opt}}; f)} \alpha \|g - g_0\|_2 \geq \eta_* \right\} \\
&\lesssim \sum_{j=1}^{\infty} \left\{ \exp(-2^{2j} n u^2 V_n^2) + \exp(-2^{2j} n V_n^2) + \frac{1}{u^p 2^{2j}} \right\} \\
&\lesssim e^{-n V_n^2} + \frac{1}{u^p},
\end{aligned}$$

where the second inequality follows from the fact that $u \geq 1$. This proves the claim. □

C.1.5 Proof of Theorem 3.4.3

For any $u \geq 1$, denote

$$\eta_* = c_8 \left(\eta_{\mathrm{s}} + \eta_{\mathrm{b}} + \eta_{\mathrm{a}} + \delta_{\mathrm{s}} + \eta_{\mathrm{opt}} + \delta_4^2 + \sigma_0 \sqrt{\frac{u}{n}} \right),$$

where c_8 is given by

$$c_8 = \max(\sqrt{24 \cdot 7 \bar{p}}, \sqrt{28 \cdot 24}, 192 c_{16}, 192 c_{17}, 192 c_{18}) \geq 1.$$

Recall the definition of notations $\mathscr{F}_0(\delta)$ and Δ_n in the proof of Theorem 3.4.1. By employing the peeling argument and following a similar line of reasoning that leads to (C.10) in conjunction

with Lemma C.1.2 and the definition of η_*, it can be shown that

$$\mathbb{P}\left\{\exists f \in \mathscr{F}_0(\delta_4) \text{ such that } \sup_{g\in\mathscr{T}_n(\eta_{\text{opt}};f)} \alpha\|g-g_0\|_2 \geq \eta_*\right\} \leq \mathbb{P}\left\{\sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)} |\Delta_n(f,g)| \geq \frac{1}{64}2^{2j}\eta_*^2\right\}, \tag{C.18}$$

where Δ_n is defined in (C.9). Moreover, we have for each $j \geq 1$ that

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)} |\Delta_n(f,g)| \geq \frac{1}{64}2^{2j}\eta_*^2\right\} \\
&\leq \mathbb{P}\left[\sup_{g\in\mathscr{G}_n(2^j\eta_*)}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\left\{\int_0^{\alpha\Delta_g(X_i)}\psi_\tau(\omega_i)\mathrm{d}t\right\}\right| \geq \frac{1}{192}2^{2j}\eta_*^2\right] \\
&\quad + \mathbb{P}\left\{\sup_{g\in\mathscr{G}_n(2^j\eta_*)}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\left[\int_0^{\alpha\Delta_g(X_i)}\left\{\psi_\tau(\omega_i+t)-\psi_\tau(\omega_i)\right\}\mathrm{d}t\right]\right| \geq \frac{1}{192}2^{2j}\eta_*^2\right\} \\
&\quad + \mathbb{P}\left\{\sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\left[\int_0^{\alpha\Delta_g(X_i)}\left\{\psi_\tau(\omega_i+Z_i(f)-Z_i(f_0)+t)-\psi_\tau(\omega_i+t)\right\}\mathrm{d}t\right]\right| \right. \\
&\qquad\qquad \left. \geq \frac{1}{128}2^{2j}\eta_*^2\right\} \\
&=: \mathrm{P}_1+\mathrm{P}_2+\mathrm{P}_3.
\end{aligned}$$

To bound P_1, we choose $\eta = 2^j\eta_*/\alpha$ and $x = 2^{2j}u$. Then, $\eta \geq V_n$, $0 \leq x \leq n\eta^2$ and $V_n + e^{-\tau^2/(2\sigma_0^2)} + \sqrt{x/n} \leq 2^j\eta_*/c_8$. Thus, applying Lemma C.1.6 yields

$$\mathbb{P}\left\{\sup_{g\in\mathscr{G}_n(2^j\eta_*/\alpha)}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\left\{\int_0^{\alpha\Delta_g(X_i)}\psi_\tau(\omega_i)\mathrm{d}t\right\}\right| \geq \frac{c_{17}}{c_8}2^{2j}\eta_*^2\right\} \leq e^{-2^{2j}u^2},$$

which, combined with the choice of c_8, further implies

$$\mathrm{P}_1 \leq e^{-2^{2j}u}.$$

Moreover, for the same choice of η and x, Lemma C.1.4 and Lemma C.1.5 imply that $\mathrm{P}_2 \leq e^{-2^{2j}u}$

and $\mathrm{P}_2 \leq e^{-2^{2j}u}$, respectively.

Combining the upper bounds on $\mathrm{P}_1, \mathrm{P}_2$ and P_3 with (C.18) implies

$$\begin{aligned}\mathbb{P}\left\{\exists f \in \mathscr{F}_0(\delta_4) \text{ such that } \sup_{g \in \mathscr{T}_n(\eta_{\mathrm{opt}}; f)} \|g - g_0\|_2 \geq \eta_*\right\} &\leq 3\sum_{j=1}^{\infty} e^{-2^{2j}u} \\ &\leq 3\sum_{j=1}^{\infty} e^{-ju} \\ &\leq 3(1-e^{-1})^{-1}e^{-u},\end{aligned}$$

which completes the proof. □

C.1.6 Proof of Theorem 3.4.4

Following a similar line to the proof of Theorem 3.4.1, we start with the peeling argument. To begin with, let $\delta_* = c_9(\delta_{\mathrm{s}} + \delta_{\mathrm{a}} + \delta_{\mathrm{opt}} + \sqrt{u/n})$ for given $u \geq 1$, where c_9 is given by

$$c_9 = \max\{(\sqrt{8c_{20}/c_{19}}, \sqrt{2/c_{19}}, 16c_{21}/c_{19}\} \geq 1. \tag{C.19}$$

Here, c_{19} and c_{20} are given in Lemma C.1.7 and c_{21} is given in Lemma C.1.8. We then define the donut-shaped sets for integers $j = 1, 2, \ldots$ as

$$\mathscr{D}_{n,j} := \mathscr{F}_n(2^j\delta_*) \setminus \mathscr{F}_n(2^{j-1}\delta_*) = \{f \in \mathscr{F}_n : 2^{j-1}\delta_* < \|f - f_0\|_2 \leq 2^j\delta_*\},$$

so that we can write

$$\mathbb{P}\left\{\exists f \in \mathscr{S}_n(\delta_{\mathrm{opt}}) \text{ such that } \|f - f_0\|_2 \geq \delta_*\right\} \leq \sum_{j=1}^{\infty} \mathbb{P}\left\{\exists f \in \mathscr{S}_n(\delta_{\mathrm{opt}}) \cap \mathscr{D}_{n,j}\right\}. \tag{C.20}$$

Therefore, it reduces to bounding each probability $\mathbb{P}\{\exists f \in \mathscr{S}_n(\delta_{\mathrm{opt}}) \cap \mathscr{D}_{n,j}\}$ separately. Following Lemma C.1.7, any $f \in \mathscr{D}_{n,j}$ satisfies

$$c_{19} 2^{2j-2} \delta_*^2 \leq c_{19} \|f - f_0\|_2^2 \leq \mathscr{Q}_\alpha(f) - \mathscr{Q}_\alpha(f_0). \tag{C.21}$$

We next derive an upper bound of the right-hand side of (C.21). By the definition of δ_{a}, there exists $f_n \in \mathscr{F}_n$ such that $\|f_n - f_0\|_2 \leq 2\delta_{\mathrm{a}}$. Now, for any $f \in \mathscr{S}_n(\delta_{\mathrm{opt}}) \cap \mathscr{D}_{n,j}$, we have

$$\begin{aligned}
&\mathscr{Q}_\alpha(f) - \mathscr{Q}_\alpha(f_0) \\
&= \mathscr{Q}_\alpha(f) - \widehat{\mathscr{Q}}_\alpha(f) + \widehat{\mathscr{Q}}_\alpha(f) - \widehat{\mathscr{Q}}_\alpha(f_n) + \widehat{\mathscr{Q}}_\alpha(f_n) - \mathscr{Q}_\alpha(f_n) + \mathscr{Q}_\alpha(f_n) - \mathscr{Q}_\alpha(f_0) \\
&\leq \mathscr{Q}_\alpha(f) - \widehat{\mathscr{Q}}_\alpha(f) + \widehat{\mathscr{Q}}_\alpha(f_n) - \mathscr{Q}_\alpha(f_n) + \mathscr{Q}_\alpha(f_n) - \mathscr{Q}_\alpha(f_0) + \delta_{\mathrm{opt}}^2,
\end{aligned}$$

where the last line follows from the definition of $\mathscr{S}_n(\delta_{\mathrm{opt}})$. By Lemma C.1.7, it follows that $\mathscr{Q}_\alpha(f_n) - \mathscr{Q}_\alpha(f_0) \leq 4c_{20}\delta_{\mathrm{a}}^2$. Denoting

$$\Delta_n(f) = \frac{1}{n} \sum_{i=1}^{n} (1 - \mathbb{E}) \big\{ \rho_\alpha(Y_i - f(X_i) - \rho_\alpha(Y_i - f_0(X_i) \big\},$$

the earlier inequality is further bounded as

$$\mathscr{Q}_\alpha(f) - \mathscr{Q}_\alpha(f_0) \leq \Delta_n(f_n) - \Delta_n(f) + 4c_{20}\delta_{\mathrm{a}}^2 + \delta_{\mathrm{opt}}^2. \tag{C.22}$$

Note that $f_n \in \mathscr{F}_n(2^j\delta_*)$ for any $j \geq 1$ because $2\delta_{\mathrm{a}} \leq 2^j\delta_*$ for any $j \geq 1$. Combining this with (C.22) and (C.21), we obtain upper bounds of the probability $\mathbb{P}\{\exists f \in \mathscr{S}_n(\delta_{\mathrm{opt}}) \cap \mathscr{D}_{n,j}\}$ as

$$\begin{aligned}
&\mathbb{P}\{\exists f \in \mathscr{S}_n(\delta_{\mathrm{opt}}) \cap \mathscr{D}_{n,j}\} \\
&\leq \mathbb{P}\left\{\exists f \in \mathscr{D}_{n,j} \text{ such that } \Delta_n(f_n) - \Delta_n(f) \geq \frac{c_{19}}{4}2^{2j}\delta_*^2 - 4c_{20}\delta_{\mathrm{a}}^2 - \delta_{\mathrm{opt}}^2\right\} \\
&\leq \mathbb{P}\left\{\sup_{f \in \mathscr{F}_n(2^j\delta_*)} |\Delta_n(f)| \geq \frac{c_{19}}{16}2^{2j}\delta_*^2\right\}, \qquad \text{(C.23)}
\end{aligned}$$

where the last line follows from the choice of c_9 in (C.19) that

$$\frac{c_{19}}{8}2^{2j}\delta_*^2 \geq \frac{c_{19}}{2}c_9^2(\delta_{\mathrm{a}} + \delta_{\mathrm{opt}})^2 \geq 4c_{20}\delta_{\mathrm{a}}^2 + \delta_{\mathrm{opt}}^2$$

for any $j \geq 1$.

We next bound the probability $\mathbb{P}\{\sup_{f \in \mathscr{F}_n(2^j\delta_*)} |\Delta_n(f)| \geq c_{19}2^{2j}\delta_*^2/16\}$. We choose $\delta = 2^j\delta_*$ and $x = 2^{2j}u$ to apply Lemma C.1.8. Since $c_9 \geq 1$, we have $\delta \geq \delta_{\mathrm{s}}$ and $0 \leq x \leq n\delta^2$. Then, Lemma C.1.8 yields

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{f \in \mathscr{F}_n(2^j\delta_*)} |\Delta_n(f)| \geq \frac{c_{21}}{c_9}2^{2j}\delta_*^2\right\} \\
&= \mathbb{P}\left[\sup_{f \in \mathscr{F}_n(2^j\delta_*)} \left|\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\{\rho_\alpha(Y_i - f(X_i)) - \rho_\alpha(Y_i - f_0(X_i))\}\right| \geq \frac{c_{21}}{c_9}2^{2j}\delta_*^2\right] \\
&\leq \mathbb{P}\left[\sup_{f \in \mathscr{F}_n(2^j\delta_*)} \left|\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\{\rho_\alpha(Y_i - f(X_i)) - \rho_\alpha(Y_i - f_0(X_i))\}\right| \geq c_{21}\delta\left(\delta_{\mathrm{s}} + \sqrt{\frac{x}{n}}\right)\right] \\
&\leq \exp(-x) = \exp(-2^{2j}u).
\end{aligned}$$

Since c_9 satisfies $c_{21}/c_9 \leq c_{19}/16$, the above probability bound yields

$$\mathbb{P}\left\{\sup_{f \in \mathscr{F}_n(2^j\delta_*)} |\Delta_n(f)| \geq \frac{c_{19}}{16}2^{2j}\delta_*^2\right\} \leq \exp(-2^{2j}u).$$

Combining this with (C.20) and (C.23) implies

$$\begin{aligned}\mathbb{P}\Big\{\exists f \in \mathscr{S}_n(\delta_{\mathrm{opt}}) \ \text{ such that } \ \|f - f_0\|_2 \geq \delta_*\Big\} &\leq \sum_{j=1}^{\infty} \exp(-2^{2j}u) \\ &\leq \sum_{j=1}^{\infty} \exp(-ju) \\ &\leq (1-e^{-1})^{-1} e^{-u},\end{aligned}$$

which proves the claim. □

C.1.7 Proof of Theorem 3.4.5

The proof proceeds by specifying each term in the error bound in Theorem 3.4.4. By the assumption, we have $\delta_{\mathrm{opt}} \leq \delta_n$. For the approximation error, we can utilize Proposition 3.2.1 since the probability measure of X_i is absolutely continuous with respect to the Lebesgue measure. Applying Proposition 3.2.1 with our chosen values of L_0 and N_0, there exists a universal constant $C_1 > 0$ such that for any $f_0 \in \mathscr{H}(d, l, M_0, \mathscr{P})$,

$$\delta_{\mathrm{a}} = \inf_{f \in \mathscr{F}_n} \|f - f_0\|_2 \leq c_3 (L_0 N_0)^{-2\gamma^*} \leq C_1 c_3 \left(\frac{\log^6 n}{n}\right)^{\gamma^*/(2\gamma^*+1)}.$$

Furthermore, from the choice of L and N, we have

$$LN \leq c_1 c_2 \lceil L_0 \log L_0 \rceil \lceil N_0 \log N_0 \rceil \leq 4 c_1 c_2 (L_0 N_0) \log L_0 \log N_0.$$

Then, it follows that $\log(LN) \leq C_2 c_1 c_2 \log(L_0 N_0)$ for some universal constant $C_2 > 0$. Combining this with the choice of L_0 and N_0 gives

$$\begin{aligned}\delta_{\mathrm{s}} = LN\sqrt{\frac{\log(LN)\log n}{n}} &\leq C_3 (c_1 c_2)^{3/2} \frac{(L_0 N_0 \log L_0 \log N_0)\{\log(L_0 N_0)\log n\}^{1/2}}{\sqrt{n}} \\ &\leq C_4 (c_1 c_2)^{3/2} \frac{(L_0 N_0)\log^3 n}{\sqrt{n}} \leq C_5 (c_1 c_2)^{3/2} \left(\frac{\log^6 n}{n}\right)^{\gamma^*/(2\gamma^*+1)}\end{aligned}$$

for some universal positive constants $C_3 - C_5$. Remark that the prefactors $c_1 - c_3$ have a polynomial dependence on $t_{\max}$ so that the prefactors $C_1 c_3$ and $C_5 (c_1 c_2)^{3/2}$ in the bounds of δ_{a} and δ_{s} also demonstrate a polynomial dependence on $t_{\max}$. Therefore, there exists a positive constant $c_{10} > 0$, which depends on $t_{\max}$ polynomially and satisfies that for any $u \geq 1$,

$$c_9 \left(\delta_{\mathrm{s}} + \delta_{\mathrm{a}} + \delta_{\mathrm{opt}} + \frac{u}{\sqrt{n}} \right) \leq c_{10} \left(\delta_n + \sqrt{\frac{u}{n}} \right).$$

Plugging these values into the deviation bound in Theorem 3.4.4 establishes the claim. □

C.1.8 Proof of Theorem 3.4.6

In a similar manner to the proof of Theorem 3.4.5, we proceed to specify each term that constitutes the bound in Theorem 3.4.1. To begin with, recall that we choose τ as

$$\tau \asymp \nu_p^{1/p} \left(\frac{n}{\log^6 n} \right)^{2\gamma^*(1-\zeta_p)/(2\gamma^*+\zeta_p)} \quad \text{with} \quad \zeta_p = 1 - \frac{1}{2p-1}.$$

Then, for all sufficiently large n satisfying

$$\left(\frac{n}{\log^6 n} \right)^{\frac{\gamma^*}{2\gamma^*+\zeta_p}} \gtrsim \max \left\{ \nu_p^{1/p}, M_0 / \nu_p^{1/p} \right\}^{p-1/2}, \tag{C.24}$$

we have $\tau \gtrsim \nu_p^{2/p}$ and $\tau \geq c_4$. Furthermore, we have $\tau \nu_p^{-1/p} \lesssim n$. Thus, following a similar argument in the proof of Theorem 3.4.5 yields

$$\begin{aligned} \eta_{\mathrm{s}} = (\nu_p^{1/p} + \sqrt{\tau}) \sqrt{\frac{(LN)^2 \log(LN) \log(n^2 \tau \nu_p^{-1/p})}{n}} &\leq C_1 (c_1 c_2)^{3/2} \sqrt{\tau} \frac{(L_0 N_0) \log^3 n}{\sqrt{n}} \\ &\leq C_2 (c_1 c_2)^{3/2} \nu_p^{1/(2p)} \left(\frac{\log^6 n}{n} \right)^{\frac{\gamma^* \zeta_p}{2\gamma^*+\zeta_p}} \end{aligned}$$

for some universal constants $C_1, C_2 > 0$. In addition, we have

$$\eta_{\mathrm{b}} = \frac{\nu_p}{(\tau/2)^{p-1}} \lesssim \nu_p^{1/p} \left(\frac{\log^6 n}{n}\right)^{\frac{2\gamma^*(1-\zeta_p)(p-1)}{2\gamma^*+\zeta_p}} = \nu_p^{1/p} \left(\frac{\log^6 n}{n}\right)^{\frac{\gamma^*\zeta_p}{2\gamma^*+\zeta_p}},$$

and there exist universal constants $C_3, C_4 > 0$ such that

$$\begin{aligned}\delta_{\mathrm{s}} = \sqrt{\frac{\mathrm{Pdim}(\mathscr{F}_n)\log n}{n}} \leq C_3 \sqrt{\frac{(LN)^2 \log(LN)\log n}{n}} &\leq C_4 (c_1 c_2)^{3/2} \left(\frac{\log^6 n}{n}\right)^{\frac{\gamma^*}{2\gamma^*+\zeta_p}} \\ &\leq C_4 (c_1 c_2)^{3/2} \left(\frac{\log^6 n}{n}\right)^{\frac{\gamma^*\zeta_p}{2\gamma^*+\zeta_p}}, \quad \text{(C.25)}\end{aligned}$$

where the first inequality follows from Lemma C.2.4. Regarding the approximation error η_{a}, Proposition 3.2.1 implies that there exists a universal positive constant C_5 satisfying

$$\eta_{\mathrm{a}} = \inf_{g \in \mathscr{G}_n} \|g - g_0\|_2 \leq c_3 (L_0 N_0)^{-2\gamma^*} \leq C_5 c_3 \left(\frac{\log^6 n}{n}\right)^{\gamma^*\zeta_p/(2\gamma^*+\zeta_p)} \quad \forall g_0 \in \mathscr{H}(d, l, M_0, \mathscr{P}).$$

By the definition, the optimization error $\eta_{\mathrm{opt}} \leq \eta_n^{\mathrm{AH}}$. Next, we apply Theorem 3.4.4 to find an upper bound of $\|f - f_0\|_4$ for $f \in \mathscr{S}(\delta_{\mathrm{opt}})$. By the definition, $\delta_{\mathrm{opt}} \leq \eta_n^{\mathrm{AH}}$, and following the same argument for deriving an upper bound of the approximation η_{a} gives

$$\inf_{f \in \mathscr{F}_n} \|f - f_0\|_2 \leq C_5 c_3 \left(\frac{\log^6 n}{n}\right)^{\gamma^*\zeta_p/(2\gamma^*+\zeta_p)}.$$

Combining the two bounds with (C.25) and applying Theorem 3.4.4, we have for any $f_0 \in \mathscr{H}(d, l, M_0, \mathscr{P})$ that

$$\mathbb{P}\left[\sup_{f \in \mathscr{S}(\delta_{\mathrm{opt}})} \|f - f_0\|_2 \geq C_6 \left\{ \max\left(\nu_p^{1/p}, 1\right) \cdot \left(\frac{\log^6 n}{n}\right)^{\gamma^*\zeta_p/(2\gamma^*+\zeta_p)} + \sqrt{\frac{u}{n}} \right\}\right] \lesssim e^{-u},$$

where C_6 has a polynomial dependence on $t_{\max}$. Since $\|f_0\|_\infty \le M_0$ and $\|f\|_\infty \le M_0$ for any $f \in \mathscr{F}_n$, this implies

$$\mathbb{P}\left[\sup_{f\in\mathscr{S}(\delta_{\mathrm{opt}})} \|f-f_0\|_4^2 \ge 2M_0\cdot C_6\left\{\max\left(\nu_p^{1/p},1\right)\cdot\left(\frac{\log^6 n}{n}\right)^{\gamma^*\zeta_p/(2\gamma^*+\zeta_p)} + \sqrt{\frac{u}{n}}\right\}\right] \lesssim e^{-u}.$$

Finally, under the scaling condition C.24, we have

$$(\nu_p^{1/p}+\sqrt{\tau})\sqrt{\frac{u}{n}} \lesssim \sqrt{\tau}\sqrt{\frac{u}{n}} \lesssim \nu_p^{1/(2p)}\frac{\sqrt{u}}{n^{(2\gamma^*+1)\zeta_p/(4\gamma^*+2\zeta_p)}} \le \nu_p^{1/(2p)}\sqrt{\frac{u}{n^{\zeta_p}}}.$$

Putting the pieces together into the bound (3.15), there exists $c_{11}>0$ with a polynomial dependence on $t_{\max}$ satisfying

$$\mathbb{P}\left\{\sup_{f\in\mathscr{S}_n(\delta_{\mathrm{opt}}),\, g\in\mathscr{T}_{n,\tau}(\eta_{\mathrm{opt}};f)} \alpha\|g-g_0\|_2 \ge c_{11}\left[\eta_n^{\mathrm{AH}} + \max\left\{\nu_p^{1/(2p)},1\right\}\sqrt{\frac{u}{n^{\zeta_p}}}\right]\right\} \lesssim e^{-u}.$$

This concludes the proof of the claim. □

C.1.9 Proof of Theorem 3.4.7

To apply Theorem 3.4.2, we follow a similar line of argument in the proof of Theorem 3.4.6. To begin with, from the choice of L and N, there exist some positive constants $C_1, C_2 > 0$ satisfying

$$\begin{aligned} V_n = \sqrt{\frac{(LN)^2\log(LN)\log n}{n}} &\le C_1(c_1c_2)^{3/2}\frac{(L_0N_0)\log^3 n}{\sqrt{n}} \\ &\le C_2(c_1c_2)^{3/2}\left(\frac{\log^6 n}{n}\right)^{\frac{\gamma^*}{2\gamma^*+\xi_p}}. \end{aligned} \tag{C.26}$$

Given that n is sufficiently large so that $V_n \leq 1$, we have

$$\begin{aligned}\eta_{\mathrm{s}} = \nu_p^{1/p} V_n + \nu_p^{1/(2p)} V_n^{1-1/p} &\leq 2C_2(c_1c_2)^{3/2} \max\left\{\nu_p^{1/p}, \nu_p^{1/(2p)}\right\} \cdot \left(\frac{\log^6 n}{n}\right)^{\frac{\gamma^* \xi_p}{2\gamma^* + \xi_p}} \\ &\leq 2C_2(c_1c_2)^{3/2} \max\left(\nu_p^{1/p}, 1\right) \cdot \left(\frac{\log^6 n}{n}\right)^{\frac{\gamma^* \xi_p}{2\gamma^* + \xi_p}}.\end{aligned}$$

Furthermore, Proposition 3.2.1 implies

$$\eta_{\mathrm{a}} \leq C_3 c_3 \left(\frac{\log^6 n}{n}\right)^{\gamma^* \xi_p/(2\gamma^* + \xi_p)},$$

and the optimization error satisfies $\eta_{\mathrm{opt}} \leq \eta_n^{\mathrm{LS}}$. Turning to deriving a high-probability bound of $\|f - f_0\|_4$ for $f \in \mathscr{S}(\delta_{\mathrm{opt}})$, note that $\delta_{\mathrm{opt}} \leq \eta_n^{\mathrm{LS}}$ and

$$\inf_{f \in \mathscr{F}_n} \|f - f_0\|_2 \leq C_4 c_3 \left(\frac{\log^6 n}{n}\right)^{\gamma^* \xi_p/(2\gamma^* + \xi_p)} \qquad \forall f_0 \in \mathscr{H}(d, l, M_0, \mathscr{P}).$$

Combining these two bounds with (C.26), Theorem 3.4.4 implies that there exists a constant C_5 with a polynomial dependence on $t_{\max}$ satisfying

$$\mathbb{P}\left[\sup_{f \in \mathscr{S}(\delta_{\mathrm{opt}})} \|f - f_0\|_2 \geq C_5 \left\{\max\left(\nu_p^{1/p}, 1\right) \cdot \left(\frac{\log^6 n}{n}\right)^{\gamma^* \xi_p/(2\gamma^* + \xi_p)} + \sqrt{\frac{x}{n}}\right\}\right] \lesssim e^{-x}$$

for any $x \geq 1$. Taking $x = n \cdot u\{\log^6(n)/n\}^{2\gamma^* \xi_p/(2\gamma^* + \xi_p)} \geq n^{1/p}$ in this bound and recalling the boundedness of $\mathscr{F}_n$ and f_0, we further have

$$\mathbb{P}\left\{\sup_{f \in \mathscr{S}(\delta_{\mathrm{opt}})} \|f - f_0\|_4^2 \geq 4C_5 M_0 \sqrt{u} \cdot \max\left(\nu_p^{1/p}, 1\right) \left(\frac{\log^6 n}{n}\right)^{\gamma^* \xi_p/(2\gamma^* + \xi_p)}\right\} \lesssim e^{-n^{1/p}}.$$

Remark that it suffices to consider the case $u \leq n^2$. Otherwise, the deviation bound becomes trivial by the uniform boundedness of g_0 and $\mathscr{G}_n$. Then, putting the pieces together and applying Theorem 3.4.2, there exists a positive constant c_{12} with a polynomial dependence on $t_{\max}$, which

satisfies

$$\begin{aligned}\mathbb{P}\left[\sup_{f\in\mathscr{S}_n(\delta_{\mathrm{opt}}),\, g\in\mathscr{T}_{n,\tau}(\eta_{\mathrm{opt}};f)} \alpha\|g-g_0\|_2 \geq c_{12}\cdot u\eta_n^{\mathrm{LS}}\right] &\lesssim e^{-nV_n^2}+e^{-n^{1/p}}+\frac{1}{u^{2p}}\\ &\lesssim e^{-C_6 n^{1/p}}+\frac{1}{u^p}\\ &\lesssim \frac{1}{n^{2p}}+\frac{1}{u^p}\\ &\lesssim \frac{1}{u^p}\end{aligned}$$

for sufficiently large n and $1\leq u\leq n^2$. This proves the theorem. □

C.1.10 Proof of Theorem 3.4.8

By following a similar argument as presented in the proof of Theorem 3.4.6, the theorem can be readily derived from Proposition 3.2.1, Theorem 3.4.5 and Theorem 3.4.3. □

C.1.11 Proof of Proposition 3.2.1

The following ReLU network approximation result for the function class $\mathscr{H}^{\beta}([0,1]^d, M_0)$ plays a crucial role in the proof of Proposition 3.2.1.

Lemma C.1.9 (Theorem 3.3 in Jiao et al. (2023))**.** For any $C_0>0$, assume $f\in\mathscr{H}^{\beta}([0,1]^d,C_0)$ with $\beta=r+s$, $r=\lfloor\beta\rfloor\in\mathbb{N}_0$ and $s\in(0,1]$. For any $L_0,N_0\in\mathbb{N}$ and $\delta\in(0,1/(3B)]$ with $B=\lceil(L_0N_0)^{2/d}\rceil$, there exists a function $\phi\in\mathscr{F}_{\mathrm{DNN}}(d,L,N)$ with depth $L=21(r+1)^2L_0\lceil\log_2(8L_0)\rceil$ and width $N=38(r+1)^2d^{r+1}N_0\lceil\log_2(8N_0)\rceil$ such that

$$|f(\boldsymbol{x})-\phi(\boldsymbol{x})|\leq 19C_0(r+1)^2d^{r+(\beta\vee 1)/2}(L_0N_0)^{-2\beta/d}$$

for all $\boldsymbol{x}\in[0,1]^d\setminus\Omega([0,1]^d,B,\delta)$, where

$$\Omega([0,1]^d,B,\delta)=\bigcup_{j=1}^{d}\left\{\boldsymbol{x}=(x_1,\ldots,x_d)^{\mathrm{T}}\in[0,1]^d : x_j\in\bigcup_{b=1}^{B-1}(b/B-\delta,b/B)\right\}.$$

We also need the following lemma which is derived from the discussions in Section B.1 of Fan et al. (2022).

Lemma C.1.10. Assume that $g_i \in \mathscr{F}_{\mathrm{DNN}}(d, L_i, N_i)$ for $1 \le i \le t$ for some $t \in \mathbb{N}$ and $h \in \mathscr{F}_{\mathrm{DNN}}(t, L, N)$. Then, we have

$$h(g_1, \ldots, g_t) \in \mathscr{F}_{\mathrm{DNN}}\left(d, L + \max_{1 \le i \le t} L_i, N \vee \sum_{i=1}^{t} N_i\right).$$

The proof of Proposition 3.2.1 is based on and refines the argument presented in the proof of Proposition 3.5 in Fan et al. (2022). The primary distinction lies in the use of Lemma C.1.9, which results in a polynomial dependence on $t_{\max}$ for the prefactors in our approximation bound and the width N. In contrast, the approximation error from Proposition 3.5 in Fan et al. (2022) exhibits an exponential dependence on t_{max}. Furthermore, the proof of Proposition 3.2.1 requires a more delicate analysis to manage unfavorable subsets in which the approximation bound is not valid.

Proof of Proposition 3.2.1. To begin with, define $\beta_{\max} = \sup_{(\beta,t)\in\mathscr{P}} \beta$ and $t_{\max} = \sup_{(\beta,t)\in\mathscr{P}} t$. We first show that there exist positive constants $c_1 - c_3$ that depend on d polynomially such that for any $f_0 \in \mathscr{H}(d, l, M_0, \mathscr{P})$ and $\delta_0 \in (0,1)$, there exists a neural network

$$f^* \in \mathscr{F}_{\mathrm{DNN}}(d, c_1 \lceil L_0 \log L_0 \rceil, c_2 \lceil N_0 \log N_0 \rceil, M_0)$$

such that

$$|f_0(\boldsymbol{x}) - f^*(\boldsymbol{x})| \le c_3 (L_0 N_0)^{-2\gamma^*} \text{ for all } \boldsymbol{x} \in [0,1]^d \setminus \Xi_0, \tag{C.27}$$

where $\Xi_0 \subseteq [0,1]^d$ is defined below and the Lebesgue measure of Ξ_0 is less than δ_0.

STEP 1. CONSTRUCTION OF NEURAL NETWORKS. For a fixed $f_0 \in \mathscr{H}(d, l, M_0, \mathscr{P})$ with $l > 1$, we denote $h_1^{(l)}(\boldsymbol{x}) = f_0(\boldsymbol{x})$. By the definition of $\mathscr{H}(d, l, M_0, \mathscr{P})$, $h_1^{(l)}(\boldsymbol{x})$ is recursively

computed consisting of various hierarchical interaction models at level $i \in \{1,\dots,l-1\}$. Let R_i denote the number of hierarchical composition models at level i, which are necessary to compute $h_1^{(l)}$. For each level $i \in \{1,\dots,l\}$, we denote $h_j^{(i)} : \mathbb{R}^d \to \mathbb{R}$ to be the j-th ($j \in \{1,\dots,R_i\}$) hierarchical composition model at the i-th level. From the definition, each function $h_j^{(i)}$ depends on functions at level $i-1$ through a function $g_j^{(i)} \in \mathscr{H}^{\beta_j^{(i)}}(\mathbb{R}^{t_j^{(i)}}, M_0)$ with $(\beta_j^{(i)}, t_j^{(i)}) \in \mathscr{P}$. Then, $h_1^{(l)}$ is recursively described as

$$h_j^{(i)}(\boldsymbol{x}) = g_j^{(i)}\left(h^{(i-1)}_{\sum_{k=1}^{j-1} t_k^{(i)}+1}(\boldsymbol{x}), \dots, h^{(i-1)}_{\sum_{k=1}^{j} t_k^{(i)}}(\boldsymbol{x})\right) \tag{C.28}$$

for $j \in \{1,\dots,R_i\}$ and $i \in \{2,\dots,l\}$, and

$$h_j^{(1)}(\boldsymbol{x}) = g_j^{(1)}\left(x_{j_1}, \dots, x_{j_{t_j^{(1)}}}\right)$$

for some $\{j_1,\dots,j_{t_j^{(1)}}\} \subset \{1,\dots,d\}$ and $\boldsymbol{x} \in [0,1]^d$. Furthermore, we can recursively calculate that

$$R_l = 1 \ \text{ and } \ R_i = \sum_{j=1}^{R_{i+1}} t_j^{(i+1)} \ \text{ for } \ i \in \{1,\dots,l-1\},$$

so that $R_i \le t_{\max}^{l-i}$ for $i \in \{1,\dots,l\}$.

To approximate f_0, we construct a sequence of deep ReLU neural networks, approximating the sequence of functions $h_j^{(i)}$. For the given δ_0, we start with $i = 1$ and $j \in \{1,\dots,R_1\}$. Note that it suffices to approximate each function $g_j^{(1)}$ on the domain $[0,1]^{t_j^{(1)}}$. Define $B_j^{(1)} = \lceil (L_0 N_0)^{2/t_j^{(1)}} \rceil$ and choose

$$\delta_j^{(1)} = \delta_0 / (3l \cdot R_1 t_j^{(1)} B_j^{(1)}) \in (0, 1/(3B_j^{(1)})].$$

By applying Lemma C.1.9 with $C_0 = M_0, \beta = \beta_j^{(1)}$ and $\delta = \delta_j^{(1)}$, there exists a function $\widetilde{g}_j^{(1)}$ in

$\mathscr{F}_{\mathrm{DNN}}(t_j^{(1)}, L_j^{(1)}, N_j^{(1)})$ with some $L_j^{(1)}, N_j^{(1)} \in \mathbb{N}$ such that

$$\left|\widetilde{g}_j^{(1)}(\boldsymbol{y}) - g_j^s(\boldsymbol{y})\right| \leq C_j^{(1)}(L_0 N_0)^{-2\beta_j^{(1)}/t_j^{(1)}} \leq C_j^{(1)}(L_0 N_0)^{-2\gamma^*} \tag{C.29}$$

for all $\boldsymbol{y} \in [0,1]^{t_j^{(1)}} \setminus \Omega([0,1]^{t_j^{(1)}}, B_j^{(1)}, u_j^{(1)})$, where $C_j^{(1)} = 19M_0(\lfloor \beta_j^{(1)} \rfloor + 1)^2 d^{\lfloor \beta_j^{(1)} \rfloor + \beta_j^{(1)}/2}$ and Ω is defined in Lemma C.1.9. Here, the last inequality holds by the definition of γ^* and recall that $\mathscr{P} \in [1,\infty) \times \mathbb{N}^+$ so that $\beta_j^{(1)} \geq 1$. Remark that for any $t \in \mathbb{N}, L_1 \leq L_2$ and $N_1 \leq N_2$, $\mathscr{F}_{\mathrm{DNN}}(t, L_1, N_1) \subset \mathscr{F}_{\mathrm{DNN}}(t, L_2, N_2)$. Therefore, we can regard $\widetilde{g}_j^{(1)}$ to be a function in $\mathscr{F}_{\mathrm{DNN}}(t_j^{(1)}, L', N')$, where

$$L' = C_1 \lceil L_0 \log L_0 \rceil \text{ and } N' = C_2 \lceil N_0 \log N_0 \rceil$$

with $C_1 = 63(\lfloor \beta_{\max} \rfloor + 1)^2$ and $C_2 = 114(\lfloor \beta_{\max} \rfloor + 1)^2 t_{\max}^{\lfloor \beta_{\max} \rfloor + 1}$. Remark that the range of $\widetilde{g}_j^{(1)}$ may not be contained in $[-M_0, M_0]$. To correct this, we truncate each neural networks $\widetilde{g}_j^{(1)}$ as

$$\widehat{g}_j^{(1)} := \max[\min\{\widetilde{g}_j^{(1)}(\boldsymbol{z}), M_0\}, -M_0] = \sigma(2M_0 - \sigma(M_0 - \widetilde{g}_j^{(1)}(\boldsymbol{z}))) - M_0,$$

where $\sigma(\cdot)$ is the ReLU activation function. Note that if $g \in \mathscr{F}_{\mathrm{DNN}}(t, L_1, N_1)$ for some $t, L_1, N_1 \in \mathbb{N}$, then for any $a, b \in \mathbb{R}$, $a\sigma(g) + b \in \mathscr{F}_{\mathrm{DNN}}(t, L_1 + 1, N_1)$. Therefore, $\widehat{g}_j^{(1)} \in \mathscr{F}_{\mathrm{DNN}}(t_j^{(1)}, L' + 2, N')$. Now, we define

$$\widehat{h}_j^{(1)}(\boldsymbol{x}) = \widehat{g}_j^{(1)}\left(x_{j_1}, \ldots, x_{j_{t_j^{(1)}}}\right) \text{ for } \boldsymbol{x} \in [0,1]^d.$$

Since $\|g_j^{(1)}\|_\infty \le M_0$, (C.29) implies that

$$\begin{aligned}\left|\widehat{h}_j^{(1)}(\boldsymbol{x}) - h_j^{(1)}(\boldsymbol{x})\right| &= \left|\widehat{g}_j^{(1)}\left(x_{j_1},\ldots,x_{j_{t_j^{(1)}}}\right) - g_j^{(1)}\left(x_{j_1},\ldots,x_{j_{t_j^{(1)}}}\right)\right| \\ &\le C_j^{(1)}(L_0N_0)^{-2\gamma^*} \\ &\le C_3(L_0N_0)^{-2\gamma^*} \end{aligned} \tag{C.30}$$

for all $\boldsymbol{x} \in [0,1]^d \setminus \Xi_j^{(1)}$, where C_3 is defined as

$$C_3 = 19\cdot 2^{\lfloor\beta_{\max}\rfloor} M_0^{\lfloor\beta_{\max}\rfloor+1}(\lfloor\beta_{\max}\rfloor+1)^2 t_{\max}^{\lfloor\beta_{\max}\rfloor+\beta_{\max}/2}., \tag{C.31}$$

and

$$\Xi_j^{(1)} = \bigcup_{k=1}^{t_j^{(1)}} \left\{\boldsymbol{x} = (x_1,\ldots,x_d)^{\mathrm{T}} \in [0,1]^d : x_k \in \bigcup_{b=1}^{B_j^{(1)}-1} (b/B_j^{(1)} - \delta, b/B_j^{(1)})\right\}.$$

Note that the Lebesgue measure of $\cup_{j=1}^{R_1}\Xi_j^{(1)}$ is not larger than $\delta_0/(3l)$.

Next, we recursively construct a neural network $\widehat{h}_j^{(i)}$ for $i \in \{2,\ldots,l\}$ and $j \in \{1,\ldots,R_i\}$ to approximate $h_j^{(i)}$. Suppose that $\widehat{h}_{j'}^{(i-1)}$ is defined for $j' \in \{1,\ldots,R_{i-1}\}$. Define $B_j^{(i)} = \lceil (L_0N_0)^{2/t_j^{(i)}} \rceil$ and choose $\delta_j^{(i)} \in (0, 1/(3B_j^{(i)})]$ to be determined. Note that it suffices to approximate $g_j^{(i)}$ on the domain $[-M_0,M_0]^{t_j^{(i)}}$. Define the function

$$\bar{g}_j^{(i)}(\boldsymbol{z}) = g_j^{(i)}(2M_0\boldsymbol{z} - M_0) \ \text{ for } \ \boldsymbol{z} \in [0,1]^{t_j^{(i)}}.$$

Then, it is easy to see that $\bar{g}_j^{(i)}$ is contained in $\mathscr{H}^{\beta_j^{(i)}}([0,1]^{t_j^{(i)}}, 2^{\beta_j^{(i)}} M_0^{\beta_j^{(i)}+1})$, and satisfies

$$g_j^{(i)}(\boldsymbol{y}) = \bar{g}_j^{(i)}\left(\frac{\boldsymbol{y}+M_0}{2M_0}\right) \ \text{ for } \ \boldsymbol{y} \in [-M_0,M_0]^{t_j^{(i)}}. \tag{C.32}$$

Applying Lemma C.1.9 with $C_0 = 2^{\beta_j^{(i)}} M_0^{\beta_j^{(i)}+1}, \beta = \beta_j^{(i)}$ and $\delta = \delta_j^{(i)}$, a similar argument as in

the case of $i = 1$ gives a function $\widetilde{g}_j^{(i)} \in \mathscr{F}_{\mathrm{DNN}}(t_j^{(i)}, L', M')$ such that

$$\left|\widetilde{g}_j^{(i)}(\boldsymbol{z}) - \bar{g}_j^{(i)}(\boldsymbol{z})\right| \leq C_3 (L_0 N_0)^{-2\gamma^*}$$

for all $\boldsymbol{z} \in [0,1]^{t_j^{(i)}} \setminus \Omega([0,1]^{t_j^{(i)}}, B_j^{(i)}, \delta_j^{(i)})$. To ensure that the range of the approximating neural network is in $[-M_0, M_0]$, we truncate $\widetilde{g}_j^{(i)}$ as

$$\widehat{g}_j^{(i)} := \max[\min\{\widetilde{g}_j^{(i)}(\boldsymbol{z}), M_0\}, -M_0] = \sigma(2M_0 - \sigma(M_0 - \widetilde{g}_j^{(i)}(\boldsymbol{z}))) - M_0,$$

so that $\widehat{g}_j^{(i)} \in \mathscr{F}_{\mathrm{DNN}}(t_j^{(i)}, L'+2, N')$. Also, by (C.32), we have

$$\left|\widehat{g}_j^{(i)}\left(\frac{\boldsymbol{y}+M_0}{2M_0}\right) - \bar{g}_j^{(i)}\left(\frac{\boldsymbol{y}+M_0}{2M_0}\right)\right| \leq C_3 (L_0 N_0)^{-2\gamma^*} \tag{C.33}$$

for all $\boldsymbol{y} \in [-M_0, M_0]^{t_j^{(i)}}$ except the small subset. Now, we construct a neural network as

$$\widehat{h}_j^{(i)}(\boldsymbol{x}) = \widehat{g}_j^{(i)}\left(\frac{\widehat{h}^{(i-1)}_{\sum_{k=1}^{j-1} t_k^{(i)}+1}(\boldsymbol{x}) + M_0}{2M_0}, \ldots, \frac{\widehat{h}^{(i-1)}_{\sum_{k=1}^{j} t_k^{(i)}}(\boldsymbol{x}) + M_0}{2M_0}\right),$$

which approximates $h_j^{(i)}$ defined in (C.28). To determine the value of $\delta_j^{(i)}$ given neural networks $\widehat{h}_{j'}^{(i-1)}$ for $1 \leq j' \leq R_{i-1}$, consider the map $\boldsymbol{r}_j^{(i)} : [0,1]^d \to \mathbb{R}^{t_j^{(i)}}$ defined as

$$\boldsymbol{r}_j^{(i)}(\boldsymbol{x}) := \left(\frac{\widehat{h}^{(i-1)}_{\sum_{k=1}^{j-1} t_k^{(i)}+1}(\boldsymbol{x}) + M_0}{2M_0}, \ldots, \frac{\widehat{h}^{(i-1)}_{\sum_{k=1}^{j} t_k^{(i)}}(\boldsymbol{x}) + M_0}{2M_0}\right) \text{ for } \boldsymbol{x} \in [0,1]^d.$$

Then, we choose $\delta_j^{(i)}$ such that the Lebesgue measure of $\cup_{j=1}^{R_i} \Xi_j^{(i)}$ is less than δ_0/l, where

$$\Xi_j^{(i)} := \left(\boldsymbol{r}_j^{(i)}\right)^{-1}\left(\Omega\left([0,1]^{t_j^{(i)}}, B_j^{(i)}, \delta_j^{(i)}\right)\right).$$

The existence of $\delta_j^{(i)}$ is guaranteed, as each $\boldsymbol{r}_j^{(i)}$ is a continuous function. Finally, we set a neural

network $f^* = \widehat{h}_1^{(l)}$ recursively, which approximates f_0. Remark that by the definition, we have $\|f^*\|_\infty \leq M_0$. Also, denoting

$$\Xi_0 = \cup_{i=1}^{l} \cup_{j=1}^{R_i} \Xi_j^{(i)},$$

it follows that the Lebesgue measure of Ξ_0 is less than δ_0 by construction.

STEP 2. CALCULATING WIDTHS AND DEPTHS. To calculate the width and depth of f^*, we sequentially specify width and depth of $\widehat{h}_j^{(i)}$ from $i = 1$ to $i = l$. For $i = 1$, from the construction of $\widehat{h}_j^{(1)}$, we have $\widehat{h}_j^{(1)} \in \mathscr{F}_{\mathrm{DNN}}(t_j^{(1)}, L'+2, N')$. Recursively, for $2 \leq i \leq l$, combining Lemma C.1.10 with the inequality $R_i \leq t_{\max}^{l-i}$ implies that $\widehat{h}_j^{(i)} \in \mathscr{F}_{\mathrm{DNN}}(t_j^{(i)}, i(L'+2), t_{\max}^{i-1}N')$. Therefore, we have $f^* \in \mathscr{F}(d, L, N)$, where the depth L satisfies

$$l(L'+2) \leq c_1 \lceil L_0 \log L_0 \rceil =: L$$

with $c_1 = 2lC_1$, and the width N satisfies

$$t_{\max}^{l-1} N' \leq c_2 \lceil N_0 \log N_0 \rceil =: N,$$

where $c_2 = t_{\max}^{l-1} C_2$.

STEP 3. CALCULATING APPROXIMATION ERRORS. Now, we calculate the approximation error bound of f^*. To this end, we show by induction on i that

$$\left|\widehat{h}_j^{(i)}(\boldsymbol{x}) - h_j^{(i)}(\boldsymbol{x})\right| \leq C_3 (M_0 t_{\max}^{1/2} + 1)^{i-1} (L_0 N_0)^{-2\gamma^*} \quad \text{for } \boldsymbol{x} \in [0,1]^d \setminus \Xi_0. \tag{C.34}$$

Starting with the case of $i = 1$, (C.34) holds for $j = 1, \ldots, R_1$ by (C.30). Suppose that (C.34) holds for some $i-1$ and every $j = 1, \ldots, R_{i-1}$. Denoting

$$\boldsymbol{w} = \left(h_{\sum_{k=1}^{j-1} t_k^{(i)}+1}^{(i-1)}(\boldsymbol{x}), \ldots, h_{\sum_{k=1}^{j} t_k^{(i)}}^{(i-1)}(\boldsymbol{x})\right) \text{ and } \widehat{\boldsymbol{w}} = \left(\widehat{h}_{\sum_{k=1}^{j-1} t_k^{(i)}+1}^{(i-1)}(\boldsymbol{x}), \ldots, \widehat{h}_{\sum_{k=1}^{j} t_k^{(i)}}^{(i-1)}(\boldsymbol{x})\right),$$

we have that for any $\boldsymbol{x} \in [0,1]^d$,

$$\begin{aligned}\left|\widehat{h}_j^{(i)}(\boldsymbol{x}) - h_j^{(i)}(\boldsymbol{x})\right| &= \left|\widehat{g}_j^{(i)}\Big(\frac{\widehat{\boldsymbol{w}}+M_0}{2M_0}\Big) - \bar{g}_j^{(i)}\Big(\frac{\boldsymbol{w}+M_0}{2M_0}\Big)\right| \\ &\le \left|\widehat{g}_j^{(i)}\Big(\frac{\widehat{\boldsymbol{w}}+M_0}{2M_0}\Big) - \bar{g}_j^{(i)}\Big(\frac{\widehat{\boldsymbol{w}}+M_0}{2M_0}\Big)\right| + \left|\bar{g}_j^{(i)}\Big(\frac{\widehat{\boldsymbol{w}}+M_0}{2M_0}\Big) - \bar{g}_j^{(i)}\Big(\frac{\boldsymbol{w}+M_0}{2M_0}\Big)\right|.\end{aligned}$$

Now, (C.33) gives that

$$\left|\widehat{g}_j^{(i)}\Big(\frac{\widehat{\boldsymbol{w}}+M_0}{2M_0}\Big) - \bar{g}_j^{(i)}\Big(\frac{\widehat{\boldsymbol{w}}+M_0}{2M_0}\Big)\right| \le C_3(L_0N_0)^{-2\gamma^*},$$

when $\boldsymbol{x} \in [0,1]^d \setminus \Xi_0$. Moreover, note that $\mathscr{P} \subseteq [1,\infty) \times \mathbb{N}$ so that $g_j^{(i)}$ is M_0-Lipschitz by the definition of the Hölder function class. Therefore, when $\boldsymbol{x} \in [0,1]^d \setminus \Xi_0$, we have

$$\begin{aligned}\left|\bar{g}_j^{(i)}\Big(\frac{\widehat{\boldsymbol{w}}+M_0}{2M_0}\Big) - \bar{g}_j^{(i)}\Big(\frac{\boldsymbol{w}+M_0}{2M_0}\Big)\right| &= |g_j^{(i)}(\widehat{\boldsymbol{w}}) - g_j^{(i)}(\boldsymbol{w})| \\ &\le M_0\|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2 \\ &\le M_0 t_{\max}^{1/2}\|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_\infty \\ &\le M_0 t_{\max}^{1/2}(1+M_0 t_{\max}^{1/2})^{i-2} C_3(L_0N_0)^{-2\gamma^*},\end{aligned}$$

where the last inequality follows from the induction hypothesis. Together with earlier inequalities, we have for $\boldsymbol{x} \in [0,1]^d \setminus \Xi_0$ that

$$\begin{aligned}\left|\widehat{h}_j^{(i)}(\boldsymbol{x}) - h_j^{(i)}(\boldsymbol{x})\right| &\le C_3(L_0N_0)^{-2\gamma^*} + M_0 t_{\max}^{1/2}(1+M_0 t_{\max}^{1/2})^{i-2} C_3(L_0N_0)^{-2\gamma^*} \\ &\le C_3(1+M_0 t_{\max}^{1/2})^{i-1}(L_0N_0)^{-2\gamma^*}.\end{aligned}$$

Therefore, inductively, we have

$$|f^*(\boldsymbol{x}) - f_0(\boldsymbol{x})| = |\widehat{h}_1^{(l)}(\boldsymbol{x}) - h_1^{(l)}(\boldsymbol{x})| \le \underbrace{C_3(1+M_0 t_{\max}^{1/2})^{l-1}}_{=:c_3}(L_0N_0)^{-2\gamma^*}$$

for any $\boldsymbol{x} \in [0,1]^d \setminus \Xi_0$, which establishes the claim. Remark that from the definition of C_3 in (C.31), c_3 also has a polynomial dependence on $t_{\max}$.

To complete the proof, fix a function $f_0 \in \mathscr{H}(d,l,M_0,\mathscr{P})$ and $\varepsilon > 0$. Since the given measure μ is absolutely continuous with respect to the Lebesgue measure, there exists $\delta_0 \in (0,1)$ satisfying that any measurable set $\mathscr{E}$ whose Lebesgue measure is less than δ_0 satisfies $\mu(\mathscr{E}) < \varepsilon$. Then, there exist a measurable set Ξ_0 whose Lebesgue measure is less than δ_0, and a neural network $f^* \in \mathscr{F}_{\mathrm{DNN}}(d, c_1\lceil L_0 \log L_0 \rceil, c_2 \lceil N_0 \log N_0 \rceil, M_0)$ which satisfies (C.27). Therefore, it follows that

$$\begin{aligned}
&\left\{ \int_{[0,1]^d} |f_0(\boldsymbol{x}) - f^*(\boldsymbol{x})|^2 \mu(\mathrm{d}\boldsymbol{x}) \right\}^{1/2} \\
&\leq \left\{ \int_{[0,1]^d \setminus \Xi_0} |f_0(\boldsymbol{x}) - f^*(\boldsymbol{x})|^2 \mu(\mathrm{d}\boldsymbol{x}) \right\}^{1/2} + \left\{ \int_{\Xi_0} |f_0(\boldsymbol{x}) - f^*(\boldsymbol{x})|^2 \mu(\mathrm{d}\boldsymbol{x}) \right\}^{1/2} \\
&\leq c_3 (L_0 N_0)^{-2\gamma^*} + 2M_0 \varepsilon^{1/2}.
\end{aligned}$$

Since ε is arbitrary, this completes the proof. □

C.1.12 Proof of Proposition 3.4.1

For simplicity, we only consider the case when $\tau = 0.5$. Also, we assume that X_i follows the uniform distribution and ε_i is independent with X_i and follows the normal distribution $\mathscr{N}(0,\sigma^2)$ with $\sigma^2 = (2\pi \underline{p}^2)^{-1}$. Remark that $p_{\varepsilon_i|X_i}(0) = \underline{p}$. When $t^* \leq d$, we have $\mathscr{H}^{\beta^*}([0,1]^{t^*}, M_0) \subseteq \mathscr{H}(d,l,\mathscr{P},M_0)$, which implies

$$\inf_{\widehat{f}_n} \sup_{\substack{f_0 \in \mathscr{H}(d,l,\mathscr{P},M_0) \\ X \sim \mathbb{P}_X}} \mathbb{E}\|\widehat{f}_n - f_0\|_2 \geq \inf_{\widehat{f}_n} \sup_{\substack{f_0 \in \mathscr{H}^{\beta^*}([0,1]^{t^*},M_0) \\ X \sim \mathrm{Unif}([0,1]^d)}} \mathbb{E}\|\widehat{f}_n - f_0\|_2.$$

Here, the supremum on the left-hand side is taken over all data generating processes (X,Y) satisfying

$$Y = f_0(X) + \varepsilon,$$

where $f_0 \in \mathscr{H}(d,l,\mathscr{P},M_0)$, and the quantile regression noise ε satisfies $\mathbb{P}(\varepsilon \leq 0|X) = 0.5$ and Condition 3. The supremum on the right-hand side is taken over all data generating processes (X,Y) with $f_0 \in \mathscr{H}^{\beta^*}([0,1]^{t^*},M_0)$ and $\varepsilon \sim \mathscr{N}(0,\sigma^2)$. Then, applying Theorem 3.2 of Györfi *et al.* (2002) establishes the claim. □

C.2 Proof of Technical Lemmas

We frequently utilize Talagrand's inequality throughout the proofs of technical lemmas to obtain non-asymptotic bounds of suprema of empirical processes. The following refined Talagrand inequality is derived from Theorem 7.3 in Bousquet (2003) combining with the basic inequalities that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $2\sqrt{ab} \leq a+b$ for any $a,b \geq 0$.

Lemma C.2.1 (Talagrand's inequality)**.** Let $X_1,\ldots,X_n$ be i.i.d. random variables from some distribution P_X and $\mathscr{F}$ be a measurable class of functions such that $\mathbb{E}f(X) = 0$ for any $f \in \mathscr{F}$. Assume $\sup_{f\in\mathscr{F}} \|f\|_\infty \leq A$ and let σ be a positive constant such that $\sigma^2 \geq \sup_{f\in\mathscr{F}} \mathbb{E}f^2(X_i)$. Then, for any $x > 0$,

$$\mathbb{P}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)\right| \geq 2\mathbb{E}\left\{\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)\right|\right\} + \sigma\sqrt{\frac{2x}{n}} + \frac{4Ax}{3n}\right] \leq e^{-x}.$$

We next introduce the definition of uniform covering number followed by Lemma C.2.2 which bounds the uniform covering number of a function class with the finite pseudo dimension.

Definition C.2.1 (Uniform covering number)**.** Let $n \in \mathbb{N}^+$ and $\mathscr{F} = \{f : \mathscr{X} \to \mathbb{R}\}$ be a function class. For a given $\varepsilon > 0$, the uniform covering number under L_∞-norm for the function class $\mathscr{F}$

is defined as

$$N_\infty(\varepsilon, \mathscr{F}, n) = \sup_{(x_1,\ldots,x_n)\in\mathscr{X}^n} N(\varepsilon, \mathscr{F}|_{x_1,\ldots,x_n}, \|\cdot\|_\infty),$$

where $\mathscr{F}|_{x_1,\ldots,x_n} = \{(f(x_1),\ldots,f(x_n))^{\mathrm{T}} : f \in \mathscr{F}\} \subset \mathbb{R}^n$ and $N(\varepsilon, \mathscr{W}, \|\cdot\|_\infty)$ is the ε-covering number of a subset $\mathscr{W} \subset \mathbb{R}^n$ under the supremum norm $\|\cdot\|_\infty$.

Lemma C.2.2 (Uniform covering number bound)**.** Let $\mathscr{F}$ be a set of real functions bounded by $A \geq 1$ with finite pseudo dimension $\mathrm{Pdim}(\mathscr{F}) < \infty$. For any $\varepsilon \in (0, A)$, we have

$$\log N_\infty(\varepsilon, \mathscr{F}, n) \leq \mathrm{Pdim}(\mathscr{F}) \cdot \log(enA/\varepsilon).$$

Proof. By Theorem 12.2 of Anthony and Bartlett (1999), we have

$$N_\infty(\varepsilon, \mathscr{F}, n) \leq \sum_{i=1}^{\mathrm{Pdim}(\mathscr{F})} \binom{n}{i} \left(\frac{A}{\varepsilon}\right)^i.$$

Therefore, when $n \geq \mathrm{Pdim}(\mathscr{F})$, it follows that $N_\infty(\varepsilon, \mathscr{F}, n) \leq \{enA/(\varepsilon \mathrm{Pdim}(\mathscr{F})\}^{\mathrm{Pdim}(\mathscr{F})}$, so the inequality holds. Meanwhile, when $n < \mathrm{Pdim}(\mathscr{F})$, we have

$$N_\infty(\varepsilon, \mathscr{F}, n) \leq \sum_{i=1}^{n} \binom{n}{i} \left(\frac{A}{\varepsilon}\right)^i = \left(1 + \frac{A}{\varepsilon}\right)^n,$$

which establishes the claim since $\varepsilon \in (0, A)$. □

We also need the following maximal inequality to prove technical lemmas.

Lemma C.2.3 (A maximal inequality (Chernozhukov, Chetverikov and Kato, 2014))**.** Denote $S = [0,1]^d \times \mathbb{R}$ and let $\mathscr{F}$ be a measurable class of functions $S \to \mathbb{R}$, to which a measurable envelope F is attached. Assume that $\|F\|_2 < \infty$ and let $\sigma^2 > 0$ be any positive constant such that $\sup_{f\in\mathscr{F}} \mathbb{E} f(X, \varepsilon)^2 \leq \sigma^2 \leq \|F\|_2^2$. Furthermore, we assume that there exists constants $A \geq e$ and

$v \geq 1$ such that $\sup_Q N(\varepsilon \|F\|_{Q,2}, \mathscr{F}, \|\cdot\|_{Q,2}) \leq (A/\varepsilon)^v$ for any $0 < \varepsilon \leq 1$, where the supremum is taken over all n-discrete probability measures Q on $\mathscr{S}$ and $N(\varepsilon, \mathscr{F}, \|\cdot\|_{Q,2})$ is the ε-covering number of $\mathscr{F}$ under the $L_2(Q)$ norm. Then,

$$\mathbb{E}\left\{ \sup_{f\in\mathscr{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(Y_i, X_i) - \mathbb{E} f(Y_i, X_i) \right| \right\} \lesssim \sigma \sqrt{v \log\left(\frac{A\|F\|_2}{\sigma}\right)} + \frac{v\|\bar{F}\|_2}{\sqrt{n}} \log\left(\frac{A\|F\|_2}{\sigma}\right),$$

where $\bar{F} = \max_{1\leq i\leq n} F(X_i, \varepsilon_i)$.

Finally, the next lemma bounds the pseudo dimension of the class of deep ReLU neural networks, which allows us to apply Lemma C.2.2 when $\mathscr{F}$ is a class of ReLU deep neural networks.

Lemma C.2.4. Let $\mathscr{F} = \mathscr{F}_{\text{DNN}}(d, L, N, M)$ be the function class of deep ReLU neural networks truncated at $M > 0$. Then, it follows that

$$\text{Pdim}(\mathscr{F}) \lesssim (LN)^2 \log(LN).$$

Proof. Proof of Lemma C.2.4. Denote W to be the number of all parameters of the network $\mathscr{F}_{\text{DNN}}(d, L, N)$. Then, we have $\text{Pdim}(\mathscr{F}_{\text{DNN}}(d, L, N)) \lesssim WL\log(W)$ by Theorem 7 of Bartlett *et al.* (2019). Since $W \lesssim LN^2$, it follows that

$$\text{Pdim}(\mathscr{F}_{\text{DNN}}(d, L, N)) \lesssim L^2N^2 \log(LN).$$

To calculate the pseudo dimension of the truncated neural network, note that the truncation function $\mathscr{T}_M(\cdot)$ is a non-decreasing function. Therefore, applying Theorem 11.3 of Anthony and Bartlett (1999) completes the proof. □

C.2.1 Proof of Lemma C.1.1

To begin with, we fix real-valued functions f and g. By the definition of the joint excess risk, we can represent $\mathscr{R}_\tau$ as follows:

$$\mathscr{R}_\tau(f,g) = \mathbb{E}\ell_\tau(Z_i(f) - \alpha g(X_i)).$$

We first derive the lower bound of the excess joint risk $\mathscr{R}_\tau$. Recall that $\ell'_\tau = \psi_\tau$, and ψ_τ is absolutely continuous and has a derivative $\psi'_\tau(t) = \mathbb{1}(|t| \le \tau)$. From the fundamental theorem of calculus, it follows that for every $a, b \in \mathbb{R}$,

$$\ell_\tau(a+b) - \ell_\tau(a) = \psi_\tau(a)b + \int_0^b \psi'_\tau(a+t)(b-t)\mathrm{d}t.$$

Therefore, denoting $\Delta_g(X_i) = g_0(X_i) - g(X_i)$, it follows that

$$\begin{aligned}
\mathscr{R}_\tau(f,g) - \mathscr{R}_\tau(f,g_0) &= \mathbb{E}\ell_\tau(Z_i(f) - \alpha g(X_i)) - \mathbb{E}\ell_\tau(Z_i(f) - \alpha g_0(X_i)) \\
&= \underbrace{\mathbb{E}\big\{\psi_\tau(Z_i(f) - \alpha g_0(X_i)) \cdot \alpha\Delta_g(X_i)\big\}}_{=:\mathrm{I}} \\
&\quad + \underbrace{\mathbb{E}\left[\int_0^{\alpha\Delta_g(X_i)} \psi'_\tau(Z_i(f) - \alpha g_0(X_i) + t)\{\alpha\Delta_g(X_i) - t\}\mathrm{d}t\right]}_{=:\mathrm{II}}. \qquad \text{(C.35)}
\end{aligned}$$

We next bound I and II separately.

We first bound the term I. Let $\mathbb{E}_{X_i}$ be the conditional expectation given X_i. Observe that we can write

$$\mathbb{E}_{X_i}\big\{\psi_\tau(Z_i(f) - \alpha g_0(X_i))\big\} = \mathbb{E}_{X_i}\big\{\psi_\tau(Z_i(f) - \alpha g_0(X_i)) - \psi_\tau(\omega_i)\big\} + \mathbb{E}_{X_i}\psi_\tau(\omega_i). \qquad \text{(C.36)}$$

To bound the first term on the right-hand side of (C.36), the fundamental theorem of calculus

and the definition of ω_i in (C.1) imply

$$\begin{aligned}\mathbb{E}_{X_i}\big\{\psi_\tau(Z_i(f)-\alpha g_0(X_i))-\psi_\tau(\omega_i)\big\} &= \mathbb{E}_{X_i}\bigg\{\int_0^{Z_i(f)-Z_i(f_0)}\psi'_\tau(\omega_i+t)\mathrm{d}t\bigg\}\\ &= \mathbb{E}_{X_i}\bigg[\int_0^{Z_i(f)-Z_i(f_0)}\big\{1-\mathbb{1}(|\omega_i+t|>\tau)\big\}\mathrm{d}t\bigg]. \quad \text{(C.37)}\end{aligned}$$

Denote $\Delta_f(X_i)=f_0(X_i)-f(X_i)$, and $p_{\varepsilon_i|X_i}$ to be the conditional density function of ε_i given X_i. Then, we have

$$\begin{aligned}\mathbb{E}_{X_i}\bigg\{\int_0^{Z_i(f)-Z_i(f_0)}1\cdot\mathrm{d}t\bigg\} &= \mathbb{E}_{X_i}\big\{Z_i(f)-Z_i(f_0)\big\}\\ &= \mathbb{E}\big[\{\varepsilon_i+\Delta_f(X_i)\}\mathbb{1}\{\varepsilon_i\le-\Delta_f(X_i)\}-\alpha\Delta_f(X_i)-\varepsilon_i\mathbb{1}(\varepsilon_i\le 0)\big]\\ &= \int_{-\infty}^{\Delta_f(X_i)}\big\{t+\Delta_f(X_i)\big\}p_{\varepsilon_i|X_i}(t)\mathrm{d}t-\int_{-\infty}^{0}t\,p_{\varepsilon_i|X_i}(t)\mathrm{d}t-\alpha\Delta_f(X_i)\\ &= \int_0^{-\Delta_f(X_i)}t\,p_{\varepsilon_i|X_i}(t)\mathrm{d}t+\Delta_f(X_i)\int_0^{-\Delta_f(X_i)}p_{\varepsilon_i|X_i}(t)\mathrm{d}t,\end{aligned}$$

where the last line follows from the model assumption $\mathbb{P}(\varepsilon_i\le 0|X_i)=\alpha$. Combining this with Condition 1 gives

$$\bigg|\mathbb{E}_{X_i}\bigg\{\int_0^{Z_i(f)-Z_i(f_0)}1\cdot\mathrm{d}t\bigg\}\bigg| = \bigg|\int_0^{-\Delta_f(X_i)}\big\{t+\Delta_f(X_i)\big\}p_{\varepsilon_i|X_i}(t)\mathrm{d}t\bigg| \le \frac{\bar{p}}{2}\big\{\Delta_f(X_i)\big\}^2. \quad \text{(C.38)}$$

To establish a bound of the remaining term on the right-hand side of (C.37), we find an upper bound of $Z_i(f)-Z_i(f_0)$. We first assume that $\Delta_f(X_i)\le 0$. From the definition of $Z_i(f)$, we have

$$\begin{aligned}&\big|Z_i(f)-Z_i(f_0)\big|\\ &= \big|\{Y_i-f(X_i)\}\mathbb{1}\{Y_i\le f(X_i)\}-\{Y_i-f_0(X_i)\}\mathbb{1}\{Y_i\le f_0(X_i)\}+\alpha\{f(X_i)-f_0(X_i)\}\big|\\ &\le \big|\{\Delta_f(X_i)\}\mathbb{1}(Y_i\le f_0(X_i)\}+\{Y_i-f(X_i)\}\mathbb{1}\{f_0(X_i)<Y_i\le f(X_i)\}-\alpha\Delta_f(X_i)\big|\\ &\le \big|\Delta_f(X_i)\big|,\end{aligned}$$

where the first inequality follows from the assumption $\Delta_f(X_i) \leq 0$ and the second inequality is derived from the following inequality

$$f_0(X_i) - f(X_i) \leq \{Y_i - f(X_i)\}\mathbb{1}\{f_0(X_i) < Y_i \leq f(X_i)\} \leq 0.$$

Exchanging the roles of f and f_0 gives the same inequality when $\Delta_f(X_i) > 0$, leading to

$$\big|Z_i(f) - Z_i(f_0)\big| \leq \big|\Delta_f(X_i)\big| = \big|f(X_i) - f_0(X_i)\big|. \tag{C.39}$$

Therefore, we have

$$\begin{aligned}
&\left|\mathbb{E}_{X_i}\left\{\int_0^{Z_i(f)-Z_i(f_0)} \mathbb{1}(|\omega_i + t| > \tau)\mathrm{d}t\right\}\right| \\
&\leq \mathbb{E}_{X_i}\left\{\int_0^{|\Delta_f(X_i)|} \mathbb{1}(|\omega_i + t| > \tau)\mathrm{d}t\right\} \\
&\leq \mathbb{E}_{X_i}\left\{\int_0^{|\Delta_f(X_i)|} \mathbb{1}(|\omega_i| > \tau/2) + \mathbb{1}(|\Delta_f(X_i)| > \tau/2)\mathrm{d}t\right\} \\
&= \mathbb{E}_{X_i}\left\{\int_0^{|\Delta_f(X_i)|} \mathbb{1}(|\omega_i| > \tau/2)\mathrm{d}t\right\},
\end{aligned} \tag{C.40}$$

where the last step follows, provided $\tau \geq 4M_0$ so that $|\Delta_f(X_i)| = |f(X_i) - f_0(X_i)| \leq 2M_0 \leq \tau/2$. By Markov's inequality and Condition 1, it follows that

$$\mathbb{P}(|\omega_i| > \tau/2|X_i) \leq \frac{\mathbb{E}_{X_i}(|\omega_i|^p)}{(\tau/2)^p} \leq \frac{2^p \nu_p}{\tau^p}. \tag{C.41}$$

Combining this with Fubini's theorem gives

$$\left|\mathbb{E}_{X_i}\left\{\int_0^{Z_i(f)-Z_i(f_0)} \mathbb{1}(|\omega_i + t| > \tau)\mathrm{d}t\right\}\right| \leq \frac{2^p \nu_p}{\tau^p}|\Delta_f(X_i)| \leq \frac{2^{p-2}\nu_p}{\tau^{p-1}},$$

where the last inequality follows given $\tau \geq 8M_0$. Finally, for $\mathbb{E}_{X_i}\{\psi_\tau(\omega_i)\}$, note that $|\psi_\tau(t) - t| =$

$(|t|-\tau)\mathbb{1}(|t|>\tau)$. Since $\mathbb{E}_{X_i}(\omega_i)=0$, we obtain

$$\begin{aligned}\big|\mathbb{E}_{X_i}\{\psi_\tau(\omega_i)\}\big| &= \big|\mathbb{E}_{X_i}\big\{\psi_\tau(\omega_i)-\omega_i\big\}\big| \\ &\le \mathbb{E}_{X_i}\big\{(|\omega_i|-\tau)\mathbb{1}(|\omega_i|>\tau)\big\} \\ &\le \frac{\mathbb{E}_{X_i}(|\omega_i|^p)}{\tau^{p-1}} \le \frac{\nu_p}{\tau^{p-1}}. \end{aligned} \tag{C.42}$$

Putting the pieces into (C.36), we have

$$\big|\mathbb{E}_{X_i}\big\{\psi_\tau(Z_i(f)-\alpha g_0(X_i))\big\}\big| \le \frac{\bar{p}}{2}\big\{\Delta_f(X_i)\big\}^2 + \frac{2^{p-1}\nu_p}{\tau^{p-1}},$$

which, combined with Hölder's inequality, further implies

$$|\mathrm{I}| \le \alpha\|g-g_0\|_2\bigg(\frac{\bar{p}}{2}\|f-f_0\|_4^2 + \frac{2^{p-1}\nu_p}{\tau^{p-1}}\bigg), \tag{C.43}$$

when $\tau \ge 8M_0$.

We next turn to bound II. By the definition of ψ'_τ and ω_i, we have

$$\begin{aligned}\mathrm{II} &= \mathbb{E}\bigg[\int_0^{\alpha\Delta_g(X_i)} \mathbb{1}\{|Z_i(f)-\alpha g_0(X_i)+t|\le\tau\}\{\alpha\Delta_g(X_i)-t\}\mathrm{d}t\bigg] \\ &= \mathbb{E}\bigg\{\int_0^{\alpha\Delta_g(X_i)} \big[1-\mathbb{1}\{|\omega_i+Z_i(f)-Z_i(f_0)+t|>\tau\}\big]\{\alpha\Delta_g(X_i)-t\}\mathrm{d}t\bigg\}.\end{aligned}$$

Furthermore, $|Z_i(f)-Z_i(f_0)| \le |f(X_i)-f_0(X_i)| \le 2M_0$ from (C.39). Therefore, we obtain

$$\begin{aligned}&\mathbb{E}_{X_i}\bigg\{\int_0^{\alpha\Delta_g(X_i)} \big[1-\mathbb{1}\{|\omega_i+Z_i(f)-Z_i(f_0)+t|>\tau\}\big]\{\alpha\Delta_g(X_i)-t\}\mathrm{d}t\bigg\} \\ &\ge \mathbb{E}_{X_i}\bigg\{\int_0^{\alpha\Delta_g(X_i)} \big[1-\mathbb{1}\{|\omega_i|>\tau/2\}-\mathbb{1}\{|\Delta_f(X_i)|+|\alpha\Delta_g(X_i)|>\tau/2\}\big]\{\alpha\Delta_g(X_i)-t\}\mathrm{d}t\bigg\} \\ &= \mathbb{E}_{X_i}\bigg\{\int_0^{\alpha\Delta_g(X_i)} \big[1-\mathbb{1}\{|\omega_i|>\tau/2\}\big]\{\alpha\Delta_g(X_i)-t\}\mathrm{d}t\bigg\},\end{aligned} \tag{C.44}$$

as long as $\tau \geq 8M_0$. By Markov's inequality and Condition 1,

$$\mathbb{P}(|\omega_i| > \tau/2 | X_i) \leq \frac{2^p \nu_p}{\tau^p} \leq \frac{1}{2},$$

provided that $\tau \geq 2(2\nu_p)^{1/p}$. Therefore, taking the expectation, we obtain

$$\mathrm{II} \geq \mathbb{E}\left[\alpha^2 \frac{\{\Delta_g(X_i)\}^2}{2}\{1 - \mathbb{P}(|\omega_i| > \tau/2|X_i)\}\right] \geq \mathbb{E}\left[\frac{\alpha^2\{\Delta_g(X_i)\}^2}{4}\right] = \frac{\alpha^2 \|g - g_0\|_2^2}{4}, \quad \text{(C.45)}$$

as long as $\tau \geq \max\{8M_0, 2(2\nu_p)^{1/p}\}$.

Combining (C.43) with (C.45) yields that when $\tau \geq \max\{8M_0, 2(2\nu_p)^{1/p}\}$,

$$\mathscr{R}_\tau(f,g) - \mathscr{R}_\tau(f,g_0) \geq \frac{\alpha^2}{4}\|g - g_0\|_2^2 - \alpha\|g - g_0\|_2\left(\frac{\bar{p}}{2}\|f - f_0\|_4^2 + \frac{2^{p-1}\nu_p}{\tau^{p-1}}\right).$$

Next, we derive the upper bound of the excess joint risk. From the decomposition (C.35), we have the upper bound of the term $|\mathrm{I}|$ as in (C.43). In addition, $0 \leq \psi_\tau'(\cdot) \leq 1$, so that

$$\begin{aligned}
\mathrm{II} &= \mathbb{E}\left[\int_0^{\alpha\Delta_g(X_i)} \psi_\tau'(Z_i(f) - \alpha g_0(X_i) + t)\{\alpha\Delta_g(X_i) - t\}\mathrm{d}t\right] \\
&\leq \mathbb{E}\left[\int_0^{\alpha\Delta_g(X_i)} \{\alpha\Delta_g(X_i) - t\}\mathrm{d}t\right] = \frac{\alpha^2}{2}\|g - g_0\|_2^2. \qquad \text{(C.46)}
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\mathscr{R}_\tau(f,g) - \mathscr{R}_\tau(f,g_0) &\leq \mathbb{E}\left[\int_0^{\alpha\Delta_g(X_i)} \psi_\tau'(Z_i(f) - \alpha g_0(X_i) + t)\{\alpha\Delta_g(X_i) - t\}\mathrm{d}t\right] \\
&\quad + \left|\mathbb{E}\left\{\psi_\tau(Z_i(f) - \alpha g_0(X_i)) \cdot \alpha\Delta_g(X_i)\right\}\right| \\
&\leq \frac{\alpha^2}{2}\|g - g_0\|_2^2 + \alpha\|g - g_0\|_2\left(\frac{\bar{p}}{2}\|f - f_0\|_4^2 + \frac{2^{p-1}\nu_p}{\tau^{p-1}}\right),
\end{aligned}$$

which completes the proof. □

C.2.2 Proof of Lemma C.1.2

The proof follows a similar structure to that of Lemma C.1.1 with the exception that we employ more refined bounds for (C.41) and (C.42) by utilizing the sub-Gaussian property of ω_i.

Recall that $\mathbb{E}_{X_i}$ represents the conditional expectation given X_i. By Markov's inequality, we have

$$\begin{aligned}\mathbb{P}(|\omega_i| > \tau/2|X_i) &= \mathbb{P}\{\exp(\omega_i^2/\sigma_0^2) > \exp(\tau^2/(4\sigma_0^2))|X_i\} \\ &\leq e^{-t^2/(4\sigma_0^2)}\mathbb{E}_{X_i}\{\exp(\omega_i^2/\sigma_0^2)\} \leq 2e^{-\tau^2/(4\sigma_0^2)}, \end{aligned} \tag{C.47}$$

where the last inequality follows from Condition 2. To find a refined bound of $\mathbb{E}_{X_i}\psi_\tau(\omega_i)$, note that $xe^{x^2/2} \leq e^{x^2}$ for any $x \geq 0$. Since $\mathbb{E}_{X_i}\omega_i = 0$, it follows that

$$\begin{aligned}\left|\mathbb{E}_{X_i}\psi_\tau(\omega_i)\right| &= \left|\mathbb{E}_{X_i}\left\{\psi_\tau(\omega_i) - \omega_i\right\}\right| \leq \mathbb{E}_{X_i}\left\{|\omega_i|\mathbb{1}(|\omega_i| > \tau)\right\} \\ &= \sigma_0\mathbb{E}_{X_i}\left\{|\omega_i/\sigma_0|\mathbb{1}(|\omega_i/\sigma_0| > \tau/\sigma_0)\right\} \\ &= \sigma_0\mathbb{E}_{X_i}\left[|\omega_i/\sigma_0|\mathbb{1}\left\{\exp\left(\frac{\omega_i^2}{2\sigma_0^2}\right) > \exp\left(\frac{\tau^2}{2\sigma_0^2}\right)\right\}\right] \\ &\leq \sigma_0 e^{-\tau^2/(2\sigma_0^2)}\mathbb{E}_{X_i}\left\{|\omega_i/\sigma_0|\exp\left(\frac{\omega_i^2}{2\sigma_0^2}\right)\right\} \\ &\leq \sigma_0 e^{-\tau^2/(2\sigma_0^2)}\mathbb{E}_{X_i}\{\exp(\omega_i^2/\sigma_0^2)\} \leq 2\sigma_0 e^{-\tau^2/(2\sigma_0^2)}. \end{aligned} \tag{C.48}$$

Based on these two bounds, we prove the lemma. Provided that $\tau \geq 4M_0$, (C.40) and (C.47) give

$$\begin{aligned}\left|\mathbb{E}_{X_i}\left\{\int_0^{Z_i(f)-Z_i(f_0)}\mathbb{1}(|\omega_i + t| > \tau)\mathrm{d}t\right\}\right| &\leq \mathbb{E}_{X_i}\left\{\int_0^{|\Delta_f(X_i)|}\mathbb{1}(|\omega_i| > \tau/2)\mathrm{d}t\right\} \\ &\leq 2e^{-\tau^2/(2\sigma_0^2)}|\Delta_f(X_i)| \\ &\leq 4M_0 e^{-\tau^2/(2\sigma_0^2)}, \end{aligned}$$

which, together with (C.36), (C.37), (C.38) and (C.48), further implies

$$\begin{aligned}&\left|\mathbb{E}\big\{\psi_\tau(Z_i(f)-\alpha g_0(X_i))\cdot\alpha\Delta_g(X_i)\big\}\right| \\ &\qquad\le \alpha\|g-g_0\|_2\left\{\frac{\bar{p}}{2}\|f-f_0\|_4^2+(4M_0+2\sigma_0)e^{-\tau^2/(2\sigma_0^2)}\right\}. \end{aligned} \tag{C.49}$$

Next, we have from (C.44) that

$$\begin{aligned}&\mathbb{E}\left[\int_0^{\alpha\Delta_g(X_i)}\mathbb{1}\{|Z_i(f)-\alpha g_0(X_i)+t|\le\tau\}\{\alpha\Delta_g(X_i)-t\}\mathrm{d}t\right] \\ &\qquad\ge\mathbb{E}\left\{\mathbb{E}_{X_i}\left[\int_0^{\alpha\Delta_g(X_i)}\big\{1-\mathbb{1}(|\omega_i|>\tau/2)\big\}\{\alpha\Delta_g(X_i)-t\}\mathrm{d}t\right]\right\},\end{aligned}$$

as long as $\tau\ge 8M_0$. By (C.47), note that $\mathbb{P}(|\omega_i|>\tau/2|X_i)\le 1/2$ provided that $\tau\ge 2\sigma_0\sqrt{\log 4}$. Therefore, the earlier expectation bound is further lower bounded as

$$\begin{aligned}&\mathbb{E}\left[\int_0^{\alpha\Delta_g(X_i)}\mathbb{1}\{|Z_i(f)-\alpha g_0(X_i)+t|\le\tau\}\{\alpha\Delta_g(X_i)-t\}\mathrm{d}t\right] \\ &\ge\mathbb{E}\left[\alpha^2\frac{\{\Delta_g(X_i)\}^2}{2}\{1-\mathbb{P}(|\omega_i|>\tau/2|X_i)\}\right]\ge\frac{\alpha^2\|g-g_0\|_2}{4}.\end{aligned}$$

Together, this bound, (C.49) and (C.35) give the lower bound of joint Huber loss.

For the upper bound of joint Huber loss, combining the decomposition (C.35) with (C.46) and (C.49) yields the upper bound. □

C.2.3 Proof of Lemma C.1.3

Recall the definition of ω_i in (C.1). Denote

$$m_g(X_i,\varepsilon_i)=\Delta_g(X_i)\psi_\tau(\omega_i)$$

for any $g \in \mathscr{G}_n$. From the definition of $\psi_\tau(\cdot)$ and the boundedness of $g \in \mathscr{G}_n$ and g_0, we obtain

$$\sup_{g\in\mathscr{G}_n(\eta)} |m_g(X_i,\varepsilon_i)| = \sup_{g\in\mathscr{G}_n(\eta)} |\Delta_g(X_i)| \cdot \tau \le 2M_0\tau,$$

which further implies $\sup_{g\in\mathscr{G}_n(\eta)} |m_g^B(X_i,\varepsilon_i) - \mathbb{E}m_g^B(X_i,\varepsilon_i)| \le 4M_0\tau =: A$. Moreover, since $\psi_\tau(\omega_i) \le |\omega_i|$, it follows that

$$\begin{aligned}
\sup_{g\in\mathscr{G}_n(\eta)} \mathbb{E}\big\{m_g(X_i,\varepsilon_i)\big\}^2 &\le \sup_{g\in\mathscr{G}_n(\eta)} \mathbb{E}\{\Delta_g^2(X_i)\omega_i^2\} \\
&\le \nu_p^{2/p} \sup_{g\in\mathscr{G}_n(\eta)} \mathbb{E}\{g(X_i) - g_0(X_i)\}^2 \\
&\le \nu_p^{2/p}\eta^2,
\end{aligned}$$

where the second inequality follows from Jensen's inequality. We thus have

$$\sup_{g\in\mathscr{G}_n(\eta)} \mathbb{E}\big\{m_g(X_i,\varepsilon_i) - \mathbb{E}m_g(X_i,\varepsilon_i)\big\}^2 \le \sup_{g\in\mathscr{G}_n(\eta)} \mathbb{E}\{m_g(X_i,\varepsilon_i)\}^2 \le \nu_p^{2/p}\eta^2 =: \sigma^2.$$

Denoting $E(\eta) := \mathbb{E}\sup_{g\in\mathscr{G}_n(\eta)} |n^{-1}\sum_{i=1}^n m_g(X_i) - \mathbb{E}m_g(X_i)|$, Lemma C.2.1 implies

$$\mathbb{P}\left\{ \sup_{g\in\mathscr{G}_n(\eta)} \left| \frac{1}{n}\sum_{i=1}^n m_g(X_i,\varepsilon_i) - \mathbb{E}m_g(X_i,\varepsilon_i) \right| \ge 2E(\eta) + \sigma\sqrt{\frac{2x}{n}} + \frac{4Ax}{3n} \right\} \le e^{-x} \tag{C.50}$$

for any $x \ge 0$.

To establish an upper bound of $E(\eta)$, we first find an upper bound of the uniform covering number for the function class $\mathscr{M}_n(\eta) := \{m_g : g \in \mathscr{G}_n(\eta)\}$. For any $g, g' \in \mathscr{G}_n(\eta)$, it follows that

$$\big|m_g(X_i,\varepsilon_i) - m_{g'}(X_i,\varepsilon_i)\big| = \big|\psi_\tau(\varepsilon_i)\big\{\Delta_g(X_i) - \Delta_{g'}(X_i)\big\}\big| \le \tau|g(X_i) - g'(X_i)|.$$

Combining this with Lemma C.2.2 and Lemma C.2.4 yields

$$\begin{aligned}\log N_\infty(\varepsilon, \mathscr{M}_n(\eta), n) &\le \log N_\infty(\varepsilon/(\tau), \mathscr{G}_n, n) \\ &\lesssim \log\left(\frac{2enM_0\tau}{\varepsilon}\right)(NL)^2\log(NL). \end{aligned} \tag{C.51}$$

Now, let $F(X_i, \varepsilon_i) := 2M_0\tau$. Then, F is an envelop function of the function class $\mathscr{M}_n(\eta)$. Denoting $\bar{F} := \max_{1\le i\le n} F(X_i, \varepsilon_i) = F$, we have

$$\|F\|_{Q,2} = \|F\|_2 = 2M_0\tau, \quad \text{and} \quad \|\bar{F}\|_2 = 2M_0\tau$$

for any n-discrete probability measure Q. Therefore, for any n-discrete probability measure Q,

$$\log N(\varepsilon\|F\|_{Q,2}, \mathscr{M}_n(\eta), \|\cdot\|_{Q,2}) \le \log N_\infty(\varepsilon\|F\|_{Q,2}, \mathscr{M}_n(\eta), n) \lesssim (LN)^2\log(LN)\log\left(\frac{en}{\varepsilon}\right).$$

Combining this with Lemma C.2.3, it follows that for any $\eta \ge 1/n$ and $\tau/\nu_p^{1/p} \ge 1$,

$$\begin{aligned} E(\eta) &\lesssim \sigma\cdot LN\sqrt{\frac{\log(LN)}{n}\log\left(\frac{en\cdot 2M_0\tau}{\sigma}\right)} + \frac{(LN)^2\log(LN)\cdot 2M_0\tau}{n}\log\left(\frac{en\cdot 2M_0\tau}{\sigma}\right) \\ &\lesssim \left\{\nu_p^{1/p}\eta\cdot LN\sqrt{\frac{\log(LN)}{n}\log\left(\frac{n\tau}{\nu_p^{1/p}\eta}\right)} + \frac{(LN)^2\log(LN)\tau}{n}\log\left(\frac{n\tau}{\nu_p^{1/p}\eta}\right)\right\} \\ &\lesssim \left\{\nu_p^{1/p}\eta\cdot LN\sqrt{\frac{\log(LN)\log(n^2\tau\nu_p^{-1/p})}{n}} + \frac{\tau(LN)^2\log(LN)\log(n^2\tau\nu_p^{-1/p})}{n}\right\} \\ &= (\nu_p^{1/p}\eta V_{n,\tau,\nu_p} + \tau V_{n,\tau,\nu_p}^2). \end{aligned}$$

Therefore, there exists a universal constant $C_1 > 0$ such that

$$E(\eta) \le C_1\cdot\eta(\nu_p^{1/p} + \sqrt{\tau})V_{n,\tau,\nu_p}$$

for any $\eta \ge \max(\sqrt{\tau}V_{n,\tau,\nu_p}, 1/n)$. Also, if $0 \le x \le n\eta^2/\tau$, we have $\tau x/n \le \eta\sqrt{\tau}\sqrt{x/n}$. Putting

the pieces together in (C.50), there exists a universal constant $c_{15} > 0$ such that for any $\tau/\nu_p^{1/p} \geq 1, \eta \geq \max(\sqrt{\tau}V_{n,\tau,\nu_p}, 1)$ and $0 \leq x \leq n\eta^2/\tau$,

$$\mathbb{P}\left\{ \sup_{g\in\mathscr{G}_n(\eta)} \left| \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})m_g(X_i,\varepsilon_i) \right| \geq c_{15}\cdot\eta(\nu_p^{1/p}+\sqrt{\tau})\left(V_{n,\tau,\nu_p}+\sqrt{\frac{x}{n}}\right) \right\} \leq e^{-x},$$

which completes the proof. □

C.2.4 Proof of Lemma C.1.4

For each $g \in \mathscr{G}_n(\eta)$, define

$$m_g(X_i,\varepsilon_i) := \int_0^{\alpha\Delta_g(X_i)} \big\{\psi_\tau(\omega_i+t)-\psi_\tau(\omega_i)\big\}\mathrm{d}t,$$

and let $\mathscr{M}_n(\eta) := \{m = m_g : g \in \mathscr{G}_n(\eta)\}$. To employ Lemma C.2.1, note that

$$\begin{aligned}\sup_{m\in\mathscr{M}_n(\eta)} |m(X_i,\varepsilon_i)| &= \sup_{g\in\mathscr{G}_n(\eta)} \left| \int_0^{\alpha\Delta_g(X_i)} \big\{\psi_\tau(\omega_i+t)-\psi_\tau(\omega_i)\big\}\mathrm{d}t \right| \\ &\leq \sup_{g\in\mathscr{G}_n(\eta)} \left| \int_0^{\alpha\Delta_g(X_i)} |t|\mathrm{d}t \right| \leq 2\alpha^2 M_0^2,\end{aligned}$$

where the first inequality follows from the Lipschitz property of $\psi_\tau(\cdot)$. Thus, we have

$$\sup_{m\in\mathscr{M}_n(\eta)} |m(X_i,\varepsilon_i)-\mathbb{E}m(X_i,\varepsilon_i)| \leq 4\alpha^2M_0^2 =: A.$$

Also, by the Lipschitz property of $\psi_\tau(\cdot)$ and the boundedness, we have

$$\begin{aligned}\sup_{m\in\mathscr{M}_n(\eta)} \mathbb{E}\big\{m(X_i,\varepsilon_i)\big\}^2 &= \sup_{g\in\mathscr{G}_n(\eta)} \mathbb{E}\left[\int_0^{\alpha\Delta_g(X_i)} \big\{\psi_\tau(\omega_i+t)-\psi_\tau(\omega_i)\big\}\mathrm{d}t \right]^2 \\ &\leq \sup_{g\in\mathscr{G}_n(\eta)} \mathbb{E}\left\{ \int_0^{\alpha\Delta_g(X_i)} |t|\mathrm{d}t \right\}^2 = \sup_{g\in\mathscr{G}_n(\eta)} \frac{\alpha^4}{4}\mathbb{E}\big\{g(X_i)-g_0(X_i)\big\}^4 \\ &\leq \alpha^4M_0^2 \sup_{g\in\mathscr{G}_n(\eta)} \mathbb{E}\big\{g(X_i)-g_0(X_i)\big\}^2 \leq \alpha^4M_0^2\eta^2,\end{aligned}$$

which further implies

$$\sup_{m\in\mathscr{M}_n(\eta)} \mathbb{E}\big\{m(X_i,\varepsilon_i)-\mathbb{E}m(X_i,\varepsilon_i)\big\}^2 \le \sup_{m\in\mathscr{M}_n(\eta)} \mathbb{E}\big\{m(X_i,\varepsilon_i)\big\}^2 \le \alpha^4 M_0^2\eta^2 =: \sigma^2.$$

Then, applying Lemma C.2.1 yields

$$\mathbb{P}\left\{\sup_{m\in\mathscr{M}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^n m(X_i,\varepsilon_i)-\mathbb{E}m(X_i,\varepsilon_i)\right| \ge 2E(\eta)+\sigma\sqrt{\frac{2x}{n}}+\frac{4Ax}{3n}\right\} \le e^{-x} \tag{C.52}$$

for any $x\ge 0$, where $E(\eta)=\mathbb{E}\sup_{m\in\mathscr{M}_n(\eta)}|n^{-1}\sum_{i=1}^n(1-\mathbb{E})m(X_i,\varepsilon_i)|$.

We follow a similar argument as in the proofs of Lemma C.1.3 to derive a bound of $E(\eta)$. By the Lipschitz property of ψ_τ, we have for any $g,g'\in\mathscr{G}_n$ that

$$\begin{aligned}
\big|m_g(X_i,\varepsilon_i)-m_{g'}(X_i,\varepsilon_i)\big| &= \left|\int_{\alpha\Delta_{g'}(X_i)}^{\alpha\Delta_g(X_i)}\big\{\psi_\tau(\omega_i+t)-\psi_\tau(\omega_i)\big\}\mathrm{d}t\right| \\
&\le \left|\int_{\alpha\Delta_{g'}(X_i)}^{\alpha\Delta_g(X_i)}|t|\mathrm{d}t\right| = \frac{\alpha^2}{2}\big|\big\{\Delta_g(X_i)\big\}^2-\big\{\Delta_{g'}(X_i)\big\}^2\big| \\
&\le 2\alpha^2 M_0\big|g(X_i)-g'(X_i)\big|.
\end{aligned}$$

Together with Lemma C.2.2 and Lemma C.2.4, we obtain for any $0<\varepsilon<4\alpha^2M_0^2$ that

$$\log N_\infty(\varepsilon,\mathscr{M}_n(\eta),n) \le \log N_\infty(\varepsilon/(2\alpha^2M_0),\mathscr{G}_n,n) \lesssim (LN)^2\log(LN)\log(4\alpha^2M_0^2ne/\varepsilon).$$

We choose an envelope function $F:=4\alpha^2M_0^2$. Then, for any n-discrete probability Q,

$$\log N(\varepsilon\|F\|_{Q,2},\mathscr{M}_n(\eta),\|\cdot\|_{Q,2}) \lesssim (LN)^2\log(LN)\log(ne/\varepsilon).$$

Thus, by Lemma C.2.3, we have

$$\begin{aligned} E(\eta) &\lesssim \sigma \cdot LN \sqrt{\frac{\log(LN)}{n} \log\left(\frac{en4\alpha^2 M_0^2}{\sigma}\right)} + \frac{(LN)^2 \log(LN) \cdot 4\alpha^2 M_0^2}{n} \log\left(\frac{en4\alpha^2 M_0^2}{\sigma}\right) \\ &\lesssim \alpha^2 \left\{ \eta \cdot LN \sqrt{\frac{\log(LN)}{n} \log\left(\frac{en}{\eta}\right)} + \frac{(LN)^2 \log(LN)}{n} \log\left(\frac{en}{\eta}\right) \right\} \\ &\leq \alpha^2 (\eta V_n + V_n^2), \end{aligned}$$

as long as $\eta \geq 1/n$. Thus, we have $E(\eta) \lesssim \alpha^2 \eta \cdot V_n$ for $\eta \geq V_n$. Also, if $0 \leq x \leq n\eta^2$, then $x/n \leq \eta\sqrt{x/n}$. Putting the pieces together in (C.52), there exists a universal constant $c_{16} > 0$ satisfying

$$\begin{aligned} \mathbb{P}\Bigg\{ \sup_{g \in \mathscr{G}_n(\eta)} \left| \frac{1}{n} \sum_{i=1}^n (1 - \mathbb{E}) \left[\int_0^{\alpha \Delta_g(X_i)} \left\{ \psi_\tau(\omega_i + t) - \psi_\tau(\omega_i) \right\} \mathrm{d}t \right] \right| \\ \geq c_{16} \alpha^2 \eta \left(V_n + \sqrt{\frac{x}{n}} \right) \Bigg\} \leq e^{-x} \end{aligned}$$

for $\eta \geq V_n$. This establishes the claim. □

C.2.5 Proof of Lemma C.1.5

For each given $g \in \mathscr{G}_n(\eta)$ and $f \in \mathscr{F}_n$, define

$$m_{f,g}(X_i, \varepsilon_i) = \int_0^{\alpha \Delta_g(X_i)} \left\{ \psi_\tau(\omega_i + Z_i(f) - Z_i(f_0) + t) - \psi_\tau(\omega_i + t) \right\} \mathrm{d}t,$$

and let $\mathscr{M}_n(\eta) = \{m = m_{f,g} : f \in \mathscr{F}_n \text{ and } g \in \mathscr{G}_n(\eta)\}$. Then, we need to find a high probability bound of the following empirical process,

$$\sup_{m \in \mathscr{M}_n(\eta)} \left| \frac{1}{n} \sum_{i=1}^n (1 - \mathbb{E}) m(X_i, \varepsilon_i) \right|.$$

To apply Lemma C.2.1, it follows from the bounded property of $\mathscr{F}_n, \mathscr{G}_n, f_0$ and g_0 that

$$\begin{aligned}\sup_{m\in\mathscr{M}_n(\eta)}|m(X_i,\varepsilon_i)| &= \sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(\eta)}\left|\int_0^{\alpha\Delta_g(X_i)}\big\{\psi_\tau(\omega_i+Z_i(f)-Z_i(f_0)+t)-\psi_\tau(\omega_i+t)\big\}\mathrm{d}t\right| \\ &\le \sup_{f\in\mathscr{F}_n}\sup_{g\in\mathscr{G}_n(\eta)}\alpha|\Delta_g(X_i)||f(X_i)-f_0(X_i)| \le 4\alpha M_0^2,\end{aligned}$$

where the first inequality follows from (C.39) and

$$\big|\psi_\tau(\omega_i+Z_i(f)-Z_i(f_0)+t)-\psi_\tau(\omega_i+t)\big| \le |Z_i(f)-Z_i(f_0)|.$$

Therefore, we obtain $\sup_{m\in\mathscr{M}_n(\eta)}|m(X_i,\varepsilon_i)-\mathbb{E}m(X_i,\varepsilon_i)| \le 8\alpha M_0^2 =: A$. Moreover, it follows from (C.39) that

$$\sup_{m\in\mathscr{M}_n(\eta)}\mathbb{E}\big\{m(X_i,\varepsilon_i)\big\}^2 \le 4M_0^2\alpha^2\sup_{g\in\mathscr{G}_n(\eta)}\mathbb{E}\big\{g(X_i)-g_0(X_i)\big\}^2 \le 4M_0^2\alpha^2\eta^2,$$

which further implies

$$\sup_{m\in\mathscr{M}_n(\eta)}\mathbb{E}\big\{m(X_i,\varepsilon_i)-\mathbb{E}m(X_i,\varepsilon_i)\big\}^2 \le \sup_{m\in\mathscr{M}_n(\eta)}\mathbb{E}\big\{m(X_i,\varepsilon_i)\big\}^2 \le 4M_0^2\alpha^2\eta^2 =: \sigma^2.$$

Denoting $E(\eta) = \mathbb{E}\sup_{m\in\mathscr{M}_n(\eta)}|n^{-1}\sum_{i=1}^n m(X_i,\varepsilon_i)-\mathbb{E}m(X_i,\varepsilon_i)|$, Lemma C.2.1 gives

$$\mathbb{P}\left\{\sup_{m\in\mathscr{M}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^n m(X_i,\varepsilon_i)-\mathbb{E}m(X_i,\varepsilon_i)\right| \ge 2E(\eta)+\sigma\sqrt{\frac{2x}{n}}+\frac{4Ax}{3n}\right\} \le e^{-x} \tag{C.53}$$

for any $x \ge 0$.

Next, we turn to bounding the expectation, $E(\eta)$. We choose $F = 4\alpha M_0^2$ to be an envelope function of $\mathscr{M}_n(\eta)$. To calculate the uniform covering number of the function class

$\mathscr{M}_n(\eta)$, note that following a similar argument which leads to (C.39) gives that

$$|Z_i(f) - Z_i(f')| \leq |f(X_i) - f'(X_i)|$$

for any $f, f' \in \mathscr{F}_n$. Thus, given $f, f' \in \mathscr{F}_n$ and $g, g' \in \mathscr{G}_n$, we have

$$\begin{aligned}
&|m_{f,g}(X_i, \varepsilon_i) - m_{f',g'}(X_i, \varepsilon_i)| \\
&\leq |m_{f,g}(X_i, \varepsilon_i) - m_{f',g}(X_i, \varepsilon_i)| + |m_{f',g}(X_i, \varepsilon_i) - m_{f',g'}(X_i, \varepsilon_i)| \\
&= \left| \int_0^{\alpha\Delta_g(X_i)} \left\{ \psi_\tau(\omega_i + Z_i(f) - Z_i(f_0) + t) - \psi_\tau(\omega_i + Z_i(f') - Z_i(f_0) + t) \right\} \mathrm{d}t \right| \\
&\qquad + \left| \int_{\alpha\Delta_{g'}(X_i)}^{\alpha\Delta_g(X_i)} \left\{ \psi_\tau(\omega_i + Z_i(f') - Z_i(f_0) + t) - \psi_\tau(\omega_i + t) \right\} \mathrm{d}t \right| \\
&\leq \alpha |g(X_i) - g_0(X_i)| |Z_i(f) - Z_i(f')| + \alpha |g(X_i) - g'(X_i)| |Z_i(f') - Z_i(f_0)| \\
&\leq \alpha \cdot 2M_0 |f(X_i) - f'(X_i)| + \alpha \cdot 2M_0 |g(X_i) - g_0(X_i)|,
\end{aligned}$$

where the second inequality follows from the Lipschitz property of ψ_τ, and the last inequality holds by the bounded property. Thus, it follows that

$$N_\infty(\varepsilon \cdot 4M_0^2 \alpha, \mathscr{M}_n(\eta), n) \leq N_\infty(\varepsilon \cdot M_0, \mathscr{F}_n, n) \cdot N_\infty(\varepsilon \cdot M_0, \mathscr{G}_n, n),$$

which, combined with Lemma C.2.2 and Lemma C.2.4, implies that

$$\log N(\varepsilon \|F\|_{Q,2}, \mathscr{M}_n(\eta), \|\cdot\|_{Q,2}) \lesssim \{(LN)^2 \log(LN) + \mathrm{Pdim}(\mathscr{F}_n)\} \log(en/\varepsilon)$$

for any n-discrete probability Q. Together, this and Lemma C.2.3 give

$$E(\eta) \lesssim \sigma \sqrt{\frac{(LN)^2 \log(LN) + \mathrm{Pdim}(\mathscr{F}_n)}{n} \log\left(\frac{en \cdot 4\alpha M_0^2}{\sigma}\right)} + \frac{4\alpha M_0^2 \cdot \{(LN)^2 \log(LN) + \mathrm{Pdim}(\mathscr{F}_n)\}}{n} \log\left(\frac{en \cdot 4\alpha M_0^2}{\sigma}\right)$$
$$\lesssim \alpha(\eta W_n + W_n^2),$$

as long as $\eta \geq 1/n$. Therefore, there exists a universal constant $C_1 > 0$ satisfying $E(\eta) \leq C_1 \alpha \eta W_n$ for $\eta \geq W_n$. Combining this with (C.53), there exists a universal positive constant $c17$ such that for any $0 \leq x \leq n\eta^2$,

$$\mathbb{P}\left\{ \sup_{h \in \mathscr{H}_n(\eta)} \left| \frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}h(X_i) \right| \geq c_{17} \alpha \eta \left(W_n + \sqrt{\frac{x}{n}} \right) \right\} \leq e^{-x}.$$

This completes the proof. □

C.2.6 Proof of Lemma C.1.6

To begin with, note that

$$\sup_{g \in \mathscr{G}_n(\eta)} \left| \mathbb{E}\left[\psi_\tau(\omega_i) \{ g(X_i) - g_0(X_i) \} \right] \right| \leq \sup_{g \in \mathscr{G}_n(\eta)} \mathbb{E}\left[\left| \mathbb{E}\{ \psi_\tau(\omega_i) | X_i \} \right| \cdot \left| g(X_i) - g_0(X_i) \right| \right]$$
$$\leq 2\sigma_0 e^{-\tau^2/(2\sigma_0^2)} \cdot \eta, \tag{C.54}$$

where the last inequality follows from (C.48). Therefore, it suffices to derive a bound for the tail probabilities of

$$\sup_{g \in \mathscr{G}_n(\eta)} \left| \frac{1}{n} \sum_{i=1}^n \psi_\tau(\omega_i) \{ g(X_i) - g_0(X_i) \} \right|.$$

To this end, we first fix covariates $(X_1, \ldots, X_n)$ and let $\mathbb{E}_X$ and $\mathbb{P}_X$ be the conditional expectation and conditional probability given $(X_1, \ldots, X_n)$, respectively. Consider the stochastic process

$\{S_g : g \in \mathscr{G}_n \cup \{g_0\}\}$, where S_g is defined as

$$S_g := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big\{ \psi_\tau(\omega_i) - \mathbb{E}_X \psi_\tau(\omega_i) \big\} \big\{ g(X_i) - g_0(X_i) \big\}.$$

Since $|\psi_\tau(t)| \leq |t|$, the assumption (C.4) implies that $\mathbb{E}_X \exp(\psi_\tau^2(\omega_i)/\sigma_0^2) \leq 2$. Combining this with Proposition 2.6.1 and Lemma 2.6.8 in Vershynin (2018), there exists a universal constant $C_1 > 0$ such that

$$\mathbb{P}_X(|S_g - S_g'| \geq x) \leq 2 \exp\left(- \frac{x^2}{C_1 \sigma_0^2 \|g - g'\|_n^2} \right) \quad \text{for } g, g' \in \mathscr{G}_n \cup \{g_0\},$$

where $\|\cdot\|_n^2$ is the empirical L_2 norm defined as

$$\|g - g'\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} \big\{ g(X_i) - g'(X_i) \big\}^2.$$

Now, we denote

$$\mathscr{M}_n(v) := \mathscr{M}_n(v; (X_1, \ldots, X_n)) = \{ g \in \mathscr{G}_n \cup \{g_0\} : \|g - g_0\|_n \leq v \}$$

for any $v \geq 0$. Applying Theorem 8.1.6 in Vershynin (2018), there exists an absolute constant $C_2 > 0$ such that for every $v, x \geq 0$,

$$\mathbb{P}_X \left[\sup_{g, g' \in \mathscr{M}_n(v)} |S_g - S_{g'}| \geq C_2 \sigma_0 \left\{ \int_0^{2v} \sqrt{\log N(\varepsilon, \mathscr{G}_n \cup \{g_0\}, \|\cdot\|_n)} \mathrm{d}\varepsilon + v\sqrt{x} \right\} \right] \leq 2e^{-x}. \tag{C.55}$$

For any $(X_1, \ldots, X_n)$, it follows that

$$N(\varepsilon, \mathscr{G}_n \cup \{g_0\}, \|\cdot\|_n) \leq 1 + N_\infty(\varepsilon, \mathscr{G}_n, n).$$

Then, by combining Lemma C.2.2 and Lemma C.2.4, it follows that

$$\begin{aligned}
&\int_0^{2v} \sqrt{\log N(\varepsilon, \mathscr{G}_n \cup \{g_0\}, \|\cdot\|_n)} \mathrm{d}\varepsilon \\
&\leq \int_0^{2v} \sqrt{1 + \log N_\infty(\varepsilon, \mathscr{G}_n, n)} \mathrm{d}\varepsilon \\
&\lesssim \int_0^{2v} \sqrt{1 + (LN)^2 \log(LN) \log(enM_0/\varepsilon)} \mathrm{d}\varepsilon \\
&\lesssim LN\sqrt{\log(LN)} \left\{ v + v\sqrt{\log(enM_0)} + \int_0^{2v} \sqrt{\log(1/\varepsilon) \vee 0} \right\}.
\end{aligned}$$

By the inequality $\int_0^x \sqrt{\log(1/\varepsilon) \vee 0} \mathrm{d}\varepsilon \leq x\sqrt{(1/x) \vee 1}$, we obtain

$$\int_0^{2v} \sqrt{\log(1/\varepsilon) \vee 0} \lesssim v\sqrt{(1/v) \vee 1} \leq v\sqrt{\log n}$$

for any $v \geq 1/n$. Thus, the earlier inequality gives

$$\int_0^{2v} \sqrt{\log N(\varepsilon, \mathscr{G}_n \cup \{g_0\}, \|\cdot\|_n)} \mathrm{d}\varepsilon \lesssim v \cdot LN\sqrt{\log(LN) \log n},$$

which, combined with (C.55), further implies that for any $x \geq 0$,

$$\mathbb{P}_X \left[\sup_{g, g' \in \mathscr{M}(v)} |S_g - S_{g'}| \geq C_3 \sigma_0 \left\{ v\sqrt{(LN)^2 \log(LN) \log n} + v\sqrt{x} \right\} \right] \leq 2e^{-x},$$

as long as $v \geq 1/n$, where C_3 is a universal constant. Since $g_0 \in \mathscr{M}(v)$ for any $v \geq 0$, this tail probability further implies with probability at least $1 - 2e^{-x}$ (conditioned on $(X_1, \ldots, X_n)$) that

$$\sup_{\substack{g \in \mathscr{G}_n \\ \|g - g_0\|_n \leq v}} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \psi_\tau(\omega_i) - \mathbb{E}_X \psi_\tau(\omega_i) \right\} \left\{ g(X_i) - g_0(X_i) \right\} \right| \leq C_3 \sigma_0 v \left(V_n + \sqrt{\frac{x}{n}} \right)$$

for $v \geq 1/n$ and $x \geq 0$. Moreover, it follows from the Cauchy-Schwartz inequality that

$$\sup_{\|g-g_0\|_n \leq v} \left| \frac{1}{n} \sum_{i=1}^{n} \{\mathbb{E}_X \psi_\tau(\omega_i)\}\{g(X_i) - g_0(X_i)\} \right| \leq \left[\frac{1}{n} \sum_{i=1}^{n} \{\mathbb{E}_X \psi_\tau(\omega_i)\}^2 \right]^{1/2} v$$
$$\leq 2\sigma_0 v e^{-\tau^2/(2\sigma_0^2)},$$

where the last inequality follows from (C.48). Together, this bound and the earlier tail probability imply that with probability at least $1 - 2e^{-x}$,

$$\sup_{\substack{g \in \mathscr{G}_n \\ \|g-g_0\|_n \leq v}} \left| \frac{1}{n} \sum_{i=1}^{n} \psi_\tau(\omega_i)\{g(X_i) - g_0(X_i)\} \right| \leq \max(C_3, 2)\sigma_0 v \left\{ V_n + e^{-\tau^2/(2\sigma_0^2)} + \sqrt{\frac{x}{n}} \right\}. \quad \text{(C.56)}$$

Note that when $\tau = \infty$, we have

$$\frac{1}{n} \sum_{i=1}^{n} (1 - \mathbb{E}) \left[\int_0^{\alpha \Delta_g(X_i)} \{\psi_\tau(\omega_i + t) - \psi_\tau(\omega_i)\} \mathrm{d}t \right] = \frac{\alpha^2}{2} \left(\|g - g_0\|_n^2 - \|g - g_0\|_2^2 \right).$$

Therefore, Lemma C.1.4 with $\tau = \infty$ implies that for $\eta \geq V_n$ and $0 \leq x \leq n\eta^2$ that

$$\sup_{g \in \mathscr{G}_n(\eta)} \left| \|g - g_0\|_n^2 - \|g - g_0\|_2^2 \right| \leq 2c_{16}\eta \left(V_n + \sqrt{\frac{x}{n}} \right) \leq 4c_{16}\eta^2 \quad \text{(C.57)}$$

with probability at least $1 - e^{-x}$. Conditioned on the event where the inequality (C.57) holds, we obtain

$$\sup_{g \in \mathscr{G}_n(\eta)} \|g - g_n\|_n^2 \leq \sup_{g \in \mathscr{G}_n(\eta)} \left| \|g - g_0\|_n^2 - \|g - g_0\|_2^2 \right| + \sup_{g \in \mathscr{G}_n(\eta)} \|g - g_0\|_2^2 \leq (1 + 4c_{16})\eta^2.$$

Thus, for the event $\mathscr{B}(\eta)$ defined as

$$\mathscr{B}(\eta) := \left\{ \sup_{g \in \mathscr{G}_n(\eta)} \|g - g_0\|_n \leq (1 + 4c_{16})^{1/2} \eta \right\},$$

we have $\mathbb{P}\{\mathscr{B}(\eta)\} \geq 1 - e^{-x}$ for any $0 \leq x \leq n\eta^2$. Therefore, denoting $C_4 = \max(C_3, 2)(1 +$

$4c_{16})^{1/2}$, we obtain

$$\begin{aligned}
&\mathbb{P}\left[\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_\tau(\omega_i)\{g(X_i)-g_0(X_i)\}\right| \geq C_4\sigma_0\eta\left\{V_n+e^{-\tau^2/(2\sigma_0^2)}+\sqrt{\frac{x}{n}}\right\}\right] \\
&\leq \mathbb{P}\left[\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_\tau(\omega_i)\{g(X_i)-g_0(X_i)\}\right| \geq C_4\sigma_0\eta\left\{V_n+e^{-\tau^2/(2\sigma_0^2)}+\sqrt{\frac{x}{n}}\right\}\middle|\mathscr{B}(\eta)\right]+e^{-x} \\
&\leq 3e^{-x},
\end{aligned}$$

where the last inequality follows from (C.56) after taking $v=(1+4c_{16})^{1/2}\eta$. Combining this with (C.54) gives

$$\begin{aligned}
&\mathbb{P}\left[\sup_{g\in\mathscr{G}_n(\eta)}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\psi_\tau(\omega_i)\{g(X_i)-g_0(X_i)\}\right| \geq \underbrace{(C_4+2)}_{=:c_{18}}\sigma_0\eta\left\{V_n+e^{-\tau^2/(2\sigma_0^2)}+\sqrt{\frac{x}{n}}\right\}\right] \\
&\leq 3e^{-x} \qquad\qquad (C.58)
\end{aligned}$$

for $\eta\geq V_n$ and $0\leq x\leq n\eta^2$. This completes the proof. □

C.2.7 Proof of Lemma C.1.7

We first prove the lower bound. From the Lipschitz continuity of $p_{\varepsilon|X}(\cdot)$, it follows that $p_{\varepsilon|X}(t)\geq \underline{p}/2$ when $|t|\leq \underline{p}/(2l_0)$. Then, we apply Lemma S6 in the supplement of Padilla and Chatterjee (2022) to obtain

$$\mathscr{Q}_\alpha(f)-\mathscr{Q}_\alpha(f_0)\geq \min\left(\frac{\underline{p}}{4},\frac{\underline{p}^2}{16l_0}\right)\mathbb{E}\min\left[|f(X)-f_0(X)|,\{f(X)-f_0(X)\}^2\right].$$

On the other hand, we have $\{f(X)-f_0(X)\}^2\leq 2M_0|f(X)-f_0(X)|$ by the definition of $\mathscr{F}_n$. Combining this with the earlier inequality and the assumption that $M_0\geq 1$, we have the desired lower bound of the excess quantile risk.

We next prove the upper bound. From Knight's inequality (see, e.g., Knight (1998)), for

any $u, v \in \mathbb{R}$, it follows that

$$\rho_\tau(u-v) - \rho_\tau(u) = -v\{\tau - \mathbb{1}(u \le 0)\} + \int_0^v \{\mathbb{1}(u \le t) - \mathbb{1}(u \le 0)\}\mathrm{d}t.$$

Taking expectation this equality with $u = \varepsilon_i$ and $v = f(X_i) - f_0(X_i)$, we obtain

$$L_\tau(f) - L_\tau(f_0) = \mathbb{E}\int_0^{f(X)-f_0(X)} \int_0^t p_{\varepsilon|X}(s)\mathrm{d}s\mathrm{d}t \le \frac{\bar{p}}{2}\|f - f_0\|_2^2,$$

where the last inequality follows from Condition 3. This concludes the proof. □

C.2.8 Proof of Lemma C.1.8

For each $f \in \mathscr{F}_n(\delta)$, denote

$$m_f(X_i, \varepsilon_i) := \rho_\alpha(Y_i - f(X_i)) - \rho_\alpha(Y_i - f_0(X_i)).$$

Since $\rho_\alpha(\cdot)$ is a Lipschitz function, we have

$$\sup_{f \in \mathscr{F}_n(\delta)} |m_f(X_i, \varepsilon_i)| \le \sup_{f \in \mathscr{F}_n(\delta)} |f(X_i) - f_0(X_i)| \le 2M_0.$$

Therefore, $\sup_{f \in \mathscr{F}_n(\delta)} |m_f(X_i, \varepsilon_i) - \mathbb{E}m_f(X_i, \varepsilon_i)| \le 4M_0 =: A$. Moreover,

$$\sup_{f \in \mathscr{F}_n(\delta)} \mathbb{E}\big\{m_f(X_i, \varepsilon_i)\big\}^2 \le \sup_{f \in \mathscr{F}_n(\delta)} \mathbb{E}\big\{f(X_i) - f_0(X_i)\big\}^2 \le \delta^2,$$

which further implies

$$\sup_{f \in \mathscr{F}_n(\delta)} \mathbb{E}\big\{m_f(X_i, \varepsilon_i) - \mathbb{E}m_f(X_i, \varepsilon_i)\big\}^2 \le \sup_{f \in \mathscr{F}_n(\delta)} \mathbb{E}\big\{m_f(X_i, \varepsilon_i)\big\}^2 \le \delta^2 =: \sigma^2.$$

Denoting $E(\delta) = \mathbb{E}\sup_{f\in\mathscr{F}_n(\delta)} |n^{-1}\sum_{i=1}^n m_f(X_i,\varepsilon_i) - \mathbb{E}m_f(X_i,\varepsilon_i)|$, Lemma C.2.1 gives

$$\mathbb{P}\left\{ \sup_{f\in\mathscr{F}_n(\delta)} \left| \frac{1}{n}\sum_{i=1}^n m_f(X_i,\varepsilon_i) - \mathbb{E}m_f(X_i,\varepsilon_i) \right| \geq 2E(\delta) + \sigma\sqrt{\frac{2x}{n}} + \frac{4Ax}{3n} \right\} \leq e^{-x} \tag{C.59}$$

for any $x \geq 0$.

Now, we find an upper bound of the expectation $E(\delta)$. We denote $\mathscr{M}_n(\delta) := \{m_f(X_i,\varepsilon_i) : f \in \mathscr{F}_n(\delta)\}$. Combining the Lipschitz continuity of $\rho_\alpha(\cdot)$ with Lemma C.2.2 and Lemma C.2.4 gives that for any $\varepsilon \in (0, M_0)$,

$$\log \mathscr{N}_\infty(\varepsilon, \mathscr{M}_n(\delta), n) \leq \log \mathscr{N}_\infty(\varepsilon, \mathscr{F}_n, n) \lesssim (LN)^2 \log(LN) \log(M_0 ne/\varepsilon).$$

Also, the Lipschitz property of $\rho_\alpha(\cdot)$ implies that $F = 2M_0$ is an envelope function of $\mathscr{M}_n(\delta)$. Thus, for any n-discrete probability measure Q,

$$\log N(\varepsilon \|F\|_{Q,2}, \mathscr{M}_n(\delta), \|\cdot\|_{Q,2}) \lesssim (LN)^2 \log(LN) \log(en/(2\varepsilon)).$$

Applying Lemma C.2.3, we have

$$\begin{aligned} E(\delta) &\lesssim \sigma\sqrt{\frac{(LN)^2\log(LN)}{n}\log\left(\frac{enM_0}{\sigma}\right)} + 2M_0\frac{(LN)^2\log(LN)}{n}\log\left(\frac{enM_0}{\sigma}\right) \\ &\lesssim \delta V_n + V_n^2 \end{aligned}$$

for any $\delta \geq 1/n$. Thus, when $\delta \geq V_n$, we have $E(\delta) \lesssim \delta V_n$. By combining this and (C.59), there exists a universal positive constant $c_{21} > 0$ such that

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{f\in\mathscr{F}_n(\delta)} \left| \frac{1}{n}\sum_{i=1}^n m_f(X_i,\varepsilon_i) - \mathbb{E}m_f(X_i,\varepsilon_i) \right| \geq c_{21}\delta\left(V_n + \sqrt{\frac{x}{n}}\right) \right\} \\ &\leq e^{-x} \end{aligned}$$

holds for any $0 \leq x \leq n\delta^2$ and $\delta \geq V_n$. This completes the proof. $\square$

Appendix D

Supplementary Material for Chapter 4

D.1 Some Comparisons with Previous Studies

COMPARISON WITH EXISTING WORK ON QUANTILE KERNEL RIDGE REGRESSION. Nonparametric quantile regression using the kernel method has been extensively studied in the literature. For instance, Takeuchi et al. (2006) and Li et al. (2007) explored quantile KRR estimators and developed efficient algorithms for their implementation. They also established theoretical properties of the estimators, with a focus on excess risk analysis under strong assumptions, such as a uniformly bounded function class or bounded quantile residuals. Furthermore, their analysis is confined to the scenario where $r_q = 0$. Under the assumption of bounded response variables, Steinwart and Christmann (2011) derived convergence rates in the L_2-norm for quantile KRR estimators, which are minimax optimal.

Using kernels with eigenvalues that decay polynomially, Lian (2022) established the L_2 convergence rate for Q-KRR estimators without imposing the above restrictive assumptions. Upon closer examination, we identify a potential minor gap in the proof of Theorem 1 therein, which relies on a local strong convexity of the expected risk that for some constant $C > 0$, $\mathbb{E}\{\rho_\tau(Y_i - f(X_i)) - \rho_\tau(Y_i - f_0(X_i))\} \geq C\|f - f_0\|_2^2$ whenever $\|f - f_0\|_{\mathscr{H}} \leq 1$. However, in the proof of Theorem 1, this local strong convexity is used without ensuring that $\widehat{f}$ satisfies $\|\widehat{f} - f_0\|_{\mathscr{H}} \leq 1$. Our result extends Theorem 1 of Lian (2022) to any RKHS satisfying Condition 4.2.1. We present a self-contained proof of Theorem 4.3.1, addressing the above gap in Lian (2022).

Furthermore, our finding eliminates the need for the stringent moment/boundedness assumptions frequently imposed in the literature.

COMPARISON WITH EXISTING WORK ON FUNCTIONAL BAHADUR REPRESENTATION. Recall that $\widehat{g}_{\text{ora}}$ is a standard KRR estimator obtained by regressing $Z_i(f_0)/\tau$ on X_i. The functional Bahadur representation provided in Theorem 4.3.3 is of independent interest and, more importantly, improves the existing results in certain cases, as we explain below.

Shang and Cheng (2013) studied penalized nonparametric estimators when $\mathbb{E}(Y|X) = F(g_0(X))$ for some known link function $F(\cdot)$, and the unknown function g_0 belongs to an RKHS with eigenvalues that decay polynomially with exponent $\beta > 1$. Their estimator reduces to the least squares KRR estimator when F is the identity function. Under our notations, Theorem 3.5 and Lemma 3.1 in Shang and Cheng (2013) imply that with high probability,

$$\left\| \tau(\widehat{g}_{\text{ora}} - g_{\lambda_e}) - \frac{1}{n}\sum_{i=1}^{n} \omega_i (T_K + \lambda_e I)^{-1} K_{X_i} \right\|_\infty \lesssim \mathfrak{D}_{\lambda_e}^{1/2} \Delta_{\text{SC}}(\lambda_e, n),$$

where $\Delta_{\text{SC}}(\lambda_e) := (\lambda_e^{-1/(2\beta)} n^{-1/2} + \lambda_e^{1/2}) \lambda_e^{-(3\beta-1)/(2\beta^2)} (\log\log n)^{1/2} n^{-1/2}$. By Example B.2, $\mathfrak{D}_{\lambda_e} \asymp \lambda_e^{-1/\beta}$ when the kernel has β-polynomially decaying eigenvalues. Applying Theorem 4.3.3 with $r_e = 0$ yields $\Delta_1(\lambda_e) \asymp (\lambda_e^{-1/(2\beta)} n^{-1/2} + \lambda_e^{1/2}) \lambda_e^{-1/(2\beta)} \sqrt{\log(n)/n}$. Since $(3\beta-1)/(2\beta^2) > 1/(2\beta)$ for any $\beta > 1/2$, we have $\Delta_{\text{SC}} > \Delta_1$ for all $0 < \lambda_e < 1$. Furthermore, for $\Delta_{\text{SC}}(\lambda_e) = o(n^{-1/2})$ to be achieved, it is necessary to have $\beta > (3+\sqrt{5})/2 \approx 2.618$, whereas we only need $\beta > 1$.

D.2 Statistical Theory for Finite-rank, Polynomial and Exponential Decay Kernels

We first explore three classes of kernel functions: finite-rank kernels, polynomial decay kernels, and exponential decay kernels, while determining their respective effective dimensions.

Example 1 (Finite-rank kernels)**.** The kernel K is considered to have a finite rank of m when its

eigenvalues $\mu_j = 0$ for all $j > m$. An illustrative example is the m-th order polynomial kernel, defined as $K(x_1, x_2) = (1 + x_1^{\mathrm{T}} x_2)^m$, $x_1, x_2 \in \mathbb{R}^d$, which has a rank of $\binom{d+m}{m}$. This class of kernels encompasses linear and polynomial functions, and more broadly, any function class based on finite dictionary vectors. It is easy to see that $\mathfrak{D}_\lambda \leq m$ for any $\lambda > 0$ when the kernel has a finite rank of m.

Example 2 (Polynomial decay kernels)**.** Another commonly used class of kernels has eigenvalues that exhibit β-polynomial decay $\mu_j \asymp j^{-\beta}$ for some $\beta > 1$. Typical examples include spline kernels, Sobolev kernels, and Laplacian kernels. To calculate the effective dimensions of such kernels, note that $\mu_j/(\mu_j + \lambda) \asymp 1/(1 + \lambda j^\beta)$ when the kernel K has eigenvalues with β-polynomial decay. By bounding the sums with integrals, we obtain $\mathfrak{D}_\lambda \asymp \sum_{j=1}^{\infty} (1 + \lambda \cdot j^\beta)^{-1} \asymp c_\beta \lambda^{-1/\beta}$, where c_β is a positive constant depending only on β.

Example 3 (Exponential decay kernels)**.** The eigenvalues $\{\mu_j\}_{j \geq 1}$ exhibit β-exponential decay if, for some $\beta > 0$, they satisfy $\mu_j \asymp \exp(-c_\beta j^\beta)$ for some $c_\beta > 0$. This class of kernels includes Gaussian kernels defined as $K(x, x') = \exp(-\|x - x'\|_2^2/\sigma^2)$ for $x, x' \in \mathscr{X}$ and $\sigma^2 > 0$, where $\|\cdot\|_2$ denotes the Euclidean distance. Similar to the case of polynomial decay kernels, the effective dimension can be computed by bounding the sums with integrals. For $0 < \lambda < 1$, it holds $\mathfrak{D}_\lambda \asymp \sum_{j=1}^{\infty} (1 + \lambda \cdot e^{c_\beta j^\beta})^{-1} \asymp C \log^{1/\beta}(1/\lambda)$, where the constant C depends only on β and c_β.

By applying the theoretical results from Section 4.3, we explicitly provide upper bounds for the convergence rate of the proposed estimator for each type of RKHS, followed by pointwise asymptotic normality. For ease of presentation, we impose the following condition throughout this section.

Condition D.2.1 (Source condition)**.** The true conditional quantile and expected shortfall functions satisfy $f_0 = T_K^{r_q} f^*$ and $g_0 = T_K^{r_e} g^*$ for some $f^*, g^* \in \mathbb{B}_{\mathscr{H}}(1)$ and $0 \leq r_q, r_e \leq 1/2$.

D.2.1 Finite-rank kernels

We first consider a kernel with a finite rank of m for some $m \in \mathbb{N}$. In this case, we have $\mathfrak{D}_\lambda \leq m$ for any $\lambda > 0$. Combining this with Theorems 4.3.1 and 4.3.2, we obtain the following corollary for kernels with a finite rank.

Corollary D.2.1 (Convergence rates under finite-rank kernels)**.** Assume Conditions 4.2.1, 4.3.1, 4.3.3 and D.2.1 hold. For any $t > 0$, $(\widehat{f}, \widehat{g}) = (\widehat{f}_n(\lambda_q), \widehat{g}_n(\lambda_e))$ with $\lambda_q \asymp \lambda_e \asymp (m+t)/n$ satisfy that, with probability at least $1 - 7e^{-t}$,

$$\|\widehat{f} - f_0\|_2 \lesssim \sqrt{\frac{m+t}{n}} \quad \text{and} \quad \tau\|\widehat{g} - g_0\|_2 \lesssim \sqrt{\frac{m+t}{n}}$$

as long as $n \gtrsim m\log(n)$.

The above rates are minimax-optimal, as shown in Theorem 2 of Raskutti, Wainwright and Yu (2012). Next, we establish the pointwise asymptotic normality of the ES-KRR estimator under finite-rank kernels and verify the validity of the bootstrap procedure.

Corollary D.2.2 (Pointwise asymptotic normality under finite-rank kernels)**.** Assume that the same conditions as in Corollary D.2.1 hold with $t = \log n$, $m = o(n^{1/3})$ and $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m > 0$. If $\rho^2_{\lambda_e}(x_0) \to \rho^2(x_0)$ for some $\rho^2(x_0) > 0$, the two-step ES-KRR estimator $\widehat{g}$ satisfies $\tau\sqrt{n/m}(\widehat{g} - g_0)(x_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \rho^2(x_0))$.

Corollary D.2.3 (Validity of bootstrap under finite-rank kernels)**.** Assume that the same conditions as in Corollary D.2.2 hold. If $x_0 \in \mathscr{X}$ satisfies $\rho^2_\lambda(x_0) > C$ for any sufficiently small $\lambda > 0$ with a constant $C > 0$, then we have $|\mathbb{P}\{g_0(x_0) \in \mathscr{I}^\flat_\alpha(x_0)\} - (1-\alpha)| = o(1)$ for any $\alpha \in (0,1)$.

D.2.2 Polynomial decay kernels

We next consider an RKHS whose kernel has β-polynomially decaying eigenvalues for some $\beta > 1$, that is, $\mu_j \asymp j^{-\beta}$ for $j \geq 1$. Recall that the effective dimension satisfies $\mathfrak{D}_\lambda \asymp \lambda^{-1/\beta}$. The following corollary establishes non-asymptotic L_2-error bounds for $\widehat{f}$ and $\widehat{g}$.

Corollary D.2.4 (Convergence rates under polynomial decay kernels)**.** Assume Conditions 4.2.1–4.3.3 and D.2.1 hold. For any $t \in (0,n)$, we choose

$$\lambda_q \asymp n^{-\beta/\{(2r_q+1)\beta+1\}} + \frac{t}{n} \quad \text{and} \quad \lambda_e \asymp n^{-\beta/\{(2r_e+1)\beta+1\}} + \frac{t}{n}.$$

Then, $\widehat{f}$ and $\widehat{g}$ satisfy that, with probability at least $1-7e^{-t}$,

$$\|\widehat{f} - f_0\|_2 \lesssim n^{-(2r_q+1)\beta/\{(4r_q+2)\beta+2\}} + \sqrt{\frac{t}{n}} \quad \text{and} \quad \tau\|\widehat{g} - g_0\|_2 \lesssim n^{-e^*} + \sqrt{\frac{t}{n}},$$

where

$$e^* = \min\left\{\frac{(2r_e+1)\beta}{(4r_e+2)\beta+2}, \frac{(4r_q+2)\beta-1}{(4r_q+2)\beta+2}\right\}.$$

In particular, if (r_q, r_e, β) satisfy

$$\{(2r_q+1)\beta - 2\}\{(2r_e+1)\beta+2\} \geq -3, \tag{D.1}$$

then with the same probability,

$$\|\widehat{f} - f_0\|_2 \lesssim n^{-\frac{(2r_q+1)\beta}{(4r_q+2)\beta+2}} + \sqrt{\frac{t}{n}} \quad \text{and} \quad \tau\|\widehat{g} - g_0\|_2 \lesssim n^{-\frac{(2r_e+1)\beta}{(4r_e+2)\beta+2}} + \sqrt{\frac{t}{n}}.$$

An immediate consequence of Corollary D.2.4 is that $\|\widehat{f} - f_0\|_2 = \mathscr{O}_{\mathbb{P}}(n^{-\frac{(2r_q+1)\beta}{(4r_q+2)\beta+2}})$, implying that $\widehat{f}$ attains the minimax optimal convergence rate by Theorem 4 of Suzuki and Sugiyama (2013) with $d = M = 1$ in their notations. Given that $0 \leq r_q, r_e \leq 1/2$, a sufficient condition for (D.1) is $\beta \geq (\sqrt{3}+1)/2 \approx 1.366$; see Lemma D.4.1 for details. Thus, under this mild condition on β, even when $r_e > r_q$, $\widehat{g}$ achieves the minimax optimal convergence rate, $\tau\|\widehat{g} - g_0\|_2 = \mathscr{O}_{\mathbb{P}}(n^{-\frac{(2r_e+1)\beta}{(4r_e+2)\beta+2}})$. If (r_q, r_e, β) fail to satisfy (D.1), implying $r_e > r_q$, $\widehat{g}$ does not attain the above minimax optimal convergence rate. Nevertheless, as e^* is strictly larger than

$(2r_q+1)\beta/\{(4r_q+2)\beta+2\}$, $\widehat{g}$ achieves a faster convergence rate compared to the nuisance estimator $\widehat{f}$. This is attributed to the orthogonal property (4.5) of the score function.

Remark D.2.1 (An alternative assumption to condition (D.1))**.** With an additional strong assumption on the RKHS, $\widehat{g}$ can attain an optimal convergence rate for any (r_q, r_e, β). Specifically, assume that the function K_{X_i} satisfies the following L_4–L_2 norm equivalence: $(\mathbb{E}\langle K_{X_i}, h\rangle_{\mathscr{H}}^4)^{1/4} \leq \kappa(\mathbb{E}\langle K_{X_i}, h\rangle_{\mathscr{H}}^2)^{1/2}$ for any $h \in \mathscr{H}$, where $\kappa > 0$ is a dimension-free constant. This L_4–L_2 norm equivalence can be interpreted as the uniform boundedness of the kurtosis of the one-dimensional marginal $\langle K_{X_i}, h\rangle_{\mathscr{H}}$ for any direction $h \in \mathscr{H}$. Then, by the definition of the RKHS, the equivalence implies that $\|h\|_4^2 = (\mathbb{E}\langle K_{X_i}, h\rangle_{\mathscr{H}}^4)^{1/2} \leq \kappa^2 \mathbb{E}\langle K_{X_i}, h\rangle_{\mathscr{H}}^2 = \kappa^2 \|h\|_2^2$ for any $h \in \mathscr{H}$. Following a similar line of arguments as in the proof of Corollary D.2.4, we can conclude that $\widehat{g}$ attains the optimal convergence rate for any $0 \leq r_q, r_e, \leq 1/2$, and $\beta > 1$.

However, we note that, to the best of our knowledge, there are no specific settings in which a polynomial decay kernel satisfies the L_4–L_2 norm equivalence. One sufficient condition for the equivalence is that K_{X_i} is a Gaussian random element in $\mathscr{H}$, or more generally, $\{\langle K_{X_i}, h\rangle_{\mathscr{H}} : h \in \mathscr{H}\}$ is a sub-Gaussian function class (Lecué and Mendelson, 2013). Nevertheless, it remains uncertain when these conditions can be fulfilled. Thus, our analysis does not assume this stringent theoretical conditions.

Next, by combining the bound for the effective dimension with Corollary 4.3.1 and Theorem 4.3.6, we establish the pointwise asymptotic normality of the ES-KRR estimator and verify the validity of the confidence interval $\mathscr{I}_\alpha^\flat(x_0)$, defined in (4.15) for any $\alpha \in (0,1)$, respectively.

Corollary D.2.5 (Pointwise asymptotic normality under polynomial decay kernels)**.** Assume Conditions 4.2.1–4.3.3 and D.2.1 hold. Moreover, we assume that $(2r_q+1)\beta > 2, (2r_e+1)\beta > 3$ and $0 < r_e \leq 1/2$, and choose $\lambda_q \asymp n^{-\beta/\{(2r_q+1)\beta+1\}}$ and $\lambda_e \asymp n^{-\iota}$ with $1/(2r_e+1) < \iota < \min(1, \beta/2)$. If $\rho_{\lambda_e}^2(x_0) \to \rho^2(x_0)$ for some $\rho^2(x_0) > 0$, the two-step ES-KRR estimator $\widehat{g}$

satisfies

$$\tau\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}(\widehat{g}-g_0)(x_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\rho^2(x_0)).$$

Corollary D.2.6 (Validity of bootstrap under polynomial decay kernels)**.** Assume that the same conditions as in Corollary D.2.5 hold. If $x_0 \in \mathscr{X}$ satisfies $\rho_\lambda^2(x_0) > C$ for any sufficiently small $\lambda > 0$ with a constant $C > 0$, then we have $|\mathbb{P}\{g_0(x_0) \in \mathscr{I}_\alpha^\flat(x_0)\} - (1-\alpha)| = o(1)$ for any $\alpha \in (0,1)$.

As a concrete example, consider the p-th order periodic Sobolev space $H_0^p[0,1]$ defined as follows:

$$H_0^p[0,1] := \Bigg\{ h : [0,1] \to \mathbb{R} : h^{(j)} \text{ is absolutely continuous and satisfies } h^{(j)}(0) = h^{(j)}(1) \\ \text{for } j = 0,1,\ldots,p-1, \text{ and } \int_0^1 \{h^{(p)}(x)\}^2 \mathrm{d}x < \infty \Bigg\},$$

where p is larger than $1/2$ and $h^{(j)}$ denotes the j-th derivative of h. Assuming that $\mathbb{P}_X$ is the uniform distribution on $[0,1]$, the corresponding Sobolev kernels are

$$K(x,x') = 1 + \sum_{j=1}^{\infty} \frac{2\cos(2\pi j(x-x'))}{(2\pi j)^{2p}};$$

see Chapter 4 in Gu (2013) for details. The Sobolev kernel has uniformly bounded eigenfunctions

$$\phi_j(x) = \begin{cases} 1, & j = 0, \\ \sqrt{2}\cos(j\pi x), & j = 2k \text{ for } k = 1,2,\ldots, \\ \sqrt{2}\sin((j+1)\pi x), & j = 2k-1 \text{ for } k = 1,2,\ldots, \end{cases} \tag{D.2}$$

and the corresponding eigenvalues are

$$\mu_j = \begin{cases} 1, & j = 0, \\ (j\pi)^{-2p}, & j = 2k \text{ for } k = 1,2,\ldots, \\ \{(j+1)\pi\}^{-2p}, & j = 2k-1 \text{ for } k = 1,2,\ldots. \end{cases} \tag{D.3}$$

Moreover, by following a similar argument as in the proof of Lemma 4.1 in Zhao, Cheng and Liu (2016), there exists a lower bound for $\rho_\lambda^2(x_0)$ for any sufficiently small $\lambda > 0$ and $x_0 \in [0,1]$, affirming that this kernel satisfies the condition in Corollary D.2.6.

Lemma D.2.1 (Lower bound for $\rho_\lambda^2(x_0)$ under periodic Sobolev spaces.)**.** Assume that $\mathbb{P}_X$ is the uniform distribution over $[0,1]$ and $\mathscr{H} = H_0^p[0,1]$ with eigenfunctions (D.2) and eigenvalues (D.3). Moreover, suppose that there exists an absolute constant $\underline{\sigma} > 0$ satisfying $\mathbb{E}(\omega_i^2|X_i) \geq \underline{\sigma}^2$. Then, there exists a constant $c_{11} = c_{11}(p) > 0$ such that for any $\lambda \in (0, (2\pi)^{-2p})$ and $x_0 \in [0,1]$, we have $\rho_\lambda^2(x_0) \geq c_{11}\underline{\sigma}^2$.

D.2.3 Exponential decay kernels

Finally, we consider an RKHS whose kernel has β-exponentially decaying eigenvalues for some $\beta > 0$, that is, $\mu_j \asymp \exp(-c_\beta j^\beta)$ for some $c_\beta > 0$. Since the effective dimension of this type of kernels satisfies $\mathfrak{D}_\lambda \asymp \log^{1/\beta}(1/\lambda)$ for any $\lambda \in (0,1)$, applying Theorem 4.3.1 and Theorem 4.3.2 establishes non-asymptotic L_2-error bounds for $\widehat{f}$ and $\widehat{g}$ as follows.

Corollary D.2.7 (Convergence rates under exponential decay kernels)**.** Assume Conditions 4.2.1–4.3.3 and D.2.1 hold. For any $t > 0$, $\widehat{f}$ and $\widehat{g}$ with $\lambda_q \asymp \lambda_e \asymp \{t + \log^{1/\beta}(n)\}/n$ satisfy that, with probability at least $1 - 7e^{-t}$,

$$\|\widehat{f} - f_0\|_2 \lesssim \sqrt{\frac{t + \log^{1/\beta}(n)}{n}} \quad \text{and} \quad \tau\|\widehat{g} - g_0\|_2 \lesssim \sqrt{\frac{t + \log^{1/\beta}(n)}{n}}.$$

Corollary D.2.7 implies that

$$\|\widehat{f} - f_0\|_2 = \mathscr{O}_{\mathbb{P}}\left(\frac{\log^{1/(2\beta)}(n)}{\sqrt{n}}\right) \quad \text{and} \quad \tau\|\widehat{g} - g_0\|_2 = \mathscr{O}_{\mathbb{P}}\left(\frac{\log^{1/(2\beta)}(n)}{\sqrt{n}}\right),$$

and the rates are minimax optimal by Zhang, Duchi and Wainwright (2013) when $\beta = 2$. We next derive the pointwise asymptotic normality and verify the validity of the confidence intervals $\mathscr{I}_{\alpha}^{\flat}(x_0)$.

Corollary D.2.8 (Pointwise asymptotic normality under exponential decay kernels)**.** Assume Conditions 4.2.1–4.3.3 and D.2.1 hold. Moreover, we assume that $0 < r_e \leq 1/2$, and choose $\lambda_q \asymp \lambda_e \asymp \log^{1/\beta}(n)/n$. If $\rho^2_{\lambda_e}(x_0) \to \rho^2(x_0)$ for some $\rho^2(x_0) > 0$, the two-step ES-KRR estimator $\widehat{g}$ satisfies

$$\tau\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}(\widehat{g} - g_0)(x_0) \xrightarrow{\mathrm{d}} \mathscr{N}(0, \rho^2(x_0)).$$

Corollary D.2.9 (Validity of bootstrap under exponential decay kernels)**.** Assume that the same conditions as in Corollary D.2.8 hold. If $x_0 \in \mathscr{X}$ satisfies $\rho^2_{\lambda}(x_0) > C$ for any sufficiently small $\lambda > 0$ with a constant $C > 0$, then we have $|\mathbb{P}\{g_0(x_0) \in \mathscr{I}_{\alpha}^{\flat}(x_0)\} - (1-\alpha)| = o(1)$ for any $\alpha \in (0,1)$.

To illustrate a specific example, assume that $\mathscr{X} = [-\pi, \pi]$ and $\mathbb{P}_X$ is the uniform distribution on $\mathscr{X}$. For any $\theta > 0$, the periodic Gaussian reproducing kernel on $[-\pi, \pi]$, introduced in Smola, Schölkopf and Müller (1998), is defined as follows:

$$K(x, x') = \frac{1}{\pi}\sum_{j=1}^{\infty} \exp(-j^2\theta^2/2)\cos(j(x - x')),$$

along with the corresponding RKHS space H_{θ}^{∞} given by

$$H_{\theta}^{\infty} := \left\{h : [-\pi, \pi] \to \mathbb{R} : \sum_{j=0}^{\infty} \frac{\theta^{2j}}{j!2^j}\int_{-\pi}^{\pi}\{h^{(j)}(x)\}^2 \mathrm{d}x < \infty\right\},$$

referred to as the infinite-order periodic Sobolev space H_θ^∞ (Lin and Brown, 2004). The infinite-order periodic Sobolev space also has uniformly bounded eigenfunctions

$$\phi_j(x) = \begin{cases} \frac{1}{\sqrt{2\pi}}, & j = 0, \\ \frac{1}{\sqrt{\pi}} \cos(j\pi x/2), & j = 2k \text{ for } k = 1,2,\ldots, \\ \frac{1}{\sqrt{\pi}} \sin((j+1)\pi x/2), & j = 2k-1 \text{ for } k = 1,2,\ldots, \end{cases} \tag{D.4}$$

and the associated eigenvalues are

$$\mu_j = \begin{cases} 1, & j = 0, \\ \exp(-j^2\theta^2/8), & j = 2k \text{ for } k = 1,2,\ldots, \\ \exp(-(j+1)^2\theta^2/8), & j = 2k-1 \text{ for } k = 1,2,\ldots. \end{cases} \tag{D.5}$$

Furthermore, the following lemma demonstrates an explicit lower bound for $\rho_\lambda^2(x_0)$ for any sufficiently small λ and $x_0 \in [0,1]$, implying that this kernel satisfies the condition in Corollary D.2.9.

Lemma D.2.2 (Lower bound for $\rho_\lambda^2(x_0)$ under infinite-order periodic Sobolev spaces.)**.** Assume that $\mathbb{P}_X$ is the uniform distribution over $[-\pi,\pi]$ and $\mathscr{H} = H_\theta^\infty$ with eigenfunctions (D.4) and eigenvalues (D.5). Moreover, suppose that there exists $\underline{\sigma} > 0$ satisfying $\mathbb{E}(\omega_i^2|X_i) \geq \underline{\sigma}^2$. Then, there exists a constant $c_{12} = c_{12}(\theta) > 0$ such that for any $\lambda \in (0, \min(1/e, e^{-2\theta^2}))$ and $x_0 \in [-\pi,\pi]$, we have $\rho_\lambda^2(x_0) \geq c_{12}\underline{\sigma}^2$.

D.3 Proofs for Section 4.3

This section presents the proofs of the results in Section 4.3, and also Proposition 4.2.1. The proofs for the technical lemmas involved are deferred to Section D.5. For ease of notation, we use the expression $\sum_{i=1}^n (R_i - \mathbb{E}R_i) = \sum_{i=1}^n (1-\mathbb{E})R_i$ for any sequence of random variables $R_{i\,i=1}^{\,n}$.

D.3.1 Proof of Proposition 4.2.1

Let $\mathbf{S} : \mathscr{H} \to \mathbb{R}^n$ be the sampling operator defined as $\mathbf{S}(f) = (\langle K_{X_i}, f\rangle_{\mathscr{H}})_{i=1}^n = (f(X_i))_{i=1}^n$. Its adjoint operator $\mathbf{S}^* : \mathbb{R}^n \to \mathscr{H}$ is given by $\mathbf{S}^*(\mathbf{a}) = \sum_{i=1}^n a_i K_{X_i}$ for any $\mathbf{a} = (a_1, a_2, \ldots, a_n)^{\mathrm{T}}$. To see this, note that

$$\langle \mathbf{S}(f), \mathbf{a}\rangle = \sum_{i=1}^n a_i f(X_i) = \left\langle f, \sum_{i=1}^n a_i K_{X_i} \right\rangle_{\mathscr{H}} = \langle f, \mathbf{S}^*(\mathbf{a})\rangle_{\mathscr{H}}.$$

Recall that $\mathbf{K} = (K(X_i, X_j))_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$. Then, for any $f \in \mathscr{H}$ and $\boldsymbol{a} \in \mathbb{R}^n$,

$$\mathbf{S}^*\mathbf{S}(f) = \mathbf{S}^*((f(X_i))_{i=1}^n) = \sum_{i=1}^n f(X_i) K_{X_i} = \underbrace{\sum_{i=1}^n (K_{X_i} \otimes K_{X_i})\, f}_{=n\widehat{T}} \quad \text{and}$$

$$\mathbf{S}\mathbf{S}^*(\boldsymbol{a}) = \mathbf{S}\left(\sum_{i=1}^n a_i K_{X_i}\right) = \left(\sum_{i=1}^n a_i K_{X_i}(X_j)\right)_{j=1}^n = \left(\sum_{i=1}^n a_i K(X_i, X_j)\right)_{j=1}^n = \mathbf{K}\boldsymbol{a}.$$

Therefore, we have $\mathbf{S}^*\mathbf{S}/n = \widehat{T}$ and $\mathbf{S}\mathbf{S}^*/n = \mathbf{K}/n$. Moreover,

$$\mathbf{k}_{x_0} = (K(X_1, x_0), \ldots, K(X_n, x_0))^{\mathrm{T}} = (K_{x_0}(X_1), \ldots, K_{x_0}(X_n))^{\mathrm{T}} = \mathbf{S}(K_{x_0}),$$

which implies that $\mathbf{v}_{x_0} = (v_{x_0,1}, \ldots, v_{x_0,n})^{\mathrm{T}} \in \mathbb{R}^n$ defined in (4.13) satisfies

$$\mathbf{v}_{x_0} = (\mathbf{K}/n + \lambda_e \mathbf{I}_n)^{-1}\mathbf{k}_{x_0} = (\mathbf{S}\mathbf{S}^*/n + \lambda_e \mathbf{I}_n)^{-1}\mathbf{k}_{x_0} = (\mathbf{S}\mathbf{S}^*/n + \lambda_e \mathbf{I}_n)^{-1}\mathbf{S}(K_{x_0}).$$

Denoting $\widehat{T}_{\lambda_e} = \widehat{T} + \lambda_e I$ with the identity operator I, we have

$$\big(\widehat{T}_{\lambda_e}^{-1} K_{X_i}(x_0)\big)_{i=1}^n = \big(\langle \widehat{T}_{\lambda_e}^{-1} K_{X_i}, K_{x_0}\rangle_{\mathscr{H}}\big)_{i=1}^n = \big(\langle K_{X_i}, \widehat{T}_{\lambda_e}^{-1} K_{x_0}\rangle_{\mathscr{H}}\big)_{i=1}^n = \mathbf{S}(\widehat{T}_{\lambda_e}^{-1} K_{x_0}),$$

where the second equality is due to the fact that $\widehat{T}_{\lambda_e}$ is an invertible self-adjoint operator.

Let $U_i = W_i - 1$ be the centered random weights, and define the n-vector

$$\mathbf{u} = \big(U_1\{Z_1(\widehat{f})/\tau - \widehat{g}(X_1)\}, U_2\{Z_2(\widehat{f})/\tau - \widehat{g}(X_2)\}, \ldots, U_n\{Z_n(\widehat{f})/\tau - \widehat{g}(X_n)\}\big)^{\mathrm{T}}.$$

The above calculations imply

$$\frac{1}{n}\sum_{i=1}^{n} U_i\{Z_i(\widehat{f})/\tau - \widehat{g}(X_i)\}v_{x_0,i} = \frac{\mathbf{u}^{\mathrm{T}}\mathbf{v}}{n} = \frac{1}{n}\mathbf{u}^{\mathrm{T}}(\mathbf{S}\mathbf{S}^*/n + \lambda_e\mathbf{I}_n)^{-1}\mathbf{S}(K_{x_0})$$

and

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} U_i\{Z_i(\widehat{f})/\tau - \widehat{g}(X_i)\}\widehat{T}_{\lambda_e}^{-1}K_{X_i}(x_0)\\ &= \frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{S}(\widehat{T}_{\lambda_e}^{-1}K_{x_0}) = \frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{S}(\mathbf{S}^*\mathbf{S}/n + \lambda_e I)^{-1}(K_{x_0}).\end{aligned}$$

Hence, it suffices to prove that

$$(\mathbf{S}\mathbf{S}^*/n + \lambda_e\mathbf{I}_n)^{-1}\mathbf{S} = \mathbf{S}(\mathbf{S}^*\mathbf{S}/n + \lambda_e I)^{-1}. \tag{D.6}$$

Note that

$$\begin{aligned}(\mathbf{S}\mathbf{S}^*/n + \lambda_e\mathbf{I}_n)\mathbf{S}(\mathbf{S}^*\mathbf{S}/n + \lambda_e I)^{-1} &= (\mathbf{S}\mathbf{S}^*\mathbf{S}/n + \lambda_e\mathbf{S})(\mathbf{S}^*\mathbf{S}/n + \lambda_e I)^{-1}\\ &= \mathbf{S}(\mathbf{S}^*\mathbf{S}/n + \lambda_e I)(\mathbf{S}^*\mathbf{S}/n + \lambda_e I)^{-1}\\ &= \mathbf{S}.\end{aligned}$$

Multiplying $(\mathbf{S}\mathbf{S}^*/n + \lambda_e\mathbf{I})^{-1}$ on both sides of the above equation yields (D.6), thereby completing the proof. □

D.3.2 Proof of Theorem 4.3.1

To prove Theorem 4.3.1, we first introduce two technical lemmas as building blocks. The first lemma below establishes a lower bound for the expected excess risk using the check loss.

Lemma D.3.1. Under Condition 4.3.1 on the conditional density function $p_{\varepsilon_i|X_i}$, the population excess risk satisfies the lower bound

$$\mathbb{E}\big\{\rho_\tau(Y_i - f(X_i)) - \rho_\tau(Y_i - f_0(X_i))\big\} \geq \frac{c_{13}}{M}\|f - f_0\|_2^2$$

for all $f : \mathscr{X} \to \mathbb{R}$ satisfying $\|f - f_0\|_\infty \leq M$ with $M \geq 1$, where $c_{13} = \min\{\underline{p}/2, \underline{p}l_0/4\}$.

Without loss of generality, throughout the following we assume $c_{13} \leq 1$ for ease of presentation; otherwise, it suffices to define $c_{13} = \min\{1, \underline{p}/2, \underline{p}l_0/4\}$.

Next, we characterize the concentration properties of the empirical excess risk around the population excess risk uniformly in a local neighborhood of f_0.

Lemma D.3.2. Assume Condition 4.2.1 holds. There exists a universal constant $c_{14} > 0$ such that for any $t, \lambda > 0$ and $\delta_2, \delta_{\mathscr{H}} > 0$, the bound

$$\sup_{\substack{f \in \mathscr{H}: \|f - f_0\|_2 \leq \delta_2, \\ \|f - f_0\|_{\mathscr{H}} \leq \delta_{\mathscr{H}}}} \left| \frac{1}{n}\sum_{i=1}^n (1 - \mathbb{E})\big\{\rho_\tau(Y_i - f(X_i)) - \rho_\tau(Y_i - f_0(X_i))\big\} \right| \leq c_{14}\left(\delta\sqrt{\frac{\mathfrak{D}_\lambda}{n}} + \delta_2\sqrt{\frac{t}{n}} + \delta_{\mathscr{H}}\frac{t}{n}\right) \tag{D.7}$$

holds with probability at least $1 - e^{-t}$, where $\delta^2 = \delta_2^2 + \lambda\delta_{\mathscr{H}}^2$.

Lemma D.3.2 presents a localized version of Lemma 1 in Lian (2022), the latter addressing a ratio-type suprema of empirical processes. For ratio-type empirical processes, the applicability of the standard contraction inequality may raise concerns. Instead, we direct our attention to the original empirical process, restricting the function f to a local neighborhood of f_0. With the above preparations, we are ready to prove Theorem 4.3.1.

Proof of Theorem 4.3.1. For any $t > 0$ and $\delta_2, \delta_{\mathscr{H}} > 0$, let $\mathscr{E}(\delta_2, \delta_{\mathscr{H}})$ be the event on which (D.7) holds with $\delta^2 = \delta_2^2 + \lambda_q \delta_{\mathscr{H}}^2$, that is, $\mathscr{E}(\delta_2, \delta_{\mathscr{H}}) = \{\text{(D.7) holds}\}$. By Lemma D.3.2, $\mathbb{P}\{\mathscr{E}(\delta_2, \delta_{\mathscr{H}})\} \geq 1 - e^{-t}$. For some $C_1 \geq 1$ to be determined, assume that the KRR estimator $\widehat{f}$ defined in (4.6) satisfies

$$\sqrt{\frac{c_{13}}{8C_1}}\|\widehat{f} - f_0\|_2 + \frac{1}{2}\lambda_q^{1/2}\|\widehat{f} - f_0\|_{\mathscr{H}} > C_1\delta_n := C_1\left(\lambda_q^{r_q+1/2}\|f^*\|_{\mathscr{H}} + \sqrt{\frac{\mathfrak{D}_{\lambda_q} + t}{n}}\right). \quad \text{(D.8)}$$

By the optimality of $\widehat{f}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - \widehat{f}(X_i)) + \lambda_q\|\widehat{f}\|_{\mathscr{H}}^2 \leq \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - f_0(X_i)) + \lambda_q\|f_0\|_{\mathscr{H}}^2.$$

Since both $\rho_\tau(\cdot)$ and $\|\cdot\|_{\mathscr{H}}$ are convex, there exists some $\widetilde{f} = \nu\widehat{f} + (1-\nu)f_0$ with $\nu \in (0,1)$ such that

$$\sqrt{\frac{c_{13}}{8C_1}}\|\widetilde{f} - f_0\|_2 + \frac{1}{2}\lambda_q^{1/2}\|\widetilde{f} - f_0\|_{\mathscr{H}} = C_1\delta_n \quad \text{(D.9)}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - \widetilde{f}(X_i)) + \lambda_q\|\widetilde{f}\|_{\mathscr{H}}^2 \leq \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - f_0(X_i)) + \lambda_q\|f_0\|_{\mathscr{H}}^2.$$

It follows that

$$\begin{aligned}
&\mathbb{E}\big\{\rho_\tau(Y_i - \widetilde{f}(X_i)) - \rho_\tau(Y_i - f_0(X_i))\big\} \\
&\leq \lambda_q\big(\|f_0\|_{\mathscr{H}}^2 - \|\widetilde{f}\|_{\mathscr{H}}^2\big) + \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\big\{\rho_\tau(Y_i - f_0(X_i)) - \rho_\tau(Y_i - \widetilde{f}(X_i))\big\} \\
&\leq \lambda_q\big(\|f_0\|_{\mathscr{H}}^2 - \|\widetilde{f}\|_{\mathscr{H}}^2\big) + \sup_{\substack{\|f-f_0\|_2 \leq \delta_2 \\ \|f-f_0\|_{\mathscr{H}} \leq \delta_{\mathscr{H}}}}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\big\{\rho_\tau(Y_i - f_0(X_i)) - \rho_\tau(Y_i - f(X_i))\big\}\right|,
\end{aligned} \quad \text{(D.10)}$$

where $\delta_2 = \sqrt{8/c_{13}} C_1^{3/2} \delta_n$ and $\delta_{\mathscr{H}} = 2C_1 \lambda_q^{-1/2} \delta_n$. Since $C_1/c_{13} \geq 1$, $\delta_2^2 + \lambda_q \delta_{\mathscr{H}}^2 \leq 12 \frac{C_1^3}{c_{13}} \delta_n^2$. Provided that $\lambda_q \geq t/n$, conditioned on $\mathscr{E}(\delta_2, \delta_{\mathscr{H}})$ we have

$$\begin{aligned}
&\sup_{\substack{\|f-f_0\|_2 \leq \delta_2 \\ \|f-f_0\|_{\mathscr{H}} \leq \delta_{\mathscr{H}}}} \left| \frac{1}{n} \sum_{i=1}^n (1-\mathbb{E}) \big\{ \rho_\tau(Y_i - f_0(X_i)) - \rho_\tau(Y_i - f(X_i)) \big\} \right| \\
&\leq c_{14} \left(\sqrt{\frac{12}{c_{13}}} C_1^{3/2} \delta_n \sqrt{\frac{\mathfrak{D}_{\lambda_q}}{n}} + \sqrt{\frac{8}{c_{13}}} C_1^{3/2} \delta_n \sqrt{\frac{t}{n}} + 2C_1 \lambda_q^{-1/2} \delta_n \frac{t}{n} \right) \\
&\leq c_{14} C_1 \delta_n \left(\sqrt{\frac{12 C_1}{c_{13}}} \sqrt{\frac{\mathfrak{D}_{\lambda_q}}{n}} + \sqrt{\frac{8 C_1}{c_{13}}} \sqrt{\frac{t}{n}} + 2\sqrt{\frac{t}{n}} \right) \\
&\leq 5 c_{14} C_1 \sqrt{\frac{C_1}{c_{13}}} \delta_n \sqrt{\frac{\mathfrak{D}_{\lambda_q} + t}{n}} \leq 5 c_{14} C_1 \sqrt{\frac{C_1}{c_{13}}} \delta_n^2, \qquad \text{(D.11)}
\end{aligned}$$

where the third inequality uses the fact that $C_1/c_{13} \geq 1$. For the term $\lambda_q(\|f_0\|_{\mathscr{H}}^2 - \|\widetilde{f}\|_{\mathscr{H}}^2)$ on the right-hand side of (D.10), note that

$$\begin{aligned}
\lambda_q \big(\|f_0\|_{\mathscr{H}}^2 - \|\widetilde{f}\|_{\mathscr{H}}^2 \big) &= -2\lambda_q \langle f_0, \widetilde{f} - f_0 \rangle_{\mathscr{H}} - \lambda_q \|\widetilde{f} - f_0\|_{\mathscr{H}}^2 \\
&\leq 2\lambda_q |\langle f_0, \widetilde{f} - f_0 \rangle_{\mathscr{H}}| - \lambda_q \|\widetilde{f} - f_0\|_{\mathscr{H}}^2 \\
&= 2\lambda_q |\langle T_K^{r_q} f^*, \widetilde{f} - f_0 \rangle_{\mathscr{H}}| - \lambda_q \|\widetilde{f} - f_0\|_{\mathscr{H}}^2 \\
&\overset{\text{(i)}}{=} 2\lambda_q |\langle f^*, T_K^{r_q} (\widetilde{f} - f_0) \rangle_{\mathscr{H}}| - \lambda_q \|\widetilde{f} - f_0\|_{\mathscr{H}}^2 \\
&\overset{\text{(ii)}}{\leq} 2\lambda_q^{r_q + 1/2} \|f^*\|_{\mathscr{H}} \cdot \|\lambda_q^{-r_q+1/2} T_K^{r_q} (\widetilde{f} - f_0)\|_{\mathscr{H}} - \lambda_q \|\widetilde{f} - f_0\|_{\mathscr{H}}^2, \qquad \text{(D.12)}
\end{aligned}$$

where step (i) follows from the self-adjoint property of $T_K^{r_q}$ and step (ii) is based on Cauchy-Schwarz inequality. To further bound the first term on the right-hand side of (D.12), we claim that for any $h \in \mathscr{H}$ and $0 \leq r_q \leq 1/2$,

$$\langle \lambda_q^{1-2r_q} T_K^{2r_q} h, h \rangle_{\mathscr{H}} \leq \langle (\lambda_q I + T_K) h, h \rangle_{\mathscr{H}}. \qquad \text{(D.13)}$$

To see this, write $h(\cdot) = \sum_{j=1}^{\infty} h_j \phi_j(\cdot)$, and note that

$$\begin{aligned}
\langle \lambda_q^{1-2r_q} T_K^{2r_q} h, h \rangle_{\mathscr{H}} &= \sum_{j=1}^{\infty} \mu_j^{2r_q} \lambda_q^{1-2r_q} \frac{h_j^2}{\mu_j} = \sum_{j=1}^{\infty} \left(\lambda_q \frac{h_j^2}{\mu_j} \right)^{1-2r_q} \cdot (h_j^2)^{2r_q} \\
&\leq (1-2r_q)\lambda_q \sum_{j=1}^{\infty} \frac{h_j^2}{\mu_j} + 2r_q \sum_{j=1}^{\infty} h_j^2 \quad \left(\text{by Young's inequality} \right) \\
&\leq \lambda_q \sum_{j=1}^{\infty} \frac{h_j^2}{\mu_j} + \sum_{j=1}^{\infty} h_j^2 \\
&= \lambda_q \|h\|_{\mathscr{H}}^2 + \langle T_K h, h \rangle_{\mathscr{H}}.
\end{aligned}$$

This verifies (D.13), from which it follows by taking $h = \widetilde{f} - f_0$ that

$$\begin{aligned}
\|\lambda_q^{-r_q+1/2} T_K^{r_q} (\widetilde{f} - f_0)\|_{\mathscr{H}} &= \sqrt{\left\langle \lambda_q^{-r_q+1/2} T_K^{r_q} (\widetilde{f} - f_0), \lambda_q^{-r_q+1/2} T_K^{r_q} (\widetilde{f} - f_0) \right\rangle_{\mathscr{H}}} \\
&\leq \sqrt{\left\langle \widetilde{f} - f_0, (\lambda_q I + T_K)(\widetilde{f} - f_0) \right\rangle_{\mathscr{H}}} \\
&\leq \lambda_q^{1/2} \|\widetilde{f} - f_0\|_{\mathscr{H}} + \|\widetilde{f} - f_0\|_2.
\end{aligned}$$

Combining this bound with (D.12) yields

$$\begin{aligned}
&\lambda_q \big(\|f_0\|_{\mathscr{H}}^2 - \|\widetilde{f}\|_{\mathscr{H}}^2 \big) \\
&\leq \underbrace{2\lambda_q^{r_q+1} \|f^*\|_{\mathscr{H}} \|\widetilde{f} - f_0\|_{\mathscr{H}}}_{= 2\lambda_q^{r_q+1/2} \|f^*\|_{\mathscr{H}} \cdot \lambda_q^{1/2} \|\widetilde{f} - f_0\|_{\mathscr{H}}} + 2\lambda_q^{r_q+1/2} \|f^*\|_{\mathscr{H}} \|\widetilde{f} - f_0\|_2 - \lambda_q \|\widetilde{f} - f_0\|_{\mathscr{H}}^2 \\
&\leq 2\lambda_q^{1+2r_q} \|f^*\|_{\mathscr{H}}^2 + \frac{\lambda_q}{2} \|\widetilde{f} - f_0\|_{\mathscr{H}}^2 + 2\lambda_q^{r_q+1/2} \|f^*\|_{\mathscr{H}} \|\widetilde{f} - f_0\|_2 - \lambda_q \|\widetilde{f} - f_0\|_{\mathscr{H}}^2 \\
&= 2\lambda_q^{1+2r_q} \|f^*\|_{\mathscr{H}}^2 + 2\lambda_q^{r_q+1/2} \|f^*\|_{\mathscr{H}} \|\widetilde{f} - f_0\|_2 - \frac{\lambda_q}{2} \|\widetilde{f} - f_0\|_{\mathscr{H}}^2 \\
&\leq 2\delta_n^2 + 2C_1 \sqrt{\frac{8C_1}{c_{13}}} \delta_n^2 - \frac{\lambda_q}{2} \|\widetilde{f} - f_0\|_{\mathscr{H}}^2, \qquad \text{(D.14)}
\end{aligned}$$

where the last inequality follows from the definition of δ_n in (D.8) and (D.9).

Turning to the left-hand side of (D.10), again using (D.8) and (D.9) yields

$$\|\widetilde{f}-f_0\|_\infty \le \|\widetilde{f}-f_0\|_{\mathscr{H}} \le 2C_1\lambda_q^{-1/2}\delta_n = 2C_1\left(\lambda_q^{r_q}\|f^*\|_{\mathscr{H}} + \lambda_q^{-1/2}\sqrt{\frac{\mathfrak{D}_{\lambda_q}+t}{n}}\right) \le 4C_1,$$

where the last inequality is due to the conditions imposed on λ_q, that is, $\lambda_q \ge (\mathfrak{D}_{\lambda_q}+t)/n$ and $\lambda_q^{r_q}\|f^*\|_{\mathscr{H}} \le 1$. Combining (D.10), (D.11) and (D.14), and taking $M = 4C_1$ in Lemma D.3.1, we obtain that conditioned on $\mathscr{E}(\delta_1,\delta_2)$,

$$\frac{c_{13}}{4C_1}\|\widetilde{f}-f_0\|_2^2 \le 5c_{14}C_1\sqrt{\frac{C_1}{c_{13}}}\delta_n^2 + 2\delta_n^2 + 2C_1\sqrt{\frac{8C_1}{c_{13}}}\delta_n^2 - \frac{\lambda_q}{2}\|\widetilde{f}-f_0\|_{\mathscr{H}}^2,$$

which further implies

$$\begin{aligned}\left(\sqrt{\frac{c_{13}}{8C_1}}\|\widetilde{f}-f_0\|_2 + \frac{1}{2}\lambda_q^{1/2}\|\widetilde{f}-f_0\|_{\mathscr{H}}\right)^2 &\le \frac{c_{13}}{4C_1}\|\widetilde{f}-f_0\|_2^2 + \frac{\lambda_q}{2}\|\widetilde{f}-f_0\|_{\mathscr{H}}^2 \\ &\le \left(5c_{14}C_1\sqrt{\frac{C_1}{c_{13}}} + 2 + 2C_1\sqrt{\frac{8C_1}{c_{13}}}\right)\delta_n^2.\end{aligned}$$

Based on the above inequality, we can choose a sufficiently large $C_1 > 1$ such that $5c_{14}C_1\sqrt{C_1/c_{13}} + 2 + 2C_1\sqrt{8C_1/c_{13}} < C_1^2$. However, the construction of $\widetilde{f}$, assuming $\widehat{f}$ satisfies (D.8), ensures that

$$\left(\sqrt{\frac{c_{13}}{8C_1}}\|\widetilde{f}-f_0\|_2 + \frac{1}{2}\lambda_q^{1/2}\|\widetilde{f}-f_0\|_{\mathscr{H}}\right)^2 = C_1^2\delta_n^2.$$

This leads to a contradiction conditioned on $\mathscr{E}(\delta_2,\delta_{\mathscr{H}})$. Therefore, we must have

$$\sqrt{\frac{c_{13}}{8C_1}}\|\widehat{f}-f_0\|_2 + \frac{1}{2}\|\widehat{f}-f_0\|_{\mathscr{H}} \le C_1\delta_n$$

on the event $\mathscr{E}(\delta_2,\delta_{\mathscr{H}})$ that occurs with probability at least $1-e^{-t}$. This implies the claimed bounds with $c_1 = C_1^{3/2}\sqrt{8/c_{13}}$ and $c_2 = 2C_1$. □

D.3.3 Proof of Theorem 4.3.2

To begin with, we introduce the following notations that will be used frequently. Recall the QR residuals defined as $\varepsilon_i = Y_i - f_0(X_i)$ for $1 \le i \le n$ and $\varepsilon = Y - f_0(X)$. For any function $f : \mathscr{X} \to \mathbb{R}$, define

$$Z(f)(X, \varepsilon) = \{Y - f(X)\}\mathbb{1}\{Y \le f(X)\} + \tau f(X), \tag{D.15}$$

and $Z_i(f) = Z(f)(X_i, \varepsilon_i)$ for $1 \le i \le n$. Furthermore, define the zero-mean ES residuals

$$\omega_i = Z(f_0)(X_i, \varepsilon_i) - \tau g_0(X_i),$$

which can be equivalently expressed as

$$\omega_i = \varepsilon_i \mathbb{1}(\varepsilon_i \le 0) + \tau f_0(X_i) - \tau g_0(X_i) = \varepsilon_{i,-} - \mathbb{E}(\varepsilon_{i,-}|X_i),$$

where $\varepsilon_{i,-} = \varepsilon_i \mathbb{1}(\varepsilon_i \le 0)$. Condition 4.3.2 implies that $\log \mathbb{E}(e^{t\omega_i}|X_i) \le \sigma_0^2 t^2/2$ (almost surely) for any $t \in \mathbb{R}$.

Similar to the proof of Theorem 4.3.1, we also need to establish concentration bounds for the empirical processes that will arise in the proof. For any function $h : \mathscr{X} \to \mathbb{R}$, define the empirical L_2-norm $\|h\|_n = \sqrt{(1/n)\sum_{i=1}^n h^2(X_i)}$. The following lemma characterizes the relationship between the empirical and population L_2-norms for functions in $\mathscr{H}$.

Lemma D.3.3. Assume Condition 4.2.1 holds and let $\lambda_e \ge 9/n$. Then, the following event

$$\mathscr{S} = \mathscr{S}(\lambda_e) := \left\{ \frac{1}{2} \le \frac{\|h\|_n^2 + \lambda_e \|h\|_{\mathscr{H}}^2}{\|h\|_2^2 + \lambda_e \|h\|_{\mathscr{H}}^2} \le \frac{3}{2} \text{ for all } h \in \mathscr{H} \right\}$$

holds with probability at least $1 - 14\mathfrak{D}_{\lambda_e} e^{-3n\lambda_e/32}$. Here we use the convention $0/0 = 1$.

Next, Lemma D.3.4 and Lemma D.3.5 provide high probability bounds for a multiplier

empirical process and a product empirical process, respectively. In the majority of RKHS literature, the zero-mean error terms ω_i in the multiplier process $\{(1/n)\sum \omega_i h(X_i) : h \in \mathscr{H}\}$ are typically assumed to be bounded. However, this constraint is relaxed, as demonstrated in Lemma D.3.4. In comparison to the analysis for joint linear models, the primary technical challenge lies in controlling the multiplier process in Lemma D.3.5. Specifically, the bound (D.16), a nontrivial extension of Lemma G.1 in He et al. (2023), relies on a Gaussian comparison inequality for product empirical processes, an upper bound (under expectation) on the operator norm of sums of bounded random matrices, along with a series of intricate probabilistic inequalities. Lemma D.3.5 can be more broadly applied to the analysis of KRR estimators with nonparametrically generated response variables, and therefore is of independent interest.

Lemma D.3.4. Assume Conditions 4.2.1 and 4.3.2 hold. There exists an absolute constant $c_{15} > 0$ such that for any $t > 0$, the following event

$$\mathscr{M}_t := \left\{ \frac{1}{n}\sum_{i=1}^{n} \omega_i h(X_i) \le \frac{1}{4}\|h\|_n^2 + \frac{\lambda_e}{4}\|h\|_{\mathscr{H}}^2 + c_{15}\sigma_0^2 \frac{\mathfrak{D}_{\lambda_e} + t}{n}, \ \ \forall h \in \mathscr{H} \right\}$$

satisfies $\mathbb{P}(\mathscr{M}_t^{\mathrm{c}} \cap \mathscr{S}) \le 4e^{-t}$, where the event $\mathscr{S}$ is defined in Lemma D.3.3.

Lemma D.3.5. Assume Conditions 4.2.1, 4.3.2 and 4.3.3 hold. There exists a universal constant $c_{16} > 0$ such that for any $0 < t \lesssim \lambda_e n$ and $n \ge C_\phi^2 \mathfrak{D}_{\lambda_e} \log n$, with probability at least $1 - e^{-t}$, the bound

$$\begin{aligned} \frac{\tau}{n}\sum_{i=1}^{n}\{Z_i(f) - Z_i(f_0)\}\{g(X_i) - g_0(X_i)\} \le{}& \frac{\tau^2}{16}\|g - g_0\|_2^2 + \frac{\lambda_e \tau^2}{16}\|g - g_0\|_{\mathscr{H}}^2 \\ &+ \frac{\bar{p}\tau}{2}\|f - f_0\|_4^2\|g - g_0\|_2 + c_{16}C_\phi^2\delta_2^2 \frac{\mathfrak{D}_{\lambda_q}(\mathfrak{D}_{\lambda_e} + t)}{n} \end{aligned} \tag{D.16}$$

holds uniformly over $f, g \in \mathscr{H}$ satisfying $\|f - f_0\|_2^2 + \lambda_q\|f - f_0\|_{\mathscr{H}}^2 \le \delta_2^2$.

With the above preparations, we are ready to prove Theorem 4.3.2.

Proof of Theorem 4.3.2. Note that $\widehat{g}$ can be equivalently defined as the minimizer of $g \mapsto (1/n)\sum_{i=1}^{n}\{Z_i(\widehat{f}) - \tau g(X_i)\}^2 + \lambda_e \|\tau g\|_{\mathscr{H}}^2$. We consider the change of variables

$$h = \tau g, \quad \widehat{h} = \tau\widehat{g} \quad \text{and} \quad h_0 = \tau g_0,$$

so that $\omega_i = Z_i(f_0) - h_0(X_i)$. By the optimality of $\widehat{g}$,

$$\frac{1}{n}\sum_{i=1}^{n}\{Z_i(\widehat{f}) - \widehat{h}(X_i)\}^2 + \lambda_e\|\widehat{h}\|_{\mathscr{H}}^2 \le \frac{1}{n}\sum_{i=1}^{n}\{Z_i(\widehat{f}) - h_0(X_i)\}^2 + \lambda_e\|h_0\|_{\mathscr{H}}^2.$$

After a simple algebra, this further implies

$$\begin{aligned}\|\widehat{h} - h_0\|_n^2 \le \frac{2}{n}\sum_{i=1}^{n}\omega_i(\widehat{h} - h_0)(X_i) + \frac{2}{n}\sum_{i=1}^{n}\{Z_i(\widehat{f}) - Z_i(f_0)\}\{\widehat{h}(X_i) - h_0(X_i)\} \\ + \lambda_e\big(\|h_0\|_{\mathscr{H}}^2 - \|\widehat{h}\|_{\mathscr{H}}^2\big). \qquad \text{(D.17)}\end{aligned}$$

For any $t > 0$, let $\mathscr{S}$ and $\mathscr{M}_t$ be the events defined in Lemma D.3.3 and Lemma D.3.4, respectively. In addition, let $\mathscr{P}_t$ be the event on which the bound (D.16) holds uniformly for $f, g \in \mathscr{H}$ with $\|f - f_0\|_2^2 + \lambda_q\|f - f_0\|_{\mathscr{H}}^2 \le \delta_2^2$. Provided that $\lambda_e \ge \frac{32}{3}\frac{t+\log(14\mathfrak{D}_{\lambda_e})}{n}$, we have

$$14\mathfrak{D}_{\lambda_e}\exp(-3n\lambda_e/32) = \exp\{\log(14\mathfrak{D}_{\lambda_e}) - 3n\lambda_e/32\} \le e^{-t}.$$

Together, Lemma D.3.3, Lemma D.3.4 and Lemma D.3.5 imply

$$\mathbb{P}(\mathscr{S}^{\mathrm{c}} \cup \mathscr{M}_t^{\mathrm{c}} \cup \mathscr{P}_t^{\mathrm{c}}) \le \mathbb{P}(\mathscr{S}^{\mathrm{c}} \cup \mathscr{M}_t^{\mathrm{c}}) + \mathbb{P}(\mathscr{P}_t^{\mathrm{c}}) = \mathbb{P}(\mathscr{S}^{\mathrm{c}}) + \mathbb{P}(\mathscr{S} \cap \mathscr{M}_t^{\mathrm{c}}) + \mathbb{P}(\mathscr{P}_t^{\mathrm{c}}) \le 6e^{-t}.$$

By the definition of these events and the notations in Theorem 4.3.2, it follows from (D.17) that

conditioned on $\mathscr{M}_t \cap \mathscr{P}_t$,

$$\begin{aligned}\|\widehat{h}-h_0\|_n^2 \le \frac{1}{2}\|\widehat{h}-h_0\|_n^2 + \frac{1}{8}\|\widehat{h}-h_0\|_2^2 + \frac{5\lambda_e}{8}\|\widehat{h}-h_0\|_{\mathscr{H}}^2 \\ + \bar{p}\|\widehat{h}-h_0\|_2\|\widehat{f}-f_0\|_4^2 + \lambda_e\big(\|h_0\|_{\mathscr{H}}^2 - \|\widehat{h}\|_{\mathscr{H}}^2\big) + C_1(\gamma_{\mathrm{s}}^2+\delta_{\mathrm{s}}^2),\end{aligned} \tag{D.18}$$

where $C_1 > 0$ is a universal constant. Recall that $g_0 = T_k^{r_e} g^*$ with $0 \le r_e \le 1/2$ and $g^* \in \mathscr{H}$. Following a similar argument that leads to (D.14), we have

$$\begin{aligned}&\lambda_e\big(\|g_0\|_{\mathscr{H}}^2 - \|\widehat{g}\|_{\mathscr{H}}^2\big) \\ &\le 2\lambda_e^{r_e+1}\|g^*\|_{\mathscr{H}}\|\widehat{g}-g_0\|_{\mathscr{H}} + 2\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}\|\widehat{g}-g_0\|_2 - \lambda_e\|\widehat{g}-g_0\|_{\mathscr{H}}^2 \\ &\le 16\lambda_e^{2r_e+1}\|g^*\|_{\mathscr{H}}^2 + \frac{\lambda_e}{16}\|\widehat{g}-g_0\|_{\mathscr{H}}^2 + 2\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}\|\widehat{g}-g_0\|_2 - \lambda_e\|\widehat{g}-g_0\|_{\mathscr{H}}^2.\end{aligned} \tag{D.19}$$

Moreover, conditioned on $\mathscr{S}$ it holds

$$\frac{1}{4}\|\widehat{g}-g_0\|_2^2 \le \frac{1}{2}\|\widehat{g}-g_0\|_n^2 + \frac{\lambda_e}{4}\|\widehat{g}-g_0\|_{\mathscr{H}}^2.$$

Combining this with (D.18) and (D.19), we obtain that on the event $\mathscr{S} \cap \mathscr{M}_t \cap \mathscr{P}_t$,

$$\begin{aligned}\frac{1}{4}\|\widehat{h}-h_0\|_2^2 &\le \frac{1}{2}\|\widehat{h}-h_0\|_n^2 + \frac{\lambda_e}{4}\|\widehat{h}-h_0\|_{\mathscr{H}}^2 \\ &\le \frac{1}{8}\|\widehat{h}-h_0\|_2^2 + \|\widehat{h}-h_0\|_2\big(\bar{p}\|\widehat{f}-f_0\|_4^2 + 2\lambda_e^{r_e+1/2}\|h^*\|_{\mathscr{H}}\big) \\ &\quad - \frac{\lambda_e}{16}\|\widehat{h}-h_0\|_{\mathscr{H}}^2 + 16\lambda_e^{2r_e+1}\|h^*\|_{\mathscr{H}}^2 + C_1(\gamma_{\mathrm{s}}^2+\delta_{\mathrm{s}}^2),\end{aligned}$$

which further implies

$$\begin{aligned}&2\|\widehat{h}-h_0\|_2^2 \le 2\|\widehat{h}-h_0\|_2^2 + \lambda_e\|\widehat{h}-h_0\|_{\mathscr{H}}^2 \\ &\le \|\widehat{h}-h_0\|_2\big(16\bar{p}\|\widehat{f}-f_0\|_4^2 + 32\lambda_e^{r_e+1/2}\|h^*\|_{\mathscr{H}}\big) + 256\lambda_e^{2r_e+1}\|h^*\|_{\mathscr{H}}^2 + 16C_1(\gamma_{\mathrm{s}}^2+\delta_{\mathrm{s}}^2),\end{aligned} \tag{D.20}$$

where $h^* = \tau g^*$. By reorganizing the terms, it follows that

$$\begin{aligned}\|\widehat{h} - h_0\|_2 &\leq 8\bar{p}\|\widehat{f} - f_0\|_4^2 + 16(1+1/\sqrt{2})\lambda_e^{r_e+1/2}\|h^*\|_{\mathscr{H}} + 2\sqrt{2}C_1\sqrt{\gamma_{\mathrm{s}}^2 + \delta_{\mathrm{s}}^2} \\ &\leq c_3\big(\lambda_e^{r_e+1/2}\|h^*\|_{\mathscr{H}} + \bar{p}\|\widehat{f} - f_0\|_4^2 + \gamma_{\mathrm{s}} + \delta_{\mathrm{s}}\big)\end{aligned}$$

for some absolute constant $c_3 > 0$. Finally, combining this L_2-error bound with (D.20) results in the claimed bound on $\|\widehat{h} - h_0\|_{\mathscr{H}}$, thereby completing the proof. □

D.3.4 Proof of Theorem 4.3.3

Under Condition 4.3.3, we can define an equivalent kernel (Silverman, 1984) and use its norm to establish a tighter upper bound for the L_∞-norm of functions in $\mathscr{H}$; see Lemma D.3.6 and the subsequent discussion for details.

For any $\lambda > 0$ fixed, define a new inner product $\langle \cdot, \cdot \rangle_\lambda$ on $\mathscr{H}$ as

$$\langle f, g \rangle_\lambda = \langle f, g \rangle_2 + \lambda \langle f, g \rangle_{\mathscr{H}} \quad \text{for} \quad f, g \in \mathscr{H},$$

where $\langle f, g \rangle_2 = \langle f, g \rangle_{L_2(\mathbb{P}_X)} = \int_{\mathscr{X}} f(x)g(x)\mathrm{d}\mathbb{P}_X(x)$. Recall that $\{\phi_j\}_{j \geq 1}$ are the eigenfunctions of T_K with the associated eigenvalues $\{\mu_j\}_{j \geq 1}$. For $f = \sum_{j=1}^\infty f_j \phi_j$ and $g = \sum_{j=1}^\infty g_j \phi_j$, we have

$$\langle f, g \rangle_\lambda = \sum_{j=1}^\infty f_j g_j + \lambda \sum_{j=1}^\infty \frac{f_j g_j}{\mu_j} = \sum_{j=1}^\infty \frac{f_j g_j}{\nu_j},$$

where

$$\nu_j = \frac{1}{1 + \lambda/\mu_j} = \frac{\mu_j}{\mu_j + \lambda} \quad \text{for} \quad j \geq 1.$$

Let $\mathscr{H}_\lambda$ be the new RKHS associated with $\langle \cdot, \cdot \rangle_\lambda$. Note that $\mathscr{H}_\lambda$ is the same functional space as

$\mathscr{H}$, but with a different reproducing kernel K_λ, referred to as the equivalent kernel, defined as

$$K_\lambda(x_1,x_2) = \sum_{j=1}^{\infty} \nu_j \phi_j(x_1)\phi_j(x_2), \;\; x_1,x_2 \in \mathscr{X}.$$

In the following, we write $\|f\|_\lambda^2 = \langle f,f\rangle_\lambda$ for any $f \in \mathscr{H}_\lambda$, and let $\|\cdot\|_{\mathrm{op},\lambda}$ be the operator norm with respect to $\mathscr{H}_\lambda$, that is, for any $T : \mathscr{H} \to \mathscr{H}$,

$$\|T\|_{\mathrm{op},\lambda} = \sup_{\|f\|_\lambda \le 1} \langle Tf, f\rangle_\lambda.$$

Here λ is the regularization parameter, which is typically chosen as a small number. We remark that the above $\|\cdot\|_\lambda$ should not be confused with the L_2-norm $\|\cdot\|_2$.

We first describe a relationship between the norms $\|\cdot\|_\infty$ and $\|\cdot\|_\lambda$, the proof of which is established in (D.79) within the proof of Lemma D.3.5.

Lemma D.3.6. Under Conditions 4.2.1 and 4.3.3, the bound $\|h\|_\infty \le C_\phi \mathfrak{D}_\lambda^{1/2}\|h\|_\lambda$ holds for any $h \in \mathscr{H}$.

This lemma demonstrates that the norm $\|\cdot\|_\lambda$ can provide a tighter upper bound for the supremum norm of functions in $\mathscr{H}$ compared to using the original RKHS norm $\|\cdot\|_{\mathscr{H}}$. To illustrate, Condition 4.2.1 implies that $\|h\|_\infty \le \|h\|_{\mathscr{H}}$ for any $h \in \mathscr{H}$. However, Lemma D.3.6 shows that

$$\|h\|_\infty \le C_\phi \mathfrak{D}_\lambda^{1/2}\|h\|_\lambda = C_\phi \mathfrak{D}_\lambda^{1/2}\sqrt{\|h\|_2^2 + \lambda\|h\|_{\mathscr{H}}^2}.$$

Therefore, if $\mathfrak{D}_\lambda(\|h\|_2^2 + \lambda\|h\|_{\mathscr{H}}^2)$ is smaller in order than $\|h\|_{\mathscr{H}}^2$, using the norm $\|\cdot\|_\lambda$ results in a tighter bound for the supremum norm compared to using $\|\cdot\|_{\mathscr{H}}$.

Next, define

$$T_\lambda = T_K + \lambda I, \quad \widehat{T} = \frac{1}{n}\sum_{i=1}^{n} K_{X_i} \otimes K_{X_i} \;\; \text{and} \;\; \widehat{T}_\lambda = \widehat{T} + \lambda I. \tag{D.21}$$

Here, for $h_1, h_2 \in \mathscr{H}$, their tensor product $h_1 \otimes h_2 : \mathscr{H} \to \mathscr{H}$ is a rank-one operator satisfying $(h_1 \otimes h_2)h = \langle h_2, h \rangle_{\mathscr{H}} h_1$ for any $h \in \mathscr{H}$. Note that $\mathbb{E}(K_{X_i} \otimes K_{X_i}) = T_K$ by definition (4.3).

Lemma D.3.7. Suppose that $\|T_\lambda^{-1}(\widehat{T} - T_K)\|_{\mathrm{op},\lambda} \le \zeta < 1$ for some $\lambda > 0$. Then, there exists an operator $A : \mathscr{H} \to \mathscr{H}$ such that

$$\widehat{T}_\lambda^{-1} - T_\lambda^{-1} = A\, T_\lambda^{-1} \quad \text{with} \quad \|A\|_{\mathrm{op},\lambda} \le \frac{\zeta}{1-\zeta}.$$

Finally, we compile several concentration results necessary for proving the functional Bahadur representation.

Lemma D.3.8. Assume that Conditions 4.2.1, 4.3.2 and 4.3.3 hold. Then, there exists a universal constant $c_{17} > 0$ such that for any $t > 0$ and $\lambda \gtrsim (t + \log \mathfrak{D}_\lambda)/n$, we have

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n \omega_i T_\lambda^{-1} K_{X_i} \right\|_\lambda \ge c_{17} \sigma_0 \sqrt{\frac{\mathfrak{D}_\lambda + t}{n}} \right) \le 5e^{-t}, \tag{D.22}$$

and

$$\mathbb{P}\left\{ \left\| T_\lambda^{-1}(\widehat{T} - T_K) \right\|_{\mathrm{op},\lambda} \ge 4 \left(C_\phi^2 \mathfrak{D}_\lambda \frac{t + \log n}{n} \bigvee C_\phi \mathfrak{D}_\lambda^{1/2} \sqrt{\frac{t + \log n}{n}} \right) \right\} \le \frac{14 \mathfrak{D}_\lambda}{n} e^{-t}. \tag{D.23}$$

Moreover, if we further assume that $n \ge C_\phi^2 \mathfrak{D}_{\lambda_e} \log n$, then, with probability at least $1 - e^{-t}$, the bound

$$\left\| \frac{1}{n} \sum_{i=1}^n \{Z_i(f) - Z_i(f_0)\} T_\lambda^{-1} K_{X_i} \right\|_\lambda \le c_{18} C_\phi \delta_2 \mathfrak{D}_{\lambda_q}^{1/2} \sqrt{\frac{t + \mathfrak{D}_{\lambda_e}}{n}} + \frac{\bar{p}}{2} \|f - f_0\|_4^2 \tag{D.24}$$

holds uniformly for $f \in \mathscr{H}$ with $\|f - f_0\|_2^2 + \lambda_q \|f - f_0\|_{\mathscr{H}}^2 \le \delta_2^2$, where $c_{18} > 0$ is an absolute constant.

Now, we are ready to prove Theorem 4.3.3.

Proof of Theorem 4.3.3. To begin with, recall the population (penalized) risk minimizer g_{λ_e} given

in (4.12), which can be written as $g_{\lambda_e} = T_{\lambda_e}^{-1} T_K g_0$. We first prove the Bahadur representation of the two-step ES estimator $\widehat{g} \in \operatorname{argmin}_{g \in \mathscr{H}} \widehat{\mathscr{L}_n}(\widehat{f}, g)$. Write the empirical risk $\widehat{\mathscr{L}_n}(\widehat{f}, g)$ as

$$\begin{aligned}
\widehat{\mathscr{L}_n}(\widehat{f}, g) &= \frac{1}{n}\sum_{i=1}^{n} \{Z_i(\widehat{f})/\tau - g(X_i)\}^2 + \lambda_e \|g\|_{\mathscr{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \{Z_i(\widehat{f})/\tau - \langle g, K_{X_i}\rangle_{\mathscr{H}}\}^2 + \lambda_e \langle g, g\rangle_{\mathscr{H}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \{Z_i(\widehat{f})/\tau\}^2 + \Bigg\langle \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} K_{X_i} \otimes K_{X_i} + \lambda_e I\right)}_{=\widehat{T}_{\lambda_e}} g, g \Bigg\rangle_{\mathscr{H}} - \left\langle \frac{2}{\tau n}\sum_{i=1}^{n} Z_i(\widehat{f}) K_{X_i}, g \right\rangle_{\mathscr{H}}.
\end{aligned}$$

By taking the Fréchet derivative of $g \to \widehat{\mathscr{L}_n}(\widehat{f}, g)$, the empirical risk minimizer $\widehat{g}$ admits the closed-form expression

$$\widehat{g} = \widehat{T}_{\lambda_e}^{-1} \frac{1}{\tau n}\sum_{i=1}^{n} Z_i(\widehat{f}) K_{X_i}. \tag{D.25}$$

STEP I. DECOMPOSITION OF $\widehat{g} - g_{\lambda_e}$. From the model setup $Z_i(f_0) = \tau g_0(X_i) + \omega_i$ and the reproducing property of K_{X_i}, we derive that

$$\begin{aligned}
&\tau(\widehat{g} - g_{\lambda_e}) \\
&= \widehat{T}_{\lambda_e}^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i(\widehat{f}) K_{X_i} - \tau T_{\lambda_e}^{-1} T_K g_0 \\
&= \widehat{T}_{\lambda_e}^{-1} \frac{1}{n}\sum_{i=1}^{n} \{Z_i(\widehat{f}) - Z_i(f_0) + Z_i(f_0)\} K_{X_i} - \tau T_{\lambda_e}^{-1} T_K g_0 \\
&= \widehat{T}_{\lambda_e}^{-1} \frac{1}{n}\sum_{i=1}^{n} \{Z_i(\widehat{f}) - Z_i(f_0)\} K_{X_i} + \widehat{T}_{\lambda_e}^{-1} \frac{1}{n}\sum_{i=1}^{n} \omega_i K_{X_i} + \tau \widehat{T}_{\lambda_e}^{-1} \widehat{T} g_0 - \tau T_{\lambda_e}^{-1} T_K g_0 \\
&= T_{\lambda_e}^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \{Z_i(\widehat{f}) - Z_i(f_0)\} K_{X_i} + \frac{1}{n}\sum_{i=1}^{n} \omega_i K_{X_i}\right] + \tau \widehat{T}_{\lambda_e}^{-1} \widehat{T} g_0 - \tau T_{\lambda_e}^{-1} T_K g_0 \\
&\quad + \left(\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1}\right) \left[\frac{1}{n}\sum_{i=1}^{n} \{Z_i(\widehat{f}) - Z_i(f_0)\} K_{X_i} + \frac{1}{n}\sum_{i=1}^{n} \omega_i K_{X_i}\right].
\end{aligned} \tag{D.26}$$

Observe that

$$\begin{aligned}\widehat{T}_{\lambda_e}^{-1}\widehat{T}g_0 - T_{\lambda_e}^{-1}T_K g_0 &= (\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1})\widehat{T}g_0 + T_{\lambda_e}^{-1}(\widehat{T} - T_K)g_0 \\ &= (\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1})(\widehat{T} - T_K)g_0 + (\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1})T_K g_0 + T_{\lambda_e}^{-1}(\widehat{T} - T_K)g_0.\end{aligned} \tag{D.27}$$

Here, note that

$$\begin{aligned}&(\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1})T_K g_0 + T_{\lambda_e}^{-1}(\widehat{T} - T_K)g_0 \\ &= \widehat{T}_{\lambda_e}^{-1}(T_{\lambda_e} - \widehat{T}_{\lambda_e})T_{\lambda_e}^{-1}T_K g_0 + T_{\lambda_e}^{-1}(\widehat{T} - T_K)g_0 \\ &= -\widehat{T}_{\lambda_e}^{-1}(\widehat{T} - T_K)g_{\lambda_e} + T_{\lambda_e}^{-1}(\widehat{T} - T_K)g_0 \\ &= -(\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1})(\widehat{T} - T_K)g_{\lambda_e} - T_{\lambda_e}^{-1}(\widehat{T} - T_K)g_{\lambda_e} + T_{\lambda_e}^{-1}(\widehat{T} - T_K)g_0 \\ &= -(\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1})(\widehat{T} - T_K)g_{\lambda_e} + T_{\lambda_e}^{-1}(\widehat{T} - T_K)(g_0 - g_{\lambda_e}).\end{aligned}$$

Together, the above equality and (D.27) yield

$$\widehat{T}_{\lambda_e}^{-1}\widehat{T}g_0 - T_{\lambda_e}^{-1}T_K g_0 = (\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1})(\widehat{T} - T_K)(g_0 - g_{\lambda_e}) + T_{\lambda_e}^{-1}(\widehat{T} - T_K)(g_0 - g_{\lambda_e}),$$

which, combined with (D.26), further implies

$$\begin{aligned}&\tau(\widehat{g} - g_{\lambda_e}) - \frac{1}{n}\sum_{i=1}^{n}\omega_i T_{\lambda_e}^{-1}K_{X_i} \\ &= T_{\lambda_e}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\{Z_i(\widehat{f}) - Z_i(f_0)\}K_{X_i} + \tau(\widehat{T} - T_K)(g_0 - g_{\lambda_e})\right] \\ &\quad + (\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1})\left[\frac{1}{n}\sum_{i=1}^{n}\{Z_i(\widehat{f}) - Z_i(f_0)\}K_{X_i} + \frac{1}{n}\sum_{i=1}^{n}\omega_i K_{X_i} + \tau(\widehat{T} - T_K)(g_0 - g_{\lambda_e})\right].\end{aligned} \tag{D.28}$$

STEP II. OCCURRENCES OF "GOOD" EVENTS WITH HIGH PROBABILITY. The decomposition in (D.28) suggests that $(1/n)\sum_{i=1}^{n}\omega_i T_{\lambda_e}^{-1}K_{X_i}$ should be the leading term, while the remaining terms are of higher order. In this step, we introduce the "good" events on which the remainders

are well controlled and then establish high probability bounds for these events. To this end, let $n \geq 64C_\phi^2 \mathfrak{D}_{\lambda_e}(t+\log n)$ so that $\zeta := 4C_\phi \sqrt{\mathfrak{D}_{\lambda_e}(t+\log n)/n} \leq 1/2$. It then follows from (D.23) that the event

$$\mathscr{E}_t := \left\{ \|T_{\lambda_e}^{-1}(\widehat{T} - T_K)\|_{\mathrm{op},\lambda_e} \leq \zeta \right\}$$

occurs with probability at least $1 - e^{-t}$.

To apply (D.24), we need to determine the order of δ_2 and establish the convergence rate of $\widehat{f}$ under the L_4-norm. For the former, Theorem 4.3.1 implies that with probability at least $1 - e^{-t}$,

$$\|\widehat{f} - f_0\|_2^2 + \lambda_q \|\widehat{f} - f_0\|_{\mathscr{H}}^2 \lesssim \left(\lambda_q^{r_q+1/2} \|f^*\|_{\mathscr{H}} + \sqrt{\frac{\mathfrak{D}_{\lambda_q} + t}{n}} \right)^2.$$

By taking

$$\delta_2 \asymp \lambda_q^{r_q+1/2} \|f^*\|_{\mathscr{H}} + \sqrt{\frac{\mathfrak{D}_{\lambda_q} + t}{n}}$$

and applying Lemma D.3.6, we conclude that with probability at least $1 - e^{-t}$, $\|\widehat{f} - f_0\|_2^2 + \lambda_q \|\widehat{f} - f_0\|_{\mathscr{H}}^2 \leq \delta_2^2$,

$$\|\widehat{f} - f_0\|_\infty \leq C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \left(\|\widehat{f} - f_0\|_2^2 + \lambda_q \|\widehat{f} - f_0\|_{\mathscr{H}}^2 \right)^{1/2} \leq C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \delta_2,$$

and moreover,

$$\|\widehat{f} - f_0\|_4^2 = \left\{ \mathbb{E}_{X \sim \mathbb{P}_X} (\widehat{f} - f_0)^4(X) \right\}^{1/2} \leq \|\widehat{f} - f_0\|_\infty \|\widehat{f} - f_0\|_2 \leq C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \delta_2^2. \tag{D.29}$$

Let $\mathscr{B}_t$ be the event on which the following bounds

$$\left\| \frac{1}{n} \sum_{i=1}^n \omega_i T_{\lambda_e}^{-1} K_{X_i} \right\|_{\lambda_e} \leq \gamma_1 := c_{17} \sigma_0 \sqrt{\frac{\mathfrak{D}_{\lambda_e} + t}{n}}$$

and

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \{Z_i(\widehat{f}) - Z_i(f_0)\} T_{\lambda_e}^{-1} K_{X_i} \right\|_{\lambda_e} \leq \gamma_2 := c_{18} C_\phi \delta_2 \mathfrak{D}_{\lambda_q}^{1/2} \sqrt{\frac{t + \mathfrak{D}_{\lambda_e}}{n}} + \frac{\bar{p}}{2} C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \delta_2^2$$

hold. By Lemma D.3.8 and (D.29), the event $\mathscr{B}_t$ occurs with probability at least $1 - 7e^{-t}$. Consequently, we have $\mathbb{P}(\mathscr{E}_t \cap \mathscr{B}_t) \geq 1 - 8e^{-t}$.

STEP III. HIGH PROBABILITY BOUND FOR THE REMAINDER TERM. We combine the results in Step I and Step II to complete the proof of the Bahadur representation of $\widehat{g}$. Define

$$\mathrm{Rem}_n := \tau(\widehat{g} - g_{\lambda_e}) - \frac{1}{n} \sum_{i=1}^{n} \omega_i T_{\lambda_e}^{-1} K_{X_i},$$

which equals the right-hand side of (D.28). Conditioned on $\mathscr{E}_t \cap \mathscr{B}_t$, it is easy to see that

$$\left\| T_{\lambda_e}^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} \{Z_i(\widehat{f}) - Z_i(f_0)\} K_{X_i} + \tau(\widehat{T} - T_K)(g_0 - g_{\lambda_e}) \right] \right\|_{\lambda_e} \leq \gamma_2 + \tau\zeta \|g_0 - g_{\lambda_e}\|_{\lambda_e}.$$

On the other hand, it follows from Lemma D.3.7 that,

$$\begin{aligned}
&\left\| (\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1}) \left[\frac{1}{n} \sum_{i=1}^{n} \{Z_i(\widehat{f}) - Z_i(f_0)\} K_{X_i} + \frac{1}{n} \sum_{i=1}^{n} \omega_i K_{X_i} + \tau(\widehat{T} - T_K)(g_0 - g_{\lambda_e}) \right] \right\|_{\lambda_e} \\
&\leq \frac{\zeta}{1-\zeta} \cdot \left\| T_{\lambda_e}^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} \{Z_i(\widehat{f}) - Z_i(f_0)\} K_{X_i} + \frac{1}{n} \sum_{i=1}^{n} \omega_i K_{X_i} + \tau(\widehat{T} - T_K)(g_0 - g_{\lambda_e}) \right] \right\|_{\lambda_e} \\
&\leq 2\zeta \big(\gamma_1 + \gamma_2 + \tau\zeta \|g_0 - g_{\lambda_e}\|_{\lambda_e} \big),
\end{aligned}$$

where the second inequality is due to the fact that $\zeta \leq 1/2$. Combining the above bounds, we obtain that conditioned on $\mathscr{E}_t \cap \mathscr{B}_t$,

$$\begin{aligned}
\|\mathrm{Rem}_n\|_{\lambda_e} &\leq \gamma_2 + \tau\zeta \|g_0 - g_{\lambda_e}\|_{\lambda_e} + 2\zeta \big(\gamma_1 + \gamma_2 + \tau\zeta \|g_0 - g_{\lambda_e}\|_{\lambda_e} \big) \\
&\leq 2\gamma_2 + 2\tau\zeta \|g_0 - g_{\lambda_e}\|_{\lambda_e} + 2\zeta\gamma_1.
\end{aligned}$$

It remains to bound the (deterministic) bias term $\|g_0 - g_{\lambda_e}\|_{\lambda_e}$. Recalling that $g_0 = T_K^{r_e} g^*$ with $g^* = \sum_{j=1}^{\infty} g_j \phi_j \in \mathscr{H}$, we have

$$g_0 - g_{\lambda_e} = g_0 - T_{\lambda_e}^{-1} T_K g_0 = \sum_{j=1}^{\infty} \mu_j^{r_e} g_j \phi_j - \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda_e} \mu_j^{r_e} g_j \phi_j = \sum_{j=1}^{\infty} \frac{\lambda_e}{\mu_j + \lambda_e} \mu_j^{r_e} g_j \phi_j.$$

Hence,

$$\begin{aligned} \|g_0 - g_{\lambda_e}\|_{\lambda_e}^2 &= \sum_{j=1}^{\infty} \frac{\lambda_e^2}{(\mu_j + \lambda_e)^2} \mu_j^{2r_e} g_j^2 \cdot \frac{\mu_j + \lambda_e}{\mu_j} \\ &= \lambda_e^{1+2r_e} \sum_{j=1}^{\infty} \frac{\lambda_e^{1-2r_e} \mu_j^{2r_e}}{(\mu_j + \lambda_e)^{1-2r_e} (\mu_j + \lambda_e)^{2r_e}} \frac{g_j^2}{\mu_j} \leq \lambda_e^{1+2r_e} \sum_{j=1}^{\infty} \frac{g_j^2}{\mu_j} = \lambda_e^{2r_e+1} \|g^*\|_{\mathscr{H}}^2. \end{aligned} \tag{D.30}$$

Putting the pieces together and recalling the notations in Theorem 4.3.3, we obtain that with probability at least $1 - 8e^{-t}$,

$$\|\mathrm{Rem}_n\|_{\lambda_e} \leq 2\zeta \left(\gamma_1 + \tau \lambda_e^{r_e + 1/2} \|g^*\|_{\mathscr{H}} \right) + 2\gamma_2 \asymp \Delta_1(\lambda_e) + \Delta_2(\lambda_q, \lambda_e).$$

Combining this bound on the $\|\cdot\|_{\lambda_e}$-norm with Lemma D.3.6 completes the proof of (4.16).

To obtain the functional Bahadur representation (4.17) for the oracle estimator $\widehat{g}_{\mathrm{ora}}$, note that $\widehat{g}_{\mathrm{ora}}$ is the empirical risk minimizer of $g \to \widehat{\mathscr{L}_n}(f_0, g)$ when the true conditional quantile function is plugged in. Therefore, using a similar argument that leads to (D.25), we obtain the following closed-form expression of $\widehat{g}_{\mathrm{ora}}$:

$$\widehat{g}_{\mathrm{ora}} = \widehat{T}_{\lambda_e}^{-1} \frac{1}{\tau n} \sum_{i=1}^{n} Z_i(f_0) K_{X_i}.$$

Following a similar line of arguments that leads to (D.28), it can be shown that

$$
\begin{aligned}
&\mathrm{Rem}_{n,\mathrm{ora}} \\
&:= \tau(\widehat{g}_{\mathrm{ora}} - g_{\lambda_e}) - \frac{1}{n}\sum_{i=1}^{n} \omega_i T_{\lambda_e}^{-1} K_{X_i} \\
&= \tau T_{\lambda_e}^{-1}(\widehat{T} - T_K)(g_0 - g_{\lambda_e}) + (\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1})\left\{\frac{1}{n}\sum_{i=1}^{n} \omega_i K_{X_i} + \tau(\widehat{T} - T_K)(g_0 - g_{\lambda_e})\right\}.
\end{aligned}
$$

The rest of the proof closely resembles that in the case of $\widehat{g}_n$ and is therefore omitted here to avoid repetitive calculations. □

D.3.5 Proof of Theorem 4.3.4

Keeping the notations used in Theorem 4.3.3 and fixing $x_0 \in \mathscr{X}$, we write

$$
\rho_{\lambda_e}^2 = \rho_{\lambda_e}^2(x_0) = \frac{\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2}{\mathfrak{D}_{\lambda_e}} \text{ and } S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)
$$

throughout the proof.

We first prove the Berry-Esseen bound for the two-step estimator $\widehat{g}$. From (4.16) we see that with probability at least $1 - 8e^{-t}$,

$$
\left\| \tau(\widehat{g} - g_{\lambda_e}) - \frac{1}{n}\sum_{i=1}^{n} \omega_i T_{\lambda_e}^{-1} K_{X_i} \right\|_{\infty} \leq c_5 \mathfrak{D}_{\lambda_e}^{1/2}\{\Delta_1(\lambda_e) + \Delta_2(\lambda_q, \lambda_e)\},
$$

which implies

$$
\left|\tau\sqrt{n}(\widehat{g} - g_{\lambda_e})(x_0) - S_n\right| \leq c_5 \mathfrak{D}_{\lambda_e}^{1/2}\sqrt{n}\{\Delta_1(\lambda_e) + \Delta_2(\lambda_q, \lambda_e)\}. \tag{D.31}
$$

Applying the Berry-Esseen theorem (see, e.g. Tyurin (2011)), we obtain

$$
\sup_{u\in\mathbb{R}} \left|\mathbb{P}\left\{S_n \leq \mathrm{Var}(S_n)^{1/2} u\right\} - G(u)\right| \leq 0.5 \frac{\mathbb{E}|\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)|^3}{[\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2]^{3/2}} \frac{1}{\sqrt{n}},
$$

where $G(\cdot)$ denotes the standard normal distribution function, that is,

$$G(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} \mathrm{d}u.$$

By definition of $\rho_{\lambda_e}^2$ in (4.19),

$$[\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2]^{3/2} = \mathfrak{D}_{\lambda_e}^{3/2} \rho_{\lambda_e}^3.$$

For the third absolute moment, let $\mathbb{P}_{X_i}$ and $\mathbb{E}_{X_i}$ be the conditional probability and conditional expectation given X_i, respectively. Condition 4.3.2 implies that for any $u > 0$,

$$\begin{aligned} &\mathbb{P}_{X_i}(|\omega_i| > u) \leq \inf_{s>0} \left\{ e^{-su} \mathbb{E}_{X_i}(e^{s|\omega_i|}) \right\} \\ &\leq \inf_{s>0} \left\{ e^{-su} \mathbb{E}_{X_i}(e^{s\omega_i} + e^{-s\omega_i}) \right\} \leq \inf_{s>0} (2e^{-su} e^{\sigma_0^2 s^2/2}) = 2e^{-u^2/(2\sigma_0^2)}, \end{aligned} \tag{D.32}$$

where the first inequality follows from Markov's inequality and the last step is obtained by choosing $s = u/\sigma_0^2$. Therefore, the (conditional) third moment of $|\omega_i|$ satisfies

$$\begin{aligned} \mathbb{E}_{X_i}(|\omega_i|^3) = 3 \int_0^\infty u^2 \mathbb{P}_{X_i}(|\omega_i| > u) \mathrm{d}u &\leq 6 \int_0^\infty u^2 e^{-u^2/(2\sigma_0^2)} \mathrm{d}u \\ &= 6\sqrt{2}\sigma_0^3 \int_0^\infty u^{1/2} e^{-u} \mathrm{d}u = 3\sqrt{2\pi}\sigma_0^3, \end{aligned}$$

where the second equality is obtained by a change of variables. Moreover, note that

$$\left| T_{\lambda_e}^{-1} K_{X_i}(x_0) \right| = \left| \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda_e} \phi_j(X_i) \phi_j(x_0) \right| \leq C_\phi^2 \mathfrak{D}_{\lambda_e},$$

and $\mathbb{E}\{T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2 \leq C_\phi^2 \mathfrak{D}_{\lambda_e}$ by a similar argument as in (4.18). Combining the above bounds

yields

$$\begin{aligned}\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left\{S_n \le \mathrm{Var}(S_n)^{1/2}u\right\} - G(u)\right| &\le 0.5\frac{\mathbb{E}|\omega_i T_{\lambda_e}^{-1}K_{X_i}(x_0)|^3}{\mathfrak{D}_{\lambda_e}^{3/2}\rho_{\lambda_e}^3}\frac{1}{\sqrt{n}} \\ &\le \frac{3\sqrt{2\pi}}{2}\sigma_0^3\frac{C_\phi^2\mathfrak{D}_{\lambda_e}\mathbb{E}\{T_{\lambda_e}^{-1}K_{X_i}(x_0)\}^2}{\mathfrak{D}_{\lambda_e}^{3/2}\rho_{\lambda_e}^3}\frac{1}{\sqrt{n}} \\ &\le \frac{3\sqrt{2\pi}}{2}C_\phi^4\frac{\sigma_0^3}{\rho_{\lambda_e}^3}\sqrt{\frac{\mathfrak{D}_{\lambda_e}}{n}}. \end{aligned} \tag{D.33}$$

Let $\xi \sim \mathcal{N}(0,1)$ and recall that $\mathrm{Var}(S_n)^{1/2} = \mathfrak{D}_{\lambda_e}^{1/2}\rho_{\lambda_e}$. Denoting $\Delta_1 = \Delta_1(\lambda_e)$ and $\Delta_2 = \Delta_2(\lambda_q,\lambda_e)$ for brevity, it follows from (D.31) and (D.33) that, for any $u \in \mathbb{R}$,

$$\begin{aligned}&\mathbb{P}\left\{\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}\frac{\tau}{\rho_{\lambda_e}}(\widehat{g}-g_{\lambda_e})(x_0) \le u\right\} \\ &\le \mathbb{P}\left\{\mathrm{Var}(S_n)^{-1/2}S_n \le u + c_5\rho_{\lambda_e}^{-1}\sqrt{n}(\Delta_1+\Delta_2)\right\} + 8e^{-t} \\ &\le \mathbb{P}\left\{\xi \le u + c_5\rho_{\lambda_e}^{-1}\sqrt{n}(\Delta_1+\Delta_2)\right\} + 8e^{-t} + \frac{3\sqrt{2\pi}}{2}C_\phi^4\frac{\sigma_0^3}{\rho_{\lambda_e}^3}\sqrt{\frac{\mathfrak{D}_{\lambda_e}}{n}} \\ &\le \mathbb{P}(\xi \le u) + 8e^{-t} + \frac{3\sqrt{2\pi}}{2}C_\phi^4\frac{\sigma_0^3}{\rho_{\lambda_e}^3}\sqrt{\frac{\mathfrak{D}_{\lambda_e}}{n}} + \frac{c_5}{\sqrt{2\pi}}\rho_{\lambda_e}^{-1}\sqrt{n}(\Delta_1+\Delta_2).\end{aligned}$$

Here, the last inequality follows from the fact that $G(b)-G(a) \le (2\pi)^{-1/2}(b-a)$ for any $a \le b$. A similar argument leads to a series of reverse inequalities. The above bounds are independent of u, so that the same inequalities hold uniformly over $u \in \mathbb{R}$. Moreover, by the definition of Δ_1, $\sigma_0\sqrt{\mathfrak{D}_{\lambda_e}/n} \le \sqrt{n}\Delta_1$. This completes the proof of the Berry-Esseen bound for $\widehat{g}$ with $c_7 = c_7(C_\phi,\rho_{\lambda_e},\sigma_0) = \frac{3}{2}\sqrt{2\pi}C_\phi^4\sigma_0^2\rho_{\lambda_e}^{-3} + c_5(2\pi)^{-1/2}\rho_{\lambda_e}^{-1}$.

For the two-step oracle estimator $\widehat{g}_{\mathrm{ora}}$, the bound (4.17) yields that, with probability at least $1-6e^{-t}$,

$$\left\|\tau(\widehat{g}_{\mathrm{ora}} - g_{\lambda_e}) - \frac{1}{n}\sum_{i=1}^n \omega_i T_{\lambda_e}^{-1}K_{X_i}\right\|_\infty \le c_6\mathfrak{D}_{\lambda_e}^{1/2}\Delta_1,$$

thus implying

$$\left|\tau\sqrt{n}(\widehat{g}_{\mathrm{ora}}-g_{\lambda_e})(x_0)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\omega_i T_{\lambda_e}^{-1}K_{X_i}(x_0)\right|\leq c_6\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{n}\Delta_1.$$

By combining the above inequality with (D.33), the Berry-Esseen bound for $\widehat{g}_{\mathrm{ora}}$ can be established using a similar line of reasoning as in the case of $\widehat{g}$. The proof is omitted here for brevity. □

D.3.6 Proof of Theorem 4.3.5

Following the notation in the proof of Proposition 4.2.1, define centered random weights $U_i = W_i - \mathbb{E}W_i = W_i - 1$ for $i = 1,\ldots,n$. We begin by establishing an upper bound on the difference between $\tau\mathfrak{B}^{\flat}(x_0) = \frac{1}{n}\sum U_i Z_i(\widehat{f}) - \tau\widehat{g}(X_i)\widehat{T}_{\lambda_e}^{-1}K_{X_i}(x_0)$ and the sum $\frac{1}{n}\sum U_i\omega_i T_{\lambda_e}^{-1}K_{X_i}(x_0)$ under the $\|\cdot\|_{\lambda_e}$-norm. Recall that $\mathbb{P}^*(\cdot) = \mathbb{P}(\cdot|\mathscr{D}_n)$ denotes the conditional probability given $\mathscr{D}_n = \{(Y_i,X_i)\}_{i=1}^n$. Let $\mathbb{E}^*(\cdot) = \mathbb{E}(\cdot|\mathscr{D}_n)$ and $\mathrm{Var}^*(\cdot) = \mathrm{Var}(\cdot|\mathscr{D}_n)$ be the conditional expectation and conditional variance given $\mathscr{D}_n$, respectively.

Lemma D.3.9. Assume that Conditions 4.2.1–4.3.4 hold, and $f_0 = T_K^{r_q}f^*$ and $g_0 = T_K^{r_e}g^*$ for some $0 \leq r_q, r_e \leq 1/2$ and $f^*, g^* \in \mathscr{H}$. For any $t > 0$, let $\lambda_q \geq (\mathfrak{D}_{\lambda_q}+t)/n$, $\lambda_q^{r_q}\|f^*\|_{\mathscr{H}} \leq 1$, $\lambda_e \gtrsim (t+\log\mathfrak{D}_{\lambda_e})/n$ and $n \geq 64C_\phi^2\,\mathfrak{D}_{\lambda_e}(t+\log n)\log n$. Define $\delta_n := \delta_n(\lambda_q,n,t)$ and $\gamma_n := \gamma_n(\lambda_e,n,t)$ as

$$\delta_n = \lambda_q^{r_q+1/2}\|f^*\|_{\mathscr{H}} + \sqrt{\frac{\mathfrak{D}_{\lambda_q}+t}{n}} \quad\text{and}\quad \gamma_n = \tau\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}} + \sigma_0\sqrt{\frac{\mathfrak{D}_{\lambda_e}+t}{n}}.$$

Moreover, denote

$$\delta_{\mathrm{s}} := \delta_n\mathfrak{D}_{\lambda_q}^{1/2}\sqrt{\frac{t+\mathfrak{D}_{\lambda_e}}{n}}.$$

Then, there exists an event $\mathscr{G}(t)$ with $\mathbb{P}\{\mathscr{G}(t)\} \geq 1-12e^{-t}$ such that, with $\mathbb{P}^*$-probability at least

$1-5e^{-t}$ conditioned on $\mathscr{G}(t)$,

$$\left\| \frac{1}{n}\sum_{i=1}^{n} U_i\{Z_i(\widehat{f}) - \tau\widehat{g}(X_i)\}\widehat{T}_{\lambda_e}^{-1}K_{X_i} - \frac{1}{n}\sum_{i=1}^{n} U_i\omega_i T_{\lambda_e}^{-1}K_{X_i} \right\|_{\lambda_e} \leq c_{19}\Delta_\flat(\lambda_q,\lambda_e,n,t),$$

where

$$\Delta_\flat(\lambda_q,\lambda_e,n,t) = \gamma_n \mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{t+\log n}{n}} + (\delta_{\mathrm{s}} + \mathfrak{D}_{\lambda_q}^{1/2}\delta_n^2) = \Delta_1(\lambda_e) + \Delta_2(\lambda_q,\lambda_e),$$

and $c_{19} = c_{19}(C_\phi,\sigma_0,\sigma_W)$ is a positive constant depending only on $(C_\phi,\sigma_0,\sigma_W)$.

Now, we are ready to prove Theorem 4.3.5.

Proof of Theorem 4.3.5. Following the notations in the proof of Theorem 4.3.4 and fixing $x_0 \in \mathscr{X}$, we write $\rho_{\lambda_e}^2 = \rho_{\lambda_e}^2(x_0)$. Without loss of generality, assume that the sample size satisfies

$$n \geq 256 C_\phi^6(\sigma_0/\rho_{\lambda_e})^4 \mathfrak{D}_{\lambda_e} t. \tag{D.34}$$

Otherwise, the right-hand side of the bound (4.20) exceeds 1, so that (4.20) holds trivially.

By combining Lemmas D.3.6 and D.3.9 and denoting $\Delta_\flat = \Delta_\flat(\lambda_q,\lambda_e,n,t)$ for simplicity, it follows that, conditioned on $\mathscr{G}(t)$ defined in Lemma D.3.9, the bound

$$\left\| \frac{1}{n}\sum_{i=1}^{n} U_i\{Z_i(\widehat{f}) - \tau\widehat{g}(X_i)\}\widehat{T}_{\lambda_e}^{-1}K_{X_i} - \frac{1}{n}\sum_{i=1}^{n} U_i\,\omega_i T_{\lambda_e}^{-1}K_{X_i} \right\|_{\infty} \leq C_1 \mathfrak{D}_{\lambda_e}^{1/2}\Delta_\flat$$

holds with $\mathbb{P}^*$-probability at least $1-5e^{-t}$, where C_1 only depends on $(C_\phi,\sigma_0,\sigma_W)$. Given $x_0 \in \mathscr{X}$, define

$$S_n^\flat = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\,\omega_i T_{\lambda_e}^{-1}K_{X_i}(x_0).$$

Then, with $\mathbb{P}^*$-probability at least $1-5e^{-t}$ conditioned on $\mathscr{G}(t)$,

$$|\tau\sqrt{n}\mathfrak{B}^\flat(x_0)-S_n^\flat| = \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\{Z_i(\widehat{f})-\tau\widehat{g}(X_i)\}\widehat{T}_{\lambda_e}^{-1}K_{X_i}(x_0)-\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\,\omega_i T_{\lambda_e}^{-1}K_{X_i}(x_0)\right|$$
$$\le C_1\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{n}\Delta_\flat. \tag{D.35}$$

Applying a conditional version of the Berry-Esseen inequality, we have

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}^*\left\{S_n^\flat \le \mathrm{Var}^*(S_n^\flat)^{1/2}u\right\}-G(u)\right| \lesssim \frac{\sigma_W^3 n^{-1}\sum_{i=1}^n|\omega_i T_{\lambda_e}^{-1}K_{X_i}(x_0)|^3}{\sqrt{n}\{\mathrm{Var}^*(S_n^\flat)\}^{3/2}}, \tag{D.36}$$

where $\mathrm{Var}^*(S_n^\flat)=n^{-1}\sum_{i=1}^n\omega_i^2\{T_{\lambda_e}^{-1}K_{X_i}(x_0)\}^2$.

We then establish high probability bounds for the data-dependent quantities, $\mathrm{Var}^*(S_n^\flat)$ and $n^{-1}\sum_{i=1}^n|\omega_i T_{\lambda_e}^{-1}K_{X_i}(x_0)|^3$. Note that $\mathbb{E}\{\omega_i T_{\lambda_e}^{-1}K_{X_i}(x_0)\}^2=\rho_{\lambda_e}^2\mathfrak{D}_{\lambda_e}$ by the definition of ρ_{λ_e} in (4.19). Thus

$$\left|\mathrm{Var}^*(S_n^\flat)/\mathbb{E}\{\omega_i T_{\lambda_e}^{-1}K_{X_i}(x_0)\}^2-1\right| = \left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\frac{\omega_i^2\{T_{\lambda_e}^{-1}K_{X_i}(x_0)\}^2}{\rho_{\lambda_e}^2\mathfrak{D}_{\lambda_e}}\right|.$$

Next, we apply Bernstein's inequality to bound the right-hand side. By (D.32), we have for any $u>0$ that $\mathbb{P}_{X_i}(\omega_i^2\ge u)=\mathbb{P}_{X_i}(|\omega_i|\ge\sqrt{u})\le 2e^{-u/(2\sigma_0^2)}$, which implies $\mathbb{P}_{X_i}(\omega_i^2\ge 2\sigma_0^2 u)\le 2e^{-u}$. Then, for any $l\ge 1$, it follows that

$$\mathbb{E}_{X_i}\left(|\omega_i|^{2l}\right) = (2\sigma_0^2)^l l\int_0^\infty u^{l-1}\mathbb{P}_{X_i}\left(\omega_i^2\ge 2\sigma_0^2 u\right)\mathrm{d}u \le 2(2\sigma_0^2)^l l\int_0^\infty u^{l-1}e^{-u}\mathrm{d}u$$
$$= 2(2\sigma_0^2)^l l!. \tag{D.37}$$

Moreover,

$$\left\|T_{\lambda_e}^{-1}K_{X_i}(x_0)\right\|_\infty = \left\|\sum_{j=1}^\infty\frac{\mu_j}{\mu_j+\lambda_e}\phi_j(X_i)\phi_j(x_0)\right\|_\infty \le C_\phi^2\mathfrak{D}_{\lambda_e},$$

so that

$$\begin{aligned}
\mathbb{E}\{T_{\lambda_e}^{-1}K_{X_i}(x_0)\}^{2l} &\le \|T_{\lambda_e}^{-1}K_{X_i}(x_0)\|_\infty^{2l-2}\mathbb{E}\{T_{\lambda_e}^{-1}K_{X_i}(x_0)\}^2 \\
&= \|T_{\lambda_e}^{-1}K_{X_i}(x_0)\|_\infty^{2l-2}\mathbb{E}\left\{\sum_{j=1}^{\infty}\frac{\mu_j}{\mu_j+\lambda_e}\phi_j(X_i)\phi_j(x_0)\right\}^2 \\
&= \|T_{\lambda_e}^{-1}K_{X_i}(x_0)\|_\infty^{2l-2}\sum_{j=1}^{\infty}\left(\frac{\mu_j}{\mu_j+\lambda_e}\right)^2\phi_j^2(x_0) \\
&\le \left(C_\phi^2\mathfrak{D}_{\lambda_e}\right)^{2l-2}\sum_{j=1}^{\infty}\frac{\mu_j}{\mu_j+\lambda_e}\phi_j^2(x_0) \\
&\le \left(C_\phi^2\mathfrak{D}_{\lambda_e}\right)^{2l-2}C_\phi^2\mathfrak{D}_{\lambda_e}.
\end{aligned}$$

Combining the above inequalities, we have

$$\mathbb{E}\left[\frac{\omega_i^2\{T_{\lambda_e}^{-1}K_{X_i}(x_0)\}^2}{\rho_{\lambda_e}^2\mathfrak{D}_{\lambda_e}}\right]^2 \le 16C_\phi^6\frac{\sigma_0^4}{\rho_{\lambda_e}^4}\mathfrak{D}_{\lambda_e}$$

and for $l \ge 2$

$$\mathbb{E}\left|\frac{\omega_i^2\{T_{\lambda_e}^{-1}K_{X_i}(x_0)\}^2}{\rho_{\lambda_e}^2\mathfrak{D}_{\lambda_e}}\right|^l \le 2\left(\frac{2\sigma_0^2}{\rho_{\lambda_e}^2}\right)^l l!C_\phi^{4l-2}\mathfrak{D}_{\lambda_e}^{l-1} \le \frac{l!}{2}\cdot 16C_\phi^6\frac{\sigma_0^4}{\rho_{\lambda_e}^4}\mathfrak{D}_{\lambda_e}\cdot\left(C_\phi^4\mathfrak{D}_{\lambda_e}\frac{4\sigma_0^2}{\rho_{\lambda_e}^2}\right)^{l-2}.$$

Applying Bernstein's inequality (see, e.g. Lemma 2.2.10 in van der Vaart and Wellner (1996)), it follows that with probability at least $1-2e^{-t}$,

$$\begin{aligned}
\left|\mathrm{Var}^*(S_n^\flat)/\mathbb{E}\{\omega_i T_{\lambda_e}^{-1}K_{X_i}(x_0)\}^2 - 1\right| &\le 4\sqrt{2}C_\phi^3\frac{\sigma_0^2}{\rho_{\lambda_e}^2}\sqrt{\frac{\mathfrak{D}_{\lambda_e}t}{n}} + 8C_\phi^4\frac{\sigma_0^2}{\rho_{\lambda_e}^2}\frac{\mathfrak{D}_{\lambda_e}t}{n} \\
&\le 8C_\phi^3\frac{\sigma_0^2}{\rho_{\lambda_e}^2}\sqrt{\frac{\mathfrak{D}_{\lambda_e}t}{n}}, \qquad \text{(D.38)}
\end{aligned}$$

where the last inequality follows from the fact that $8C_\phi\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{t/n} \le 1$. Moreover, combining

(D.38) and (D.34) yields

$$\frac{1}{2}\mathbb{E}\{\omega_1 T_{\lambda_e}^{-1} K_{X_1}(x_0)\}^2 \leq \mathrm{Var}^*(S_n^\flat) = \frac{1}{n}\sum_{i=1}^n \omega_i^2 \{T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2 \leq \frac{3}{2}\mathbb{E}\{\omega_1 T_{\lambda_e}^{-1} K_{X_1}(x_0)\}^2, \quad \text{(D.39)}$$

which further implies

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n |\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)|^3 &\leq \max_{1\leq i\leq n} |\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)| \cdot \frac{1}{n}\sum_{i=1}^n |\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)|^2 \\ &\leq \frac{3}{2}\max_{1\leq i\leq n} |\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)| \cdot \mathbb{E}\{\omega_1 T_{\lambda_e}^{-1} K_{X_1}(x_0)\}^2. \quad \text{(D.40)}\end{aligned}$$

By Condition 4.3.2 and the fact that $\|T_{\lambda_e}^{-1} K_{X_i}(x_0)\|_\infty \leq C_\phi^2 \mathfrak{D}_{\lambda_e}$, we have

$$\begin{aligned}\mathbb{P}\left\{\max_{1\leq i\leq n} |\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)| \geq C_\phi^2 \sigma_0 \mathfrak{D}_{\lambda_e}\sqrt{2(t+\log n)}\right\} &\leq \mathbb{P}\left\{\max_{1\leq i\leq n} |\omega_i| \geq \sigma_0\sqrt{2(t+\log n)}\right\} \\ &\leq n\mathbb{P}\left\{|\omega_1| \geq \sigma_0\sqrt{2(t+\log n)}\right\} \\ &\leq 2e^{-t},\end{aligned}$$

where the last inequality follows from (D.32). In view of this and (D.38), the event $\mathscr{G}'(t)$, defined as

$$\mathscr{G}'(t) := \left\{\left|\frac{\mathrm{Var}^*(S_n^\flat)}{\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2} - 1\right| \leq \frac{1}{2}, \max_{1\leq i\leq n} |\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)| \leq C_\phi^2 \sigma_0 \mathfrak{D}_{\lambda_e}\sqrt{2(t+\log n)}\right\},$$

satisfies $\mathbb{P}\{\mathscr{G}'(t)\} \geq 1 - 4e^{-t}$ given the sample size requirement (D.34). Moreover, it follows from (D.36), (D.39) and (D.40) that, conditioned on the event $\mathscr{G}'(t)$,

$$\begin{aligned}\sup_{u\in\mathbb{R}}\left|\mathbb{P}^*\left\{S_n^\flat \leq \mathrm{Var}^*(S_n^\flat)^{1/2} u\right\} - G(u)\right| &\leq C_3 \frac{\sigma_W^3 C_\phi^2 \sigma_0 \sqrt{t+\log n}}{\sqrt{n}} \cdot \frac{\mathfrak{D}_{\lambda_e}\mathbb{E}\{\omega_1 T_{\lambda_e}^{-1} K_{X_1}(x_0)\}^2}{[\mathbb{E}\{\omega_1 T_{\lambda_e}^{-1} K_{X_1}(x_0)\}^2]^{3/2}} \\ &= C_3 \frac{\sigma_W^3 C_\phi^2 \sigma_0}{\rho_{\lambda_e}}\sqrt{\frac{\mathfrak{D}_{\lambda_e}(t+\log n)}{n}}, \quad \text{(D.41)}\end{aligned}$$

where the last line uses the fact that $\mathbb{E}\{\omega_1 T_{\lambda_e}^{-1} K_{X_1}(x_0)\}^2 = \rho_{\lambda_e}^2 \mathfrak{D}_{\lambda_e}$, and $C_3 > 0$ is an absolute constant.

To compare the two normal distribution functions $G(u/[\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2]^{1/2})$ and $G(u/\mathrm{Var}^*(S_n^\flat)^{1/2})$, we apply the following inequality that can be derived from Pinsker's inequality (see, e.g. Lemma A.7 in the supplement of Spokoiny and Zhilova (2015)):

$$\sup_{u\in\mathbb{R}} \left| G\left(\frac{u}{\mathrm{Var}^*(S_n^\flat)^{1/2}}\right) - G\left(\frac{u}{[\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2]^{1/2}}\right)\right| \le \frac{1}{2}\left|\frac{\mathrm{Var}^*(S_n^\flat)}{\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2} - 1\right| \quad \text{(D.42)}$$

as long as $|\mathrm{Var}^*(S_n^\flat)/\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2 - 1| \le 1/2$. Recall the notations $\Delta_1 = \Delta_1(\lambda_e)$ and $\Delta_2 = \Delta_2(\lambda_q, \lambda_e)$ in Theorem 4.3.4. Applying Lemma D.3.6 and (D.30) gives

$$|g_{\lambda_e}(x_0) - g_0(x_0)| \le \|g_{\lambda_e} - g_0\|_\infty \le C_\phi \mathfrak{D}_{\lambda_e}^{1/2} \lambda_e^{r_e+1/2} \|g^*\|_{\mathscr{H}}.$$

Combining the above calculations with Theorem 4.3.4, we conclude that for any $u \in \mathbb{R}$, there exists a constant C_2 depending only on $(\sigma_W, \sigma_0, C_\phi, \rho_{\lambda_e})$ such that, conditioned on $\mathscr{G}'(t)$,

$$\begin{aligned}
&\mathbb{P}\big\{\tau\sqrt{n}(\widehat{g} - g_0)(x_0) \le u\big\} \\
&= \mathbb{P}\big\{\tau\sqrt{n}(\widehat{g} - g_{\lambda_e})(x_0) \le u - \tau\sqrt{n}(g_{\lambda_e} - g_0)(x_0)\big\} \\
&\le G\left(\frac{u - \tau\sqrt{n}(g_{\lambda_e} - g_0)(x_0)}{[\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2]^{1/2}}\right) + c_7\sqrt{n}(\Delta_1 + \Delta_2) + 8e^{-t} \\
&\le G\left(\frac{u - C_1 \mathfrak{D}_{\lambda_e}^{1/2} \Delta_\flat \sqrt{n}}{[\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2]^{1/2}}\right) + c_7\sqrt{n}(\Delta_1 + \Delta_2) + 8e^{-t} + \frac{\sqrt{n}\{C_1\Delta_\flat + \tau C_\phi \lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}\}}{\sqrt{2\pi}\rho_{\lambda_e}} \\
&\le G\left(\frac{u - C_1 \mathfrak{D}_{\lambda_e}^{1/2} \Delta_\flat \sqrt{n}}{\mathrm{Var}^*(S_n^\flat)^{1/2}}\right) + C_2\sqrt{n}(\Delta_1 + \Delta_2 + \tau\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}) + 8e^{-t} + 4C_\phi^3 \frac{\sigma_0^2}{\rho_{\lambda_e}^2}\sqrt{\frac{\mathfrak{D}_{\lambda_e} t}{n}}.
\end{aligned} \quad \text{(D.43)}$$

Here, the second inequality follows from the fact that for any $u_1 \le u_2$,

$$G\bigg(\frac{u_2}{[\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2]^{1/2}}\bigg) - G\bigg(\frac{u_1}{[\mathbb{E}\{\omega_i T_{\lambda_e}^{-1} K_{X_i}(x_0)\}^2]^{1/2}}\bigg) \le \frac{u_2 - u_1}{(2\pi\rho_{\lambda_e}^2 \mathfrak{D}_{\lambda_e})^{1/2}},$$

and the last inequality is a consequence of (D.38) and (D.42). Moreover, combining (D.35) and (D.41) yields that conditioned on $\mathscr{G}(t) \cap \mathscr{G}'(t)$,

$$\begin{aligned} G\bigg(\frac{u - C_1 \mathfrak{D}_{\lambda_e}^{1/2} \Delta_\flat \sqrt{n}}{\mathrm{Var}^*(S_n^\flat)^{1/2}}\bigg) &\le \mathbb{P}^*\{S_n^\flat \le u - C_1 \mathfrak{D}_{\lambda_e}^{1/2} \Delta_\flat \sqrt{n}\} + C_3 \frac{\sigma_W^3 C_\phi^2 \sigma_0}{\rho_{\lambda_e}} \sqrt{\frac{\mathfrak{D}_{\lambda_e}(t + \log n)}{n}} \\ &\le \mathbb{P}^*\{\tau\sqrt{n}\mathfrak{B}^\flat(x_0) \le u\} + C_3 \frac{\sigma_W^3 C_\phi^2 \sigma_0}{\rho_{\lambda_e}} \sqrt{\frac{\mathfrak{D}_{\lambda_e}(t + \log n)}{n}} + 5e^{-t}. \end{aligned}$$

This, joint with (D.43), yields that conditioned on $\mathscr{G}(t) \cap \mathscr{G}'(t)$ that satisfies $\mathbb{P}\{\mathscr{G}(t) \cap \mathscr{G}'(t)\} \ge 1 - 16e^{-t}$,

$$\mathbb{P}\big\{\widehat{g}(x_0) - g_0(x_0) \le u\big\} - \mathbb{P}^*\big\{\mathfrak{B}^\flat(x_0) \le u\big\} \le c_9\sqrt{n}(\Delta_1 + \Delta_2 + \tau\lambda_e^{r_e + 1/2}\|g^*\|_{\mathscr{H}}) + 13e^{-t},$$

where $c_9 = c_9(C_\phi, \sigma_0, \sigma_W, \rho_{\lambda_e}) > 0$. A similar argument leads to a series of reverse inequalities, which completes the proof. □

D.3.7 Proof of Theorem 4.3.6

Following the notations in the proof of Theorem 4.3.5 and fixing $x_0 \in \mathscr{X}$, we write $\rho_{\lambda_e}^2 = \rho_{\lambda_e}^2(x_0)$, $\Delta_1 = \Delta_1(\lambda_e)$ and $\Delta_2 = \Delta_2(\lambda_q, \lambda_e)$ for brevity. We first claim the following anti-concentration inequality for $\widehat{g}(x_0) - g_0(x_0)$:

$$\mathbb{P}\big\{\widehat{g}(x_0) - g_0(x_0) > u\big\} \le \mathbb{P}\big\{\widehat{g}(x_0) - g_0(x_0) > u + \eta\big\} + \Delta_{\mathrm{anti}}(\eta) + 16e^{-t} \tag{D.44}$$

for any $u \in \mathbb{R}$ and $\eta \geq 0$, where

$$\Delta_{\text{anti}}(\eta) := 2c_7\sqrt{n}(\Delta_1 + \Delta_2) + \frac{\tau}{\sqrt{2\pi}\rho_{\lambda_e}}\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}\eta. \tag{D.45}$$

We prove (D.44) through Gaussian approximation. By Theorem 4.3.4, it holds for any $u' \in \mathbb{R}$ that

$$\mathbb{P}\big\{\widehat{g}(x_0) - g_{\lambda_e}(x_0) > u'\big\} \leq \mathbb{P}\left(\xi > \frac{\tau}{\rho_{\lambda_e}}\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}u'\right) + c_7\sqrt{n}(\Delta_1 + \Delta_2) + 8e^{-t},$$

where $\xi \sim \mathscr{N}(0,1)$. Moreover,

$$\begin{aligned}
\mathbb{P}\left(\xi > \frac{\tau}{\rho_{\lambda_e}}\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}u'\right) &\leq \mathbb{P}\left\{\xi > \frac{\tau}{\rho_{\lambda_e}}\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}(u' + \eta)\right\} + \frac{\tau}{\sqrt{2\pi}\rho_{\lambda_e}}\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}\eta \\
&\leq \mathbb{P}\big\{\widehat{g}(x_0) - g_{\lambda_e}(x_0) > u' + \eta\big\} + \frac{\tau}{\sqrt{2\pi}\rho_{\lambda_e}}\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}\eta \\
&\quad + c_7\sqrt{n}(\Delta_1 + \Delta_2) + 8e^{-t},
\end{aligned}$$

where the first inequality uses $G(b) - G(a) \leq (2\pi)^{-1/2}(b-a)$ for any $a \leq b$ and the second follows from Theorem 4.3.4 again. Combining the above inequalities, we have

$$\begin{aligned}
&\mathbb{P}\big\{\widehat{g}(x_0) - g_{\lambda_e}(x_0) > u'\big\} - \mathbb{P}\big\{\widehat{g}(x_0) - g_{\lambda_e}(x_0) > u' + \eta\big\} \\
&\quad \leq 2c_7\sqrt{n}(\Delta_1 + \Delta_2) + \frac{\tau}{\sqrt{2\pi}\rho_{\lambda_e}}\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}\eta + 16e^{-t}.
\end{aligned}$$

Since the above inequality holds for any $u' \in \mathbb{R}$, the claim (D.44) follows by taking $u' = u + g_0(x_0) - g_{\lambda_e}(x_0)$.

Now, we are ready to prove Theorem 4.3.6. Define

$$u_\alpha = u_\alpha(x_0) = \inf\big\{u \in \mathbb{R} : \mathbb{P}\{\widehat{g}(x_0) - g_0(x_0) > u\} \leq \alpha\big\} \quad \text{for } \alpha \in (0,1),$$

and denote

$$\Delta' = c_9\sqrt{n}(\Delta_1 + \Delta_2 + \tau\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}) + 13e^{-t}, \tag{D.46}$$

where c_9 is defined in Theorem 4.3.5. Theorem 4.3.5 ensures that there exists an event $\mathscr{G}(t)$ satisfying $\mathbb{P}\{\mathscr{G}(t)\} \geq 1 - 15e^{-t}$ such that conditioned on $\mathscr{G}(t)$,

$$\mathbb{P}^*\{\mathfrak{B}^\flat(x_0) > u_{\alpha-\Delta'}\} \begin{cases} = 0 < \alpha & \text{if } \alpha \leq \Delta', \\ \leq \mathbb{P}\big\{\widehat{g}(x_0) - g_0(x_0) > u_{\alpha-\Delta'}\big\} + \Delta' \leq \alpha & \text{if } \alpha > \Delta'. \end{cases}$$

This implies that $u_\alpha^\flat$ defined in (4.15) satisfies $u_\alpha^\flat \leq u_{\alpha-\Delta'}$. Similarly, conditioned on the same event $\mathscr{G}(t)$, it holds

$$\mathbb{P}\big\{\widehat{g}(x_0) - g_0(x_0) > u_\alpha^\flat\big\} \leq \mathbb{P}^*\{\mathfrak{B}^\flat(x_0) > u_\alpha^\flat\} + \Delta' \leq \alpha + \Delta',$$

which implies $u_{\alpha+\Delta'} \leq u_\alpha^\flat$. In sum, conditioned on $\mathscr{G}(t)$, we have

$$u_{\alpha+\Delta'} \leq u_\alpha^\flat \leq u_{\alpha-\Delta'}. \tag{D.47}$$

Remark that by the definition of u_α, if the distribution of $\widehat{g}(x_0) - g_0(x_0)$ is continuous at u_α, then $\mathbb{P}\{\widehat{g}(x_0) - g_0(x_0) > u_\alpha\} = \alpha$; otherwise, $\mathbb{P}\{\widehat{g}(x_0) - g_0(x_0) > u_\alpha - \eta\} \geq \alpha$ for any $\eta > 0$. Combining this observation with (D.47) and the anti-concentration inequality (D.44), we obtain

$$\begin{aligned}
&\mathbb{P}\big\{\widehat{g}(x_0) - g_0(x_0) > u_\alpha^\flat\big\} \\
&\geq \mathbb{P}\big\{\widehat{g}(x_0) - g_0(x_0) > u_{\alpha-\Delta'}\big\} - 15e^{-t} \\
&\geq \mathbb{P}\left\{\widehat{g}(x_0) - g_0(x_0) > u_{\alpha-\Delta'} - \frac{\mathfrak{D}_{\lambda_e}^{1/2}}{n}\right\} - \Delta_{\text{anti}}\left(\frac{\mathfrak{D}_{\lambda_e}^{1/2}}{n}\right) - 31e^{-t} \\
&\geq \alpha - \Delta' - \Delta_{\text{anti}}\left(\frac{\mathfrak{D}_{\lambda_e}^{1/2}}{n}\right) - 31e^{-t}.
\end{aligned}$$

Similarly, it can shown from (D.47) that

$$\mathbb{P}\{\widehat{g}(x_0) - g_0(x_0) > u_\alpha^\flat\} \le \mathbb{P}\{\widehat{g}(x_0) - g_0(x_0) > u_{\alpha+\Delta'}\} + 15e^{-t} \le \alpha + \Delta' + 15e^{-t}.$$

Combining the above two bounds with the definitions of Δ_{anti} in (D.45) and Δ' in (D.46), we conclude that for any $\alpha \in (0,1)$,

$$\begin{aligned}\left|\mathbb{P}\{\widehat{g}(x_0) - g_0(x_0) > u_\alpha^\flat\} - \alpha\right| &\le \Delta' + \Delta_{\text{anti}}\left(\frac{\mathfrak{D}_{\lambda_e}^{1/2}}{n}\right) + 31e^{-t} \\ &\le C_1\sqrt{n}(\Delta_1 + \Delta_2 + \tau\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}) + 31e^{-t},\end{aligned}$$

where $C_1 = C_1(C_\phi, \sigma_0, \sigma_W, \rho_{\lambda_e})$ is a positive constant. This proves the claim of the theorem by noting that

$$\begin{aligned}&\left|\mathbb{P}\{g_0(x_0) \in \mathscr{I}_\alpha^\flat(x_0)\} - (1-\alpha)\right| \\ &\le \left|\mathbb{P}\{\widehat{g}(x_0) - g_0(x_0) > u_{\alpha/2}^\flat\} - \alpha/2\right| + \left|\mathbb{P}\{\widehat{g}(x_0) - g_0(x_0) < u_{1-\alpha/2}^\flat\} - \alpha/2\right| \\ &= \left|\mathbb{P}\{\widehat{g}(x_0) - g_0(x_0) > u_{\alpha/2}^\flat\} - \alpha/2\right| + \left|\mathbb{P}\{\widehat{g}(x_0) - g_0(x_0) \ge u_{1-\alpha/2}^\flat\} - (1-\alpha/2)\right|.\end{aligned}$$

□

D.4 Proofs for Section D.2

In this section, we give the proofs of the results in Section D.2.

D.4.1 Proof of Corollary D.2.1

To begin with, recall that $\lambda_q \asymp \lambda_e \asymp (m+t)/n$. Since $\mathfrak{D}_{\lambda_q} \le m$, it follows by Theorem 4.3.1 that $\|\widehat{f} - f_0\|_2^2 + \lambda_q\|\widehat{f} - f_0\|_{\mathscr{H}}^2 \lesssim (m+t)/n$ with probability at least $1 - e^{-t}$, establishing the claimed bound for $\widehat{f}$. Moreover, conditioned on this event, Theorem 4.3.2 yields that,

with probability at least $1-6e^{-t}$

$$\tau\|\widehat{g}-g_0\|_2 \lesssim \sigma_0\sqrt{\frac{m+t}{n}}+C_\phi\,\mathfrak{D}_{\lambda_q}^{1/2}\frac{m+t}{n}+\bar{p}\|\widehat{f}-f_0\|_4^2. \tag{D.48}$$

Applying (D.29) gives $\|\widehat{f}-f_0\|_4^2 \lesssim C_\phi\mathfrak{D}_{\lambda_q}^{1/2}(m+t)/n$. Combining this with (D.48) and the fact that $\mathfrak{D}_{\lambda_q} \le \lambda_q^{-1}$ establishes the bound for $\widehat{g}$, which completes the proof. □

D.4.2 Proof of Corollary D.2.2

Since $\lambda_q \asymp \lambda_e \asymp (m+\log n)/n$ and $\mathfrak{D}_\lambda \le m$ for any $\lambda > 0$, we have

$$\delta_n(\lambda_q,n,\log n) \lesssim \sqrt{\frac{m+\log n}{n}} \quad \text{and} \quad \gamma_n(\lambda_e,n,\log n) \lesssim \sqrt{\frac{m+\log n}{n}}.$$

Then, it follows that

$$\Delta_1(\lambda_e) = \gamma_n\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{\log n}{n}} \lesssim \frac{\sqrt{m(\log n)(m+\log n)}}{n} = o(n^{-1/2})$$

and

$$\Delta_2(\lambda_q,\lambda_e)\delta_n\left\{\mathfrak{D}_{\lambda_q}^{1/2}\sqrt{\frac{\log n+\mathfrak{D}_{\lambda_e}}{n}}+\mathfrak{D}_{\lambda_q}^{1/2}\delta_n\right\} \lesssim \sqrt{m}\frac{m+\log n}{n} = o(n^{-1/2})$$

provided that $m^3 = o(n)$. Moreover, $\rho_{\lambda_e}^2(x_0) \to \rho^2(x_0)$ and $\mathfrak{D}_{\lambda_e}/m \to 1$ as $n \to \infty$. Thus, applying Corollary 4.3.1 gives

$$\tau\sqrt{\frac{n}{m}}(\widehat{g}-g_{\lambda_e})(x_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\rho^2(x_0)). \tag{D.49}$$

For the bias term $(g_{\lambda_e} - g_0)(x_0)$, we first remark that there exists $g^* = \sum_{j=1}^m g_j \phi_j \in \mathscr{H}$ satisfying $g_0 = T_K^{1/2} g^*$ when $g_0 \in \mathscr{H}$ and $\mu_m > 0$. To see this, note that

$$g_0 = \sum_{j=1}^m \langle g_0, \phi_j \rangle_2 \phi_j = \sum_{j=1}^m \frac{\langle g_0, \phi_j \rangle_2}{\sqrt{\mu_j}} \sqrt{\mu_j} \phi_j =: \sum_{j=1}^m g_j \sqrt{\mu_j} \phi_j,$$

where $g_j = \langle g_0, \phi_j \rangle_2 / \sqrt{\mu_j}$ for $1 \le j \le m$. Then, it is obvious that $g_0 = T_K^{1/2} g^*$ and g^* satisfies $\|g^*\|_{\mathscr{H}}^2 = \sum_{j=1}^m \langle g_0, \phi_j \rangle_2^2 / \mu_j^2 \le \mu_m^{-1} \|g_0\|_{\mathscr{H}}^2 < \infty$. Combining this with (D.30) and Lemma D.3.6 gives

$$|(g_{\lambda_e} - g_0)(x_0)| \le \|g_{\lambda_e} - g_0\|_\infty \le C_\phi \mathfrak{D}_{\lambda_e}^{1/2} \lambda_e \|g^*\|_{\mathscr{H}} \le C_\phi \mathfrak{D}_{\lambda_e}^{1/2} \lambda_e \mu_m^{-1/2} \|g_0\|_{\mathscr{H}}.$$

Thus,

$$\sqrt{\frac{n}{m}} |(g_{\lambda_e} - g_0)(x_0)| \lesssim \sqrt{n} \sqrt{\frac{\mathfrak{D}_{\lambda_e}}{m}} \cdot \frac{m + \log n}{n} = o(1),$$

where the last equality is derived by the assumption $m^3 = o(n)$. Combining this with (D.49) establishes the claim. □

D.4.3 Proof of Corollary D.2.3

We show that $\Delta_1(\lambda_e) = o(n^{-1/2})$ and $\Delta_2(\lambda_q, \lambda_e) = o(n^{-1/2})$ in the proof of Corollary D.2.2. Moreover, there exists $g^* \in \mathscr{H}$ satisfying $g_0 = T_K^{1/2} g^*$ and $\sqrt{n}\lambda_e = o(1)$ under the given assumptions. By examining the proof of Theorem 4.3.6, we observe that the prefactor term c_{10} depends on the inverse of $\rho_{\lambda_e}(x_0)$; that is, c_{10} is uniformly bounded as long as $\rho_{\lambda_e}(x_0)$ is uniformly lower bounded by some constant for any sufficiently small $\lambda_e > 0$. Thus, applying Theorem 4.3.6 with $t = \log n$ leads to

$$|\mathbb{P}\{g_0(x_0) \in \mathscr{I}_\alpha^\flat(x_0)\} - (1 - \alpha)| = o(1),$$

thereby completing the proof.

□

D.4.4 Proof of Corollary D.2.4

For completeness, we first demonstrate that $\mathfrak{D}_\lambda \asymp \lambda^{-1/\beta}$ for any $\lambda \in (0,1)$ when the kernel has β-polynomially decaying eigenvalues with $\beta > 1$. To establish this, note that

$$\mathfrak{D}_\lambda = \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda} \asymp \sum_{j=1}^{\infty} \frac{1}{1+\lambda j^\beta}.$$

For a fixed λ, $(1+\lambda j^\beta)$ is decreasing with respect to j, so we can further deduce that

$$\begin{aligned}\mathfrak{D}_\lambda \lesssim \int_0^\infty \frac{1}{1+\lambda x^\beta}\mathrm{d}x = \lambda^{-1/\beta}\int_0^\infty \frac{1}{1+x^\beta}\mathrm{d}x &= \lambda^{-1/\beta}\left(\int_0^1 \frac{1}{1+x^\beta}\mathrm{d}x + \int_1^\infty \frac{1}{1+x^\beta}\mathrm{d}x\right) \\ &\leq \lambda^{-1/\beta}\left(1+\frac{1}{\beta-1}\right) = \frac{\beta}{\beta-1}\lambda^{-1/\beta}, \quad \text{(D.50)}\end{aligned}$$

where the first equality is derived by a change of variables. Similarly,

$$\mathfrak{D}_\lambda \gtrsim \int_1^\infty \frac{1}{1+\lambda x^\beta}\mathrm{d}x \geq \lambda^{-1/\beta}\int_1^\infty \frac{1}{1+x^\beta}\mathrm{d}x \geq \frac{\lambda^{-1/\beta}}{2}\int_1^\infty \frac{1}{x^\beta}\mathrm{d}x = \frac{1}{2(\beta-1)}\lambda^{-1/\beta},$$

where the second inequality follows from a change of variables and the fact that $\lambda < 1$. Combining the above two bounds shows that $\mathfrak{D}_\lambda \asymp \lambda^{-1/\beta}$.

Since we choose

$$\lambda_q \asymp n^{-\beta/\{(2r_q+1)\beta+1\}} + \frac{t}{n},$$

Theorem 4.3.1 yields that, with probability at least $1-e^{-t}$,

$$\|\widehat{f}-f_0\|_2^2+\lambda_q\|\widehat{f}-f_0\|_{\mathscr{H}}^2 \lesssim \lambda_q^{2r_q+1}+\frac{\mathfrak{D}_{\lambda_q}+t}{n} \asymp \lambda_q^{2r_q+1}+\frac{\lambda_q^{-1/\beta}+t}{n} \lesssim n^{-(2r_q+1)\beta/\{(2r_q+1)\beta+1\}}+\frac{t}{n} \tag{D.51}$$

provided that $0<t<n$.

Turning to $\widehat{g}$, conditioned on the event where the inequality (D.51) holds, applying (D.29) gives

$$\|\widehat{f}-f_0\|_4^2 \lesssim C_\phi \mathfrak{D}_{\lambda_q}^{1/2}\left[n^{-(2r_q+1)\beta/\{(2r_q+1)\beta+1\}}+\frac{t}{n}\right] \lesssim n^{-\{(4r_q+2)\beta-1\}/\{(4r_q+2)\beta+2\}}+\sqrt{\frac{t}{n}}.$$

Then, Theorem 4.3.2 implies that, conditioned on the event where the inequality (D.51) holds, it follows with probability at least $1-6e^{-t}$ that

$$\begin{aligned}\tau\|\widehat{g}-g_0\|_2 &\lesssim \lambda_e^{r_e+1/2}+\sigma_0\sqrt{\frac{\mathfrak{D}_{\lambda_e}+t}{n}}+n^{-\{(4r_q+2)\beta-1\}/\{(4r_q+2)\beta+2\}}+\sqrt{\frac{t}{n}} \\ &\quad +C_\phi\left[n^{-(2r_q+1)\beta/\{(4r_q+2)\beta+2\}}+\sqrt{\frac{t}{n}}\right]\lambda_q^{-1/(2\beta)}\sqrt{\frac{t+\lambda_e^{-1/\beta}}{n}} \\ &\lesssim n^{-(2r_e+1)\beta/\{(4r_e+2)\beta+2\}}+n^{-\{(4r_q+2)\beta-1\}/\{(4r_q+2)\beta+2\}}+\sqrt{\frac{t}{n}}.\end{aligned}$$

This proves the claim. □

D.4.5 Sufficient conditions for (D.1)

By Corollary D.2.4, $\widehat{g}$ attains the minimax optimal convergence rate when (r_q, r_e, β) satisfy (D.1). The following lemma identifies several sufficient conditions for (D.1), but additional conditions may also exist.

Lemma D.4.1 (Sufficient conditions for (D.1))**.** Let $0 \le r_q, r_e \le 1/2$ and $\beta > 1$. Then, (r_q, r_e, β)

satisfy (D.1) when one of the following inequalities holds:

- $(2r_q+1)\beta \geq 2$;
- $r_e \leq r_q$; or
- $\beta \geq (\sqrt{3}+1)/2$.

Proof. Recall the inequality (D.1),

$$\{(2r_q+1)\beta-2\}\{(2r_e+1)\beta+2\} \geq -3.$$

First, it is obvious that $(2r_e+1)\beta+2>0$, so the inequality holds when $(2r_q+1)\beta-2\geq 0$. Second, if $r_e \leq r_q$, then it follows for any $0 \leq r_e, r_q \leq 1/2$ and $\beta > 1$ that

$$\begin{aligned}\{(2r_q+1)\beta-2\}\{(2r_e+1)\beta+2\} &= (2r_q+1)(2r_e+1)\beta^2-4(r_e-r_q)\beta-4\\ &\geq \beta^2-4>-3,\end{aligned}$$

implying the inquality (D.1).

Finally, we remark that by the quadratic formula, any $r_q, r_e \in [0,1/2]$ satisfy (D.1) if

$$\begin{aligned}\beta &\geq \sup_{0\leq r_q,r_e\leq 1/2} \frac{-2r_q+2r_e+\sqrt{4(r_q-r_e)^2+(2r_q+1)(2r_e+1)}}{(2r_q+1)(2r_e+1)}\\ &= \sup_{0\leq r_q,r_e\leq 1/2} \frac{1}{2r_q-2r_e+\sqrt{4(r_q-r_e)^2+(2r_q+1)(2r_e+1)}}\\ &= \Big[\inf_{1\leq x,y\leq 2}\big\{\sqrt{(x-y)^2+xy}+x-y\big\}\Big]^{-1} = \frac{\sqrt{3}+1}{2},\end{aligned}$$

where the last equality follows from the fact that the infimum is obtained at $x=1$ and $y=2$. This completes the proof. □

D.4.6 Proof of Corollary D.2.5

Recall that $\lambda_q \asymp n^{-\beta/\{(2r_q+1)\beta+1\}}$ and $\lambda_e \asymp n^{-\iota}$ with $1/(2r_e+1) < \iota < \min(1,\beta/3)$. Since $\mathfrak{D}_\lambda \asymp \lambda^{-1/\beta}$ for $\lambda \in (0,1)$, we have

$$\delta_n(\lambda_q, n, \log n) = \lambda_q^{r_q+1/2}\|f^*\|_{\mathscr{H}} + \sqrt{\frac{\mathfrak{D}_{\lambda_q} + \log n}{n}} \lesssim n^{-(2r_q+1)\beta/\{(4r_q+2)\beta+2\}},$$

and

$$\begin{aligned}\gamma_n(\lambda_e, n, \log n) = \tau\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}} + \sigma_0\sqrt{\frac{\mathfrak{D}_{\lambda_e} + \log n}{n}} &\lesssim n^{-\iota(r_e+1/2)} + n^{(\iota-\beta)/(2\beta)} \\ &\lesssim n^{(\iota-\beta)/(2\beta)}.\end{aligned}$$

Then, it follows that

$$\Delta_1(\lambda_e) = \gamma_n \mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{\log n}{n}} \lesssim n^{(\iota-\beta)/\beta}\log^{1/2}(n) = o(n^{-1/2}),$$

where the last equality follows by the assumption $\iota < \beta/3$. Moreover,

$$\begin{aligned}&\Delta_2(\lambda_q, \lambda_e) \\ &= \delta_n\left\{\mathfrak{D}_{\lambda_q}^{1/2}\sqrt{\frac{\log n + \mathfrak{D}_{\lambda_e}}{n}} + \mathfrak{D}_{\lambda_q}^{1/2}\delta_n\right\} \\ &\lesssim \delta_n\sqrt{\frac{\mathfrak{D}_{\lambda_q}\log(n)}{n}} + \delta_n\sqrt{\frac{\mathfrak{D}_{\lambda_q}\mathfrak{D}_{\lambda_e}}{n}} + \mathfrak{D}_{\lambda_q}^{1/2}\delta_n^2 \\ &\lesssim n^{-(2r_q+1)\beta/\{(2r_q+1)\beta+1\}}\log^{1/2}(n) + n^{-(2r_q+1)\beta/\{(2r_q+1)\beta+1\}}n^{\iota/(2\beta)} \\ &\quad + n^{\{1-(4r_q+2)\beta\}/\{(4r_q+2)\beta+2\}} \\ &= o(n^{-1/2}),\end{aligned}$$

where the last equality follows from the assumption $(2r_q+1)\beta > 2$ and $\iota < \beta/3$. Combining the above bounds with Corollary 4.3.1 gives

$$\tau\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}(\widehat{g}-g_{\lambda_e})(x_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\rho^2(x_0)). \tag{D.52}$$

For the bias term $(g_{\lambda_e}-g_0)(x_0)$, combining Lemma D.3.6 and (D.30) yields

$$|(g_{\lambda_e}-g_0)(x_0)| \le \|g_{\lambda_e}-g_0\|_\infty \le C_\phi \mathfrak{D}_{\lambda_e}^{1/2}\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}.$$

Thus,

$$\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}|(g_{\lambda_e}-g_0)(x_0)| \lesssim \sqrt{n}n^{-\iota(r_e+1/2)} = o(1),$$

where the last inequality is derived by the assumption $\iota > 1/(2r_e+1)$. Combining this with (D.52) establishes the claim. □

D.4.7 Proof of Corollary D.2.6

The proof follows a similar line of argument as the proof of Corollary D.2.3; therefore, we omit it for brevity. □

D.4.8 Proof of Corollary D.2.7

We first show that when the kernel has β-exponentially decaying eigenvalues with $\beta > 0$, $\mathfrak{D}_\lambda \asymp \log^{1/\beta}(1/\lambda)$ for any $\lambda \in (0,1)$. To establish this, we have that for some $c_\beta > 0$,

$$\mathfrak{D}_\lambda = \sum_{j=1}^{\infty}\frac{\mu_j}{\mu_j+\lambda} \asymp \sum_{j=1}^{\infty}\frac{1}{1+\lambda e^{c_\beta j^\beta}}.$$

Thus, we can deduce that

$$\begin{aligned}\mathfrak{D}_\lambda \lesssim \int_0^\infty \frac{1}{1+\lambda e^{c_\beta x^\beta}}\mathrm{d}x &= \int_0^{\{c_\beta^{-1}\log(1/\lambda)\}^{1/\beta}} \frac{1}{1+\lambda e^{c_\beta x^\beta}}\mathrm{d}x + \int_{\{c_\beta^{-1}\log(1/\lambda)\}^{1/\beta}}^\infty \frac{1}{1+\lambda e^{c_\beta x^\beta}}\mathrm{d}x \\ &\le \left\{\frac{\log(1/\lambda)}{c_\beta}\right\}^{1/\beta} + \frac{1}{\beta c_\beta}\int_0^\infty \frac{1}{(1+e^x)\{x+\log(1/\lambda)\}^{(\beta-1)/\beta}}\mathrm{d}x \\ &\le \left\{\frac{\log(1/\lambda)}{c_\beta}\right\}^{1/\beta} + \frac{\log^{-(\beta-1)/\beta}(1/\lambda)}{\beta c_\beta}\int_0^\infty \frac{1}{1+e^x}\mathrm{d}x \\ &\lesssim \log^{1/\beta}(1/\lambda), \end{aligned} \tag{D.53}$$

where the second inequality follows from a change of variables. Moreover,

$$\begin{aligned}\mathfrak{D}_\lambda \gtrsim \int_1^\infty \frac{1}{1+\lambda e^{c_\beta x^\beta}}\mathrm{d}x \ge \int_1^{\{c_\beta^{-1}\log(1/\lambda)\}^{1/\beta}} \frac{1}{1+\lambda e^{c_\beta x^\beta}}\mathrm{d}x &\ge \frac{1}{2}\big[\{c_\beta^{-1}\log(1/\lambda)\}^{1/\beta} - 1\big] \\ &\gtrsim \log^{1/\beta}(1/\lambda). \end{aligned} \tag{D.54}$$

Combining the above two bounds shows that $\mathfrak{D}_\lambda \asymp \log^{1/\beta}(1/\lambda)$.

Now, recall that $\lambda_q \asymp \{t+\log^{1/\beta}(n)\}/n$, implying $\mathfrak{D}_{\lambda_q} \lesssim \log^{1/\beta}(n)$. Thus, Theorem 4.3.1 implies that, with probability at least $1-e^{-t}$,

$$\|\widehat{f}-f_0\|_2^2 + \lambda_q\|\widehat{f}-f_0\|_{\mathscr{H}}^2 \lesssim \frac{t+\log^{1/\beta}(n)}{n}.$$

Moreover, conditioned on the event where the above inequality holds, applying (D.29) gives

$$\|\widehat{f}-f_0\|_4^2 \lesssim C_\phi \mathfrak{D}_{\lambda_q}^{1/2}\frac{t+\log^{1/\beta}(n)}{n} \lesssim \log^{1/(2\beta)}(n)\cdot\frac{t+\log^{1/\beta}(n)}{n}.$$

Combining this with Theorem 4.3.2 gives

$$\begin{aligned}\tau\|\widehat{g}-g_0\|_2 &\lesssim \sigma_0\sqrt{\frac{\mathfrak{D}_{\lambda_q}+t}{n}}+\sqrt{\frac{t+\log^{1/\beta}(n)}{n}}+\log^{1/(2\beta)}(n)\cdot\frac{t+\log^{1/\beta}(n)}{n}\\ &\quad+C_\phi\sqrt{\frac{t+\log^{1/\beta}(n)}{n}}\log^{1/\beta}(1/\lambda_q)\sqrt{\frac{t+\log^{1/\beta}(1/\lambda_e)}{n}}\\ &\lesssim\sqrt{\frac{t+\log^{1/\beta}(n)}{n}},\end{aligned}$$

which completes the proof. □

D.4.9 Proof of Corollary D.2.8

Recall that $\lambda_q \asymp \lambda_e \asymp \log^{1/\beta}(n)/n$ and $0 < r_e \leq 1/2$. Since $\mathfrak{D}_\lambda \asymp \log^{1/\beta}(1/\lambda)$ for $\lambda \in (0,1)$, we have

$$\delta_n(\lambda_q,n,\log n)=\lambda_q^{r_q+1/2}\|f^*\|_{\mathscr{H}}+\sqrt{\frac{\mathfrak{D}_{\lambda_q}+\log n}{n}}\lesssim\sqrt{\frac{\log^{1/\beta}(n)+\log(n)}{n}},$$

and

$$\gamma_n(\lambda_e,n,\log n)=\tau\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}+\sigma_0\sqrt{\frac{\mathfrak{D}_{\lambda_e}+\log n}{n}}\lesssim\sqrt{\frac{\log^{1/\beta}(n)+\log(n)}{n}}.$$

Then, it follows that

$$\Delta_1(\lambda_e)=\gamma_n\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{\log n}{n}}=o(n^{-1/2}),$$

and

$$\Delta_2(\lambda_q,\lambda_e) = \delta_n\left\{\mathfrak{D}_{\lambda_q}^{1/2}\sqrt{\frac{\log n+\mathfrak{D}_{\lambda_e}}{n}}+\mathfrak{D}_{\lambda_q}^{1/2}\delta_n\right\}$$
$$\lesssim \delta_n\sqrt{\frac{\mathfrak{D}_{\lambda_q}\log(n)}{n}}+\delta_n\sqrt{\frac{\mathfrak{D}_{\lambda_q}\mathfrak{D}_{\lambda_e}}{n}}+\mathfrak{D}_{\lambda_q}^{1/2}\delta_n^2 = o(n^{-1/2}).$$

Combining the above bounds with Corollary 4.3.1 gives

$$\tau\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}(\widehat{g}-g_{\lambda_e})(x_0) \xrightarrow{\mathrm{d}} \mathscr{N}(0,\rho^2(x_0)). \tag{D.55}$$

For the bias term $(g_{\lambda_e}-g_0)(x_0)$, combining Lemma D.3.6 and (D.30) yields

$$|(g_{\lambda_e}-g_0)(x_0)| \le \|g_{\lambda_e}-g_0\|_\infty \le C_\phi\mathfrak{D}_{\lambda_e}^{1/2}\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}}.$$

Thus,

$$\sqrt{\frac{n}{\mathfrak{D}_{\lambda_e}}}|(g_{\lambda_e}-g_0)(x_0)| \lesssim \sqrt{n}\left\{\frac{\log^{1/\beta}(n)}{n}\right\}^{r_e+1/2} = o(1),$$

where the last equality is derived by the assumption $r_e > 0$. Combining this with (D.55) establishes the claim. □

D.4.10 Proof of Corollary D.2.9

The proof follows a similar line of argument as in the proof of Corollary D.2.3, and therefore is omitted. □

D.4.11 Proof of Lemma D.2.1

To begin with, recall the eigenfunctions (D.2) and eigenvalues (D.3). Then, for any $x_0 \in [0,1]$, we have

$$\begin{aligned}
\mathbb{E}\{\omega_i(T_K+\lambda I)^{-1}K_{X_i}(x_0)\}^2 &\geq \underline{\sigma}^2\mathbb{E}\{(T_K+\lambda I)^{-1}K_{X_i}(x_0)\}^2 \\
&\geq \underline{\sigma}^2\mathbb{E}\bigg\{\sum_{j=1}^{\infty}\frac{\mu_j}{\mu_j+\lambda}\phi_j(x_0)\phi_j(X_i)\bigg\}^2 \\
&= \underline{\sigma}^2\sum_{j=1}^{\infty}\bigg(\frac{\mu_j}{\mu_j+\lambda}\bigg)^2\phi_j^2(x_0) \\
&= 2\underline{\sigma}^2\sum_{k=1}^{\infty}\frac{1}{\{1+\lambda(2k\pi)^{2p}\}^2}. \qquad \text{(D.56)}
\end{aligned}$$

For a fixed $\lambda > 0$, $\{1+\lambda(2k\pi)^{2p}\}^2$ is decreasing with respect to k, so it follows that

$$\begin{aligned}
\sum_{k=1}^{\infty}\frac{1}{\{1+\lambda(2k\pi)^{2p}\}^2} \geq \int_1^{\infty}\frac{1}{\{1+\lambda(2x\pi)^{2p}\}^2}\mathrm{d}x &\geq \frac{\lambda^{-1/(2p)}}{2\pi}\int_1^{\infty}\frac{1}{(1+x^{2p})^2}\mathrm{d}x \\
&\geq \frac{\lambda^{-1/(2p)}}{8\pi}\int_1^{\infty}\frac{1}{x^{4p}}\mathrm{d}x \\
&= \frac{\lambda^{-1/(2p)}}{8\pi}\frac{1}{4p-1}, \qquad \text{(D.57)}
\end{aligned}$$

where the second inequality follows by a change of variables and the assumption $\lambda \leq (2\pi)^{-2p}$. On the other hand, following a similar argument as in (D.50) gives

$$\begin{aligned}
\mathfrak{D}_\lambda = \sum_{j=0}^{\infty}\frac{\mu_j}{\mu_j+\lambda} = \frac{1}{1+\lambda}+2\sum_{k=1}^{\infty}\frac{1}{1+\lambda(2k\pi)^{2p}} &\leq \frac{1}{1+\lambda}+2\int_0^{\infty}\frac{1}{1+\lambda(2x\pi)^{2p}}\mathrm{d}x \\
&= \frac{1}{1+\lambda}+\frac{\lambda^{-1/(2p)}}{\pi}\int_0^{\infty}\frac{1}{1+x^{2p}}\mathrm{d}x \\
&\leq \frac{1}{1+\lambda}+\frac{\lambda^{-1/(2p)}}{\pi}\frac{2p}{2p-1} \\
&\leq \frac{4p-1}{2p-1}\lambda^{-1/(2p)}.
\end{aligned}$$

Combining this with (D.56) and (D.57) yields

$$\rho^2_{\lambda_e}(x_0) = \frac{\mathbb{E}\{\omega_i(T_K+\lambda I)^{-1}K_{X_i}(x_0)\}^2}{\mathfrak{D}_\lambda} \geq \frac{2p-1}{4\pi(4p-1)^2}\cdot\underline{\sigma}^2.$$

This implies the claimed lower bound with $c_{11} = (2p-1)/\{4\pi(4p-1)^2\}$. □

D.4.12 Proof of Lemma D.2.2

By the eigensystem (D.4) and (D.5), it follows that for any $x_0 \in [-\pi,\pi]$,

$$\begin{aligned}\mathbb{E}\{\omega_i(T_K+\lambda I)^{-1}K_{X_i}(x_0)\}^2 &\geq \underline{\sigma}^2\mathbb{E}\{(T_K+\lambda I)^{-1}K_{X_i}(x_0)\}^2 \\ &\geq \underline{\sigma}^2\mathbb{E}\bigg\{\sum_{j=1}^{\infty}\frac{\mu_j}{\mu_j+\lambda}\phi_j(x_0)\phi_j(X_i)\bigg\}^2 \\ &= \frac{\underline{\sigma}^2}{\pi}\sum_{k=1}^{\infty}\frac{1}{(1+\lambda e^{k^2\theta^2/2})^2}. \qquad \text{(D.58)}\end{aligned}$$

Then, applying a similar argument as in (D.54) gives

$$\begin{aligned}\sum_{k=1}^{\infty}\frac{1}{(1+\lambda e^{k^2\theta^2/2})^2} \geq \int_1^{\infty}\frac{1}{(1+\lambda e^{x^2\theta^2/2})^2}\mathrm{d}x &\geq \int_1^{\sqrt{2}\theta^{-1}\log^{1/2}(1/\lambda)}\frac{1}{(1+\lambda e^{x^2\theta^2/2})^2}\mathrm{d}x \\ &\geq \frac{1}{4}\bigg\{\frac{\sqrt{2}}{\theta}\log^{1/2}(1/\lambda)-1\bigg\} \\ &\geq \frac{\sqrt{2}}{8\theta}\log^{1/2}(1/\lambda), \qquad \text{(D.59)}\end{aligned}$$

where the last inequality follows from the assumption $\log(1/\lambda) \geq 2\theta^2$. On the other hand, by following a similar argument as in (D.53), we have

$$\begin{aligned}
\mathfrak{D}_\lambda &= \sum_{j=0}^{\infty} \frac{\mu_j}{\mu_j+\lambda} = \frac{1}{1+\lambda} + 2\sum_{k=1}^{\infty} \frac{1}{1+\lambda e^{k^2\theta^2/2}} \leq \frac{1}{1+\lambda} + 2\int_0^{\infty} \frac{1}{1+\lambda e^{x^2\theta^2/2}} \mathrm{d}x \\
&= \frac{1}{1+\lambda} + 2\int_0^{\sqrt{2}\theta^{-1}\log^{1/2}(1/\lambda)} \frac{1}{1+\lambda e^{x^2\theta^2/2}} \mathrm{d}x + 2\int_{\sqrt{2}\theta^{-1}\log^{1/2}(1/\lambda)}^{\infty} \frac{1}{1+\lambda e^{x^2\theta^2/2}} \mathrm{d}x \\
&\leq \frac{1}{1+\lambda} + \frac{2\sqrt{2}}{\theta}\log^{1/2}(1/\lambda) + \frac{1}{\sqrt{2}\theta}\int_0^{\infty} \frac{1}{(1+e^x)\{x+\log(1/\lambda)\}^{1/2}} \mathrm{d}x \\
&\leq \frac{1}{1+\lambda} + \frac{2\sqrt{2}}{\theta}\log^{1/2}(1/\lambda) + \frac{\log^{-1/2}(1/\lambda)}{\sqrt{2}\theta}\int_0^{\infty} \frac{1}{1+e^x} \mathrm{d}x \\
&= \frac{1}{1+\lambda} + \frac{2\sqrt{2}}{\theta}\log^{1/2}(1/\lambda) + \frac{\log(2)\log^{-1/2}(1/\lambda)}{\sqrt{2}\theta} \\
&\leq \left(1 + \frac{2\sqrt{2}}{\theta} + \frac{\log 2}{\sqrt{2}\theta}\right)\log^{1/2}(1/\lambda),
\end{aligned}$$

where the last inequality follows from the assumption $\lambda < 1/e$. Combining this with (D.58) and (D.59) yields the claimed lower bound, where c_{12} only depends on θ.

□

D.5 Proof of Technical Lemmas

D.5.1 Proof of Lemma D.3.1

By Lemma S6 in the supplement of Padilla and Chatterjee (2022), we have

$$\begin{aligned}
&\mathbb{E}\big\{\rho_\tau(Y_i - f(X_i)) - \rho_\tau(Y_i - f_0(X_i))\big\} \\
&\qquad\qquad \geq \min\left(\frac{p}{2}, \frac{l_0 p}{4}\right)\mathbb{E}\big\{|f(X_i) - f_0(X_i)| \wedge |f(X_i) - f_0(X_i)|^2\big\}.
\end{aligned}$$

When $\|f - f_0\|_\infty \leq M$ with $M \geq 1$, we get

$$|f(X_i) - f_0(X_i)| \geq |f(X_i) - f_0(X_i)|^2/M \quad \text{almost surely},$$

which further implies

$$\mathbb{E}\{\rho_\tau(Y_i - f(X_i)) - \rho_\tau(Y_i - f_0(X_i))\} \geq \frac{1}{M} \underbrace{\min\left(\frac{p}{2}, \frac{l_0 p}{4}\right)}_{=:c_{13}} \|f - f_0\|_2^2,$$

as claimed. □

D.5.2 Proof of Lemma D.3.2

To begin with, for any $\delta_2, \delta_{\mathscr{H}} > 0$, define the following random fluctuation term:

$$\Omega(\delta_2, \delta_{\mathscr{H}}) := \sup_{f \in \mathscr{H}, \|f - f_0\|_2 \leq \delta_2, \|f - f_0\|_{\mathscr{H}} \leq \delta_{\mathscr{H}}} \left| \frac{1}{n} \sum_{i=1}^{n} (1 - \mathbb{E}) \{\rho_\tau(Y_i - f(X_i)) - \rho_\tau(Y_i - f_0(X_i))\} \right|.$$

Since $\rho_\tau(\cdot)$ is a Lipschitz function and $\|f - f_0\|_\infty \leq \|f - f_0\|_{\mathscr{H}}$ by Condition 4.2.1, we have

$$\begin{aligned} \mathrm{Var}\big(\rho_\tau(Y_i - f(X_i)) - \rho_\tau(Y_i - f_0(X_i))\big) &\leq \mathbb{E}\{\rho_\tau(Y_i - f(X_i)) - \rho_\tau(Y_i - f_0(X_i))\}^2 \\ &\leq \mathbb{E}\{f(X_i) - f_0(X_i)\}^2 \leq \delta_2^2, \end{aligned}$$

and

$$\big\|\rho_\tau(Y_i - f(X_i)) - \rho_\tau(Y_i - f_0(X_i))\big\|_\infty \leq \|f - f_0\|_\infty \leq \delta_{\mathscr{H}},$$

when $\|f - f_0\|_2 \leq \delta_2$ and $\|f - f_0\|_{\mathscr{H}} \leq \delta_{\mathscr{H}}$. By Theorem 7.3 in Bousquet (2003), for any $t > 0$, $\Omega(\delta_2, \delta_{\mathscr{H}})$ satisfies

$$\Omega(\delta_2, \delta_{\mathscr{H}}) \leq 2\mathbb{E}\Omega(\delta_2, \delta_{\mathscr{H}}) + \delta_2 \sqrt{\frac{2t}{n}} + \delta_{\mathscr{H}} \frac{8t}{3n} \tag{D.60}$$

with probability at least $1 - e^{-t}$.

For the expected value $\mathbb{E}\Omega(\delta_2, \delta_{\mathscr{H}})$, applying Rademacher symmetrization and Tala-

grand's contraction principle (see, e.g. Theorem 4.12 in Ledoux and Talagrand (1991)) yields

$$\begin{aligned}\mathbb{E}\Omega(\delta_2,\delta_{\mathscr{H}}) &\le 2\mathbb{E}\left[\sup_{f\in\mathscr{H},\|f-f_0\|_2\le\delta_2,\|f-f_0\|_{\mathscr{H}}\le\delta_{\mathscr{H}}}\left|\frac{1}{n}\sum_{i=1}^n e_i\{\rho_\tau(Y_i-f(X_i))-\rho_\tau(Y_i-f_0(X_i))\}\right|\right]\\ &\le 4\mathbb{E}\left[\sup_{f\in\mathscr{H},\|f-f_0\|_2\le\delta_2,\|f-f_0\|_{\mathscr{H}}\le\delta_{\mathscr{H}}}\left|\frac{1}{n}\sum_{i=1}^n e_i\{f(X_i)-f_0(X_i)\}\right|\right],\end{aligned}$$

where $e_1,e_2,\ldots,e_n$ are independent Rademacher random variables that are independent of $\{X_i\}_{i=1}^n$. Now, for any $\lambda>0$, $\|f-f_0\|_2\le\delta_2$ and $\|f-f_0\|_{\mathscr{H}}\le\delta_{\mathscr{H}}$ imply that $\|f-f_0\|_2^2+\lambda\|f-f_0\|_{\mathscr{H}}^2\le\delta^2:=\delta_2^2+\lambda\delta_{\mathscr{H}}^2$. Thus, the expected value is further bounded as

$$\mathbb{E}\Omega(\delta_2,\delta_{\mathscr{H}})\le 4\mathbb{E}\left\{\sup_{h\in\mathscr{H},\|h\|_2^2+\lambda\|h\|_{\mathscr{H}}^2\le\delta^2}\left|\frac{1}{n}\sum_{i=1}^n e_i h(X_i)\right|\right\}.$$

For any $h\in\mathscr{H}$ with the expansion $h(\cdot)=\sum_{j=1}^\infty\theta_j\phi_j(\cdot)$, we have $\|h\|_2^2=\sum_{j=1}^\infty\theta_j^2$ and $\|h\|_{\mathscr{H}}^2=\sum_{j=1}^\infty\theta_j^2/\mu_j$. This implies that if $h=\sum_{j=1}^\infty\theta_j\phi_j$ satisfies $\|h\|_2^2+\lambda\|h\|_{\mathscr{H}}^2\le\delta^2$, then $\{\theta_j\}_{j=1}^\infty$ satisfy $\sum_{j=1}^\infty\theta_j^2/\nu_j\le\delta^2$, where $\nu_j=\mu_j/(\mu_j+\lambda)$ for $j\ge1$. Combining this with the above inequality, we have

$$\begin{aligned}\mathbb{E}\Omega(\delta_2,\delta_{\mathscr{H}}) &\le 4\mathbb{E}\left\{\sup_{\{\theta_j\}_{j=1}^\infty:\sum_{j=1}^\infty\theta_j^2/\nu_j\le\delta^2}\left|\frac{1}{n}\sum_{i=1}^n e_i\sum_{j=1}^\infty\theta_j\phi_j(X_i)\right|\right\}\\ &= \frac{4}{n}\mathbb{E}\left\{\sup_{\{\theta_j\}_{j=1}^\infty:\sum_{j=1}^\infty\theta_j^2/\nu_j\le\delta^2}\left|\sum_{j=1}^\infty\frac{\theta_j}{\sqrt{\nu_j}}\sum_{i=1}^n e_i\sqrt{\nu_j}\phi_j(X_i)\right|\right\}\\ &\overset{\text{(i)}}{\le} \frac{4}{n}\left[\mathbb{E}\left\{\sup_{\{\theta_j\}_{j=1}^\infty:\sum_{j=1}^\infty\theta_j^2/\nu_j\le\delta^2}\left|\sum_{j=1}^\infty\frac{\theta_j}{\sqrt{\nu_j}}\sum_{i=1}^n e_i\sqrt{\nu_j}\phi_j(X_i)\right|^2\right\}\right]^{1/2}\\ &\overset{\text{(ii)}}{\le} \frac{4\delta}{n}\left[\mathbb{E}\sum_{j=1}^\infty\left\{\sum_{i=1}^n e_i\sqrt{\nu_j}\phi_j(X_i)\right\}^2\right]^{1/2}\\ &\overset{\text{(iii)}}{=} \frac{4\delta}{\sqrt{n}}\left(\sum_{j=1}^\infty\nu_j\right)^{1/2}=4\delta\sqrt{\frac{\mathfrak{D}_\lambda}{n}}.\end{aligned}\tag{D.61}$$

Here, step (i) is due to Jensen's inequality, step (ii) is obtained by the Cauchy-Schwarz inequality, and step (iii) follows from the orthonormality of $\{\phi_j\}_{j=1}^\infty$. Combining this with (D.60), we have

with probability at least $1 - e^{-t}$ that

$$\Omega(\delta_2, \delta_{\mathscr{H}}) \leq c_{14} \left\{ \delta \sqrt{\frac{\mathfrak{D}_\lambda}{n}} + \delta_2 \sqrt{\frac{t}{n}} + \delta_{\mathscr{H}} \frac{t}{n} \right\},$$

where $c_{14} = 8$. This completes the proof. □

D.5.3 Proof of Lemma D.3.3

We introduce the following notations that will be used frequently. Let $\ell_2 = \ell_2(\mathbb{N})$ be the Hilbert space of square-summable infinite sequences, that is,

$$\ell_2 = \left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)^{\mathrm{T}} \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^{\infty} \theta_j^2 < \infty \right\}.$$

Denote by $\|\cdot\|_{\ell_2} = \|\cdot\|_{\ell_2(\mathbb{N})}$ the ℓ_2-norm in the space $\ell_2(\mathbb{N})$ induced by the inner product $\langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \rangle_{\ell_2} = \langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \rangle_{\ell_2(\mathbb{N})} = \sum_{j=1}^{\infty} \theta_{1,j} \theta_{2,j}$ for $\boldsymbol{\theta}_1 = (\theta_{1,1}, \theta_{1,2}, \ldots)^{\mathrm{T}}$ and $\boldsymbol{\theta}_2 = (\theta_{2,1}, \theta_{2,2}, \ldots)^{T}$. Now, we define

$$\Phi(x) = (\sqrt{\mu_1}\phi_1(x), \sqrt{\mu_2}\phi_2(x), \ldots)^{\mathrm{T}} \quad \text{for} \quad x \in \mathscr{X},$$

where $\{\phi_j\}_{j \geq 1}$ is a sequence of orthonormal eigenfunctions of T_K with an associated set of non-negative eigenvalues $\{\mu_j\}_{j \geq 1}$. Recall the relationship (4.4) and Condition 4.2.1. Thus, $\|\Phi(x)\|_{\ell_2}^2 = \sum_{j=1}^{\infty} \mu_j \phi_j^2(x) = K(x,x) \leq 1$, which implies $\Phi(x) \in \ell_2$ for any $x \in \mathscr{X}$. For any $h \in \mathscr{H}$, we can express $h(\cdot) = \langle \boldsymbol{\theta}_h, \Phi(\cdot) \rangle_{\ell_2}$, where $\boldsymbol{\theta}_h = (\theta_{h,1}, \theta_{h,2}, \ldots)^{\mathrm{T}}$ with $\theta_{h,j} = \langle h, \phi_j \rangle_2 / \sqrt{\mu_j}$. Here, we adopt the convention $0/0 = 0$. With this representation, $\|h\|_{\mathscr{H}}^2 = \|\boldsymbol{\theta}_h\|_{\ell_2}^2$ and $\|h\|_2^2 = \sum_{j=1}^{\infty} \mu_j \theta_{h,j}^2$.

For any self-adjoint operator $\mathbf{A}$ on a separable Hilbert space, $\|\mathbf{A}\|_{\mathrm{op}}$ denotes the operator norm of $\mathbf{A}$. Moreover, given two self-adjoint opeartors $\mathbf{A}$ and $\mathbf{B}$, we write $\mathbf{A} \preceq \mathbf{B}$ ($\mathbf{A} \succeq \mathbf{B}$) if and only if $\mathbf{A} - \mathbf{B}$ is negative (positive) semidefinite. In the proof, we utilize the following Bernstein inequality for a sum of self-adjoint operators. The proof can be found in Section 3.2 of Minsker

(2017).

Lemma D.5.1 (Bernstein's inequality for bounded self-adjoint operators)**.** Let $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_n$ be n independent self-adjoint operators on a separable Hilbert space. Assume that for any $1 \le i \le n$, $\mathbb{E}\mathbf{M}_i = \mathbf{0}$ and $\|\mathbf{M}_i\|_{\mathrm{op}} \le B$ for some $B \ge 0$. Moreover, there exist positive trace-class operator $\mathbf{V}$ and $\sigma > 0$ such that $(1/n)\sum_{i=1}^n \mathbb{E}\mathbf{M}_i^2 \preceq \mathbf{V}$ with $\|\mathbf{V}\|_{\mathrm{op}} \le \sigma^2$. Then, for any $u \ge B/(3n) + \sqrt{\sigma^2/n}$,

$$\mathbb{P}\bigg(\bigg\|\frac{1}{n}\sum_{i=1}^n \mathbf{M}_i\bigg\|_{\mathrm{op}} \ge u\bigg) \le 14\frac{\mathrm{tr}(\mathbf{V})}{\sigma^2}\exp\bigg(-\frac{nu^2/2}{\sigma^2 + Bu/3}\bigg).$$

In particular, for any $t \ge 1/2$,

$$\mathbb{P}\bigg(\bigg\|\frac{1}{n}\sum_{i=1}^n \mathbf{M}_i\bigg\|_{\mathrm{op}} \ge \sigma\sqrt{\frac{2t}{n}} + \frac{2B}{3}\frac{t}{n}\bigg) \le 14\frac{\mathrm{tr}(\mathbf{V})}{\sigma^2}e^{-t}.$$

With these notations and Lemma D.5.1, we are ready to prove Lemma D.3.3.

Proof of Lemma D.3.3. Define

$$\Sigma = \mathbb{E}\big\{\Phi(X)\Phi(X)^{\mathrm{T}}\big\} \quad \text{and} \quad \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \Phi(X_i)\Phi(X_i)^{\mathrm{T}},$$

which are self-adjoint operators in ℓ_2. We remark that Σ is the infinite-dimensional diagonal matrix with entries $\mu_1 \ge \mu_2 \ge \cdots \ge 0$. For any $h \in \mathscr{H}$ with the expansion $h = \langle \boldsymbol{\theta}_h, \Phi\rangle_{\ell_2}$, we can write

$$\|h\|_2^2 = \langle \boldsymbol{\theta}_h, \Sigma\boldsymbol{\theta}_h\rangle_{\ell_2} \quad \text{and} \quad \|h\|_n^2 = \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{\theta}_h, \Phi(X_i)\rangle_{\ell_2}^2 = \langle \boldsymbol{\theta}_h, \widehat{\Sigma}\boldsymbol{\theta}_h\rangle_{\ell_2},$$

which further imply

$$\|h\|_2^2 + \lambda_e\|h\|_{\mathscr{H}}^2 = \langle \boldsymbol{\theta}_h, (\Sigma + \lambda_e I)\boldsymbol{\theta}_h\rangle_{\ell_2} \quad \text{and} \quad \|h\|_n^2 + \lambda_e\|h\|_{\mathscr{H}}^2 = \langle \boldsymbol{\theta}_h, (\widehat{\Sigma} + \lambda_e I)\boldsymbol{\theta}_h\rangle_{\ell_2}.$$

Thus, it suffices to prove that

$$\frac{1}{2}(\Sigma+\lambda_e I) \preceq \widehat{\Sigma}+\lambda_e I \preceq \frac{3}{2}(\Sigma+\lambda_e I).$$

To this end, for any $1 \le i \le n$, define

$$\mathbf{M}_i = (\Sigma+\lambda_e I)^{-1/2}\{\Sigma - \Phi(X_i)\Phi(X_i)^{\mathrm{T}}\}(\Sigma+\lambda_e I)^{-1/2},$$

which is well-defined for any $\lambda_e > 0$. Since $\Sigma = \mathbb{E}\{\Phi(X_i)\Phi(X_i)^{\mathrm{T}}\}$, $\mathbf{M}_i$ satisfies

$$\begin{aligned}\|\mathbf{M}_i\|_{\mathrm{op}} &\le 2\sup_{x\in\mathscr{X}}\left\|(\Sigma+\lambda_e I)^{-1/2}\Phi(x)\Phi(x)^{\mathrm{T}}(\Sigma+\lambda_e I)^{-1/2}\right\|_{\mathrm{op}} \\ &\le \frac{2}{\lambda_e}\sup_{x\in\mathscr{X}}\left\|\Phi(x)\Phi(x)^{\mathrm{T}}\right\|_{\mathrm{op}} = \frac{2}{\lambda_e}\sup_{x\in\mathscr{X}}\left\|\Phi(x)\right\|_{\ell_2}^2 \le \frac{2}{\lambda_e},\end{aligned}$$

where the second inequality follows from $\Sigma+\lambda_e I \succeq \lambda_e I$ and the last inequality is obtained by Condition 4.2.1. Moreover,

$$\begin{aligned}\mathbb{E}\mathbf{M}_i^2 &\preceq \mathbb{E}\left\{(\Sigma+\lambda_e I)^{-1/2}\Phi(X_i)\Phi(X_i)^{\mathrm{T}}(\Sigma+\lambda_e I)^{-1/2}\right\}^2 \\ &\preceq \frac{1}{\lambda_e}\left(\Sigma+\lambda_e I\right)^{-1/2}\mathbb{E}\left\{\Phi(X_i)\Phi(X_i)^{\mathrm{T}}\right\}\left(\Sigma+\lambda_e I\right)^{-1/2} \\ &= \frac{1}{\lambda_e}(\Sigma+\lambda_e I)^{-1/2}\Sigma(\Sigma+\lambda_e I)^{-1/2} =: \mathbf{V},\end{aligned}$$

where the second inequality follows from $\Sigma+\lambda_e I \succeq \lambda_e I$ and $\|\Phi(X_i)\|_{\ell_2}^2 \le 1$ by Condition 4.2.1. Remark that $\mathbf{V}$ is the infinite-dimensional diagonal matrix with entries $\mu_1/(\lambda_e\mu_1+\lambda_e^2) \ge \mu_2/(\lambda_e\mu_2+\lambda_e^2) \ge \cdots \ge 0$, so $\mathrm{tr}(\mathbf{V}) = \mathfrak{D}_{\lambda_e}/\lambda_e$ and $\|\mathbf{V}\|_{\mathrm{op}} \le \lambda_e^{-1}$. Therefore, Lemma D.5.1 implies that

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{M}_i\right\|_{\mathrm{op}} \ge \frac{1}{2}\right) \le 14\mathfrak{D}_{\lambda_e}\exp\left(-\frac{3n}{32}\lambda_e\right),$$

as long as $n\lambda_e \geq 9$. Note that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{M}_i = (\Sigma+\lambda_e I)^{-1/2}(\Sigma-\widehat{\Sigma})(\Sigma+\lambda_e I)^{-1/2}.$$

Consequently, the previous probability bound implies that the bounds

$$\Sigma-\widehat{\Sigma} \preceq \frac{1}{2}\Sigma+\frac{1}{2}\lambda_e I, \quad \text{and} \quad \widehat{\Sigma}-\Sigma \preceq \frac{1}{2}\Sigma+\frac{1}{2}\lambda_e I$$

hold with probability at least $1-14\mathfrak{D}_{\lambda_e}\exp(-3n\lambda_e/32)$. This is equivalent to

$$\mathbb{P}\left\{\frac{1}{2}(\Sigma+\lambda_e I) \preceq \widehat{\Sigma}+\lambda_e I \preceq \frac{3}{2}(\Sigma+\lambda_e I)\right\} \geq 1-14\mathfrak{D}_{\lambda_e}\exp\left(-\frac{3n}{32}\lambda_e\right),$$

which establishes the claim.

□

D.5.4 Proof of Lemma D.3.4

To begin with, note that

$$\sup_{h\in\mathscr{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\omega_i\frac{h(X_i)}{\|h\|_n+\lambda_e^{1/2}\|h\|_{\mathscr{H}}}\right| \leq \sup_{h\in\mathscr{H},\|h\|_n\leq 1,\|h\|_{\mathscr{H}}\leq\lambda_e^{-1/2}}\left|\frac{1}{n}\sum_{i=1}^{n}\omega_i h(X_i)\right|. \tag{D.62}$$

To establish a non-asymptotic bound for the right-hand side, we begin by fixing covariates $\{X_i\}_{i=1}^n$ and define $\mathbb{E}_X$ and $\mathbb{P}_X$ as the conditional expectation and conditional probability given $\{X_i\}_{i=1}^n$, respectively. Moreover, let $S_h = n^{-1/2}\sum_{i=1}^n \omega_i h(X_i)$ for any $h\in\mathscr{H}$ and recall that $\|\cdot\|_n$ denotes the empirical L_2-norm. By following a similar line of arguments as in (D.32), Condition 4.3.2 implies that for any $u>0$ and $h,h'\in\mathscr{H}$,

$$\mathbb{P}_X\Big(|S_h-S_{h'}|\geq u\Big) \leq 2\exp\left(-\frac{u^2}{2\sigma_0^2\|h-h'\|_n^2}\right).$$

Therefore, $(S_h)_{h\in\mathscr{H}}$ is a sub-Gaussian process with respect to the metric d_n, which is defined as $d_n(h,h') = \sigma_0\|h-h'\|_n$.

Denoting $\mathbb{T} = \mathbb{T}(\lambda_e) = \{h \in \mathscr{H} : \|h\|_n \le 1, \|h\|_{\mathscr{H}} \le \lambda_e^{-1/2}\}$, it is easy to see that the diameter of $\mathbb{T}$ with respect to d_n, that is, $\sup_{h,h'\in\mathbb{T}} d_n(h,h')$, is bounded by $2\sigma_0$. Define the γ_2-functional $\gamma_2(\mathbb{T},d_n)$ as

$$\gamma_2(\mathbb{T},d_n) := \inf_{(\mathbb{T}_k)_{k=0}^{\infty}:|\mathbb{T}_0|=1,|\mathbb{T}_k|\le 2^{2^k}} \sup_{h\in\mathbb{T}} \sum_{k=0}^{\infty} 2^{k/2} \inf_{h'\in\mathbb{T}_k} d_n(h,h'),$$

where $(\mathbb{T}_k)_{k=0}^{\infty}$ is a sequence of subsets of $\mathbb{T}$ and $|\mathbb{T}_k|$ denotes the cardinality of $\mathbb{T}_k$. Applying a conditional version of the generic chaining bound (e.g., Theorem 8.5.5 in Vershynin (2018)), there exists a universal constant $C_1 > 0$ satisfying

$$\sup_{h\in\mathbb{T}} |S_h| = \sup_{h\in\mathbb{T}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i h(X_i) \right| \le C_1 \left\{ \gamma_2(\mathbb{T},d_n) + \sigma_0\sqrt{t} \right\} \tag{D.63}$$

with $\mathbb{P}_X$-probability at least $1-2e^{-t}$. To bound $\gamma_2(\mathbb{T},d_n)$, let $\xi_1,\ldots,\xi_n$ be independent standard normal random variables, which are independent of $\{X_i\}_{i=1}^n$, and define $G_h = \sigma_0 \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i h(X_i)$ for $h \in \mathscr{H}$. Then, $(G_h)_{h\in\mathbb{T}}$ is a mean-zero Gaussian process on $\mathbb{T}$ with the metric d_n conditioned on $\{X_i\}_{i=1}^n$. By a conditional version of Talagrand's majorizing measure theorem (e.g., Theorem 8.6.1 in Vershynin (2018)), there exists a universal constant $C_2 > 0$ satisfying

$$\gamma_2(\mathbb{T},d_n) \le C_2 \mathbb{E}_X \left(\sup_{h\in\mathbb{T}} |G_h| \right) = C_2\sigma_0 \mathbb{E}_X \left\{ \sup_{h\in\mathbb{T}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i h(X_i) \right| \right\}.$$

Combining this bound with (D.63), we have

$$\mathbb{P}_X \left\{ \sup_{h\in\mathbb{T}} \left| \frac{1}{n} \sum_{i=1}^{n} \omega_i h(X_i) \right| \le C_3\sigma_0 \left[\mathbb{E}_X \left\{ \sup_{h\in\mathbb{T}} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i h(X_i) \right| \right\} + \sqrt{\frac{t}{n}} \right] \right\} \ge 1-2e^{-t}, \tag{D.64}$$

where $C_3 > 0$ is an absolute constant.

To establish a high probability bound for $\mathbb{E}_X\{\sup_{h\in\mathbb{T}} |(1/n)\sum_{i=1}^n \xi_i h(X_i)|\}$, which repre-

sents the empirical Gaussian complexity of the set $\mathbb{T}$, we utilize Cauchy-Schwarz and Hoffmann-Jørgensen inequalities to bound it by the empirical Rademacher complexity of $\mathbb{T}$. Subsequently, we apply Klein's version of Talagrand's inequality to establish a high probability bound for the empirical Rademacher complexity in terms of the population Rademacher complexity as in the proof of Proposition 5 in Koltchinskii and Yuan (2010).

To begin with, we first derive a bound for the empirical Gaussian complexity in terms of eigenvalues of the normalized kernel matrix. In detail, let $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots \geq \widehat{\mu}_n \geq 0$ be the eigenvalues of the normalized kernel matrix $\mathbf{K}/n$, where $\mathbf{K} = (K(X_i, X_j))_{1 \leq i,j \leq n}$. Following a similar argument as in the proof of Lemma 13.22 in Wainwright (2019), the empirical Gaussian complexity can be written as

$$\mathbb{E}_X \left\{ \sup_{h \in \mathbb{T}} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i h(X_i) \right| \right\} = \mathbb{E}_X \left\{ \sup_{\boldsymbol{\theta} \in \mathbb{T}'} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \theta_i \right| \right\}, \tag{D.65}$$

where the set $\mathbb{T}' \subset \mathbb{R}^n$ is defined as

$$\mathbb{T}' := \left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^{\mathrm{T}} \in \mathbb{R}^n : \|\boldsymbol{\theta}\|_2^2 \leq 1, \sum_{i=1}^{n} \frac{\theta_i^2}{\widehat{\mu}_i} \leq \lambda_e^{-1} \right\}. \tag{D.66}$$

Denote $\widehat{\nu}_i = \widehat{\mu}_i / (\widehat{\mu}_i + \lambda_e)$ for $1 \leq i \leq n$, and define the ellipsoid set $\mathbb{D}(\eta) \subset \mathbb{R}^n$ for any $\eta > 0$ as follows:

$$\mathbb{D}(\eta) := \left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^{\mathrm{T}} \in \mathbb{R}^n : \sum_{i=1}^{n} \frac{\theta_i^2}{\widehat{\nu}_i} \leq \eta \right\}. \tag{D.67}$$

It is easy to see that $\mathbb{T}' \subset \mathbb{D}(2)$, which implies

$$\mathbb{E}_X \left\{ \sup_{\boldsymbol{\theta} \in \mathbb{T}'} \left| \sum_{i=1}^{n} \xi_i \theta_i \right| \right\} \leq \mathbb{E}_X \left\{ \sup_{\boldsymbol{\theta} \in \mathbb{D}(2)} \left| \sum_{i=1}^{n} \xi_i \theta_i \right| \right\}.$$

Then, by applying Jensen's inequality and the Cauchy-Schwarz inequality, we obtain

$$\left\{\mathbb{E}_X\left(\sup_{\boldsymbol{\theta}\in\mathbb{T}'}\left|\sum_{i=1}^n \xi_i\theta_i\right|\right)\right\}^2 \le \mathbb{E}_X\left\{\sup_{\boldsymbol{\theta}\in\mathbb{D}(2)}\left|\sum_{i=1}^n \xi_i\theta_i\right|^2\right\} = \mathbb{E}_X\left\{\sup_{\boldsymbol{\theta}\in\mathbb{D}(2)}\left|\sum_{i=1}^n \sqrt{\widehat{\nu}_i}\xi_i\frac{\theta_i}{\sqrt{\widehat{\nu}_i}}\right|^2\right\}$$
$$= 2\mathbb{E}_X\left(\sum_{i=1}^n \widehat{\nu}_i\xi_i^2\right) = 2\sum_{i=1}^n \widehat{\nu}_i, \quad \text{(D.68)}$$

which, combined with (D.65), further implies

$$\mathbb{E}_X\left\{\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i h(X_i)\right|\right\} \le \sqrt{\frac{2}{n}}\left(\sum_{i=1}^n \widehat{\nu}_i\right)^{1/2}. \quad \text{(D.69)}$$

Now, let $e_1, e_2, \ldots, e_n$ be independent Rademacher variables that are independent of $\{X_i\}_{i=1}^n$. Following a similar argument as in (D.65) yields

$$\mathbb{E}_X\left\{\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^n e_i h(X_i)\right|\right\} = \mathbb{E}_X\left(\sup_{\boldsymbol{\theta}\in\mathbb{T}'}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\theta_i\right|\right).$$

By the definitions of $\mathbb{T}'$ in (D.66) and $\mathbb{D}(\eta)$ in (D.67), it is obvious that $\mathbb{D}(1) \subset \mathbb{T}'$, which implies

$$\mathbb{E}_X\left\{\sup_{\boldsymbol{\theta}\in\mathbb{T}'}\left|\sum_{i=1}^n e_i\theta_i\right|\right\} \ge \mathbb{E}_X\left\{\sup_{\boldsymbol{\theta}\in\mathbb{D}(1)}\left|\sum_{i=1}^n e_i\theta_i\right|\right\}.$$

Remark that $|e_i\theta_i| \le 1$ for any $\boldsymbol{\theta} \in \mathbb{D}(1)$ and $1 \le i \le n$. Hence, applying Hoffmann-Jørgensen inequality (see, e.g. Theorem 6.20 in Ledoux and Talagrand (1991)) gives

$$\mathbb{E}_X\left\{\sup_{\boldsymbol{\theta}\in\mathbb{D}(1)}\left|\sum_{i=1}^n e_i\theta_i\right|\right\} \gtrsim \left[\mathbb{E}_X\left\{\sup_{\boldsymbol{\theta}\in\mathbb{D}(1)}\left|\sum_{i=1}^n e_i\theta_i\right|^2\right\}\right]^{1/2} - 1.$$

Note that $\mathbb{E}_X\{\sup_{\boldsymbol{\theta}\in\mathbb{D}(1)}|\sum_{i=1}^n e_i\theta_i|^2\} = \sum_{i=1}^n \widehat{\nu}_i$ by employing a similar argument as in (D.68). Putting the pieces together, we obtain

$$\mathbb{E}_X\left\{\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^n e_i h(X_i)\right|\right\} \ge \mathbb{E}_X\left\{\sup_{\boldsymbol{\theta}\in\mathbb{D}(1)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\theta_i\right|\right\} \gtrsim \frac{1}{\sqrt{n}}\left(\sum_{i=1}^n \widehat{\nu}_i\right)^{1/2} - \frac{1}{\sqrt{n}},$$

which, combined with (D.69), further implies

$$\mathbb{E}_X\left\{\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i h(X_i)\right|\right\}\leq C_4\left[\mathbb{E}_X\left\{\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^{n}e_i h(X_i)\right|\right\}+\frac{1}{\sqrt{n}}\right], \tag{D.70}$$

where $C_4>0$ is a universal constant.

Turning to establish a high-probability bound for the empirical Rademacher complexity of $\mathbb{T}$ in (D.70), note that for any $h\in\mathbb{T}$ and $1\leq i\leq n$, we have $\|e_i h(X_i)\|_\infty\leq\lambda_e^{-1/2}$ and $(1/n)\sum_{i=1}^{n}\mathbb{E}_X\{e_i^2h^2(X_i)\}=\|h\|_n^2\leq 1$. Thus, applying a conditional version of Theorem 3.3.10 in Giné and Nickl (2016) yields that the bound

$$\begin{aligned}\mathbb{E}_X\left\{\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^{n}e_i h(X_i)\right|\right\}&\lesssim\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^{n}e_i h(X_i)\right|+\sqrt{\frac{t}{n}}+\lambda_e^{-1/2}\frac{t}{n}\\&\lesssim\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^{n}e_i h(X_i)\right|+\sqrt{\frac{t}{n}}\end{aligned}$$

holds with $\mathbb{P}_X$-probability at least $1-e^{-t}$, where the last inequality follows from the fact $\lambda_e\gtrsim t/n$. Combining this with (D.64) and (D.70) and taking expectation over $\{X_i\}_{i=1}^n$, there exists an absolute constant $C_5>0$ satisfying

$$\mathbb{P}\left\{\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^{n}\omega_i h(X_i)\right|\geq C_5\sigma_0\left[\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^{n}e_i h(X_i)\right|+\sqrt{\frac{t+1}{n}}\right]\right\}\leq 3e^{-t}. \tag{D.71}$$

Now, recall the definition of the event $\mathscr{S}=\mathscr{S}(\lambda_e)$ in Lemma D.3.3. Conditioned on $\mathscr{S}$, $h\in\mathbb{T}$ satisfies $\|h\|_2\leq\sqrt{3}$. Thus, conditioned on $\mathscr{S}$,

$$\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^{n}e_i h(X_i)\right|\leq\sup_{\|h\|_2\leq\sqrt{3},\,\|h\|_{\mathscr{H}}\leq\lambda_e^{-1/2}}\left|\frac{1}{n}\sum_{i=1}^{n}e_i h(X_i)\right|.$$

Remark that for any $1\leq i\leq n$ and $h\in\mathscr{H}$ with $\|h\|_2\leq\sqrt{3}$ and $\|h\|_{\mathscr{H}}\leq\lambda_e^{-1/2}$, we have $\|e_i h(X_i)\|_\infty\leq\lambda_e^{-1/2}$ and $\mathbb{E}\{e_i^2h^2(X_i)\}\leq 3$. Therefore, by applying Theorem 7.3 in Bousquet

(2003) and recalling the fact $\lambda_e \gtrsim t/n$, it follows that, with probability at least $1-e^{-t}$,

$$\sup_{\|h\|_2\le\sqrt{3},\|h\|_{\mathscr{H}}\le\lambda_e^{-1/2}}\left|\frac{1}{n}\sum_{i=1}^n e_i h(X_i)\right| \lesssim \mathbb{E}\left\{\sup_{\|h\|_2\le\sqrt{3},\|h\|_{\mathscr{H}}\le\lambda_e^{-1/2}}\left|\frac{1}{n}\sum_{i=1}^n e_i h(X_i)\right|\right\}+\sqrt{\frac{t}{n}}. \quad \text{(D.72)}$$

We next derive a bound for the expectation in the above inequality. If we write $h(\cdot)=\sum_{j=1}^\infty \theta_j\phi_j(\cdot)$ for any function $h\in\mathscr{H}$ with $\|h\|_2\le\sqrt{3}$ and $\|h\|_{\mathscr{H}}\le\lambda_e^{-1/2}$, $\{\theta_j\}_{j=1}^\infty$ satisfy $\sum_{j=1}^\infty \theta_j^2/\nu_{j,e}\le 4$, where $\nu_{j,e}=\mu_j/(\mu_j+\lambda_e)$. Thus, we have

$$\mathbb{E}\left\{\sup_{\|h\|_2\le\sqrt{3},\|h\|_{\mathscr{H}}\le\lambda_e^{-1/2}}\left|\frac{1}{n}\sum_{i=1}^n e_i h(X_i)\right|\right\} \le \mathbb{E}\left\{\sup_{\{\theta_j\}_{j=1}^\infty:\sum_{j=1}^\infty\theta_j^2/\nu_{j,e}\le 4}\left|\frac{1}{n}\sum_{i=1}^n e_i\sum_{j=1}^\infty \theta_j\phi_j(X_i)\right|\right\}.$$

Then, following a similar argument as in (D.61) gives

$$\mathbb{E}\left\{\sup_{\|h\|_2\le 2,\|h\|_{\mathscr{H}}\le\lambda_e^{-1/2}}\left|\frac{1}{n}\sum_{i=1}^n e_i h(X_i)\right|\right\} \le \frac{2}{\sqrt{n}}\left(\sum_{j=1}^\infty \nu_{j,e}\right)^{1/2} = 2\sqrt{\frac{\mathfrak{D}_{\lambda_e}}{n}}.$$

Putting the pieces together, (D.71), (D.72) and the above inequality yield

$$\mathbb{P}\left[\left\{\sup_{h\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^n \omega_i h(X_i)\right| \ge C_6\sigma_0\sqrt{\frac{\mathfrak{D}_{\lambda_e}+t}{n}}\right\}\bigcap\mathscr{S}\right] \le 4e^{-t}, \quad \text{(D.73)}$$

where C_6 is a universal positive constant. Combining this bound with (D.62), we have

$$\mathbb{P}\left[\left\{\sup_{h\in\mathscr{H}}\left|\frac{1}{n}\sum_{i=1}^n \omega_i h(X_i)\right| \ge \frac{1}{4}(\|h\|_n^2+\lambda_e\|h\|_{\mathscr{H}}^2)+2C_6^2\sigma_0^2\frac{\mathfrak{D}_{\lambda_e}+t}{n}\right\}\bigcap\mathscr{S}\right] \le 4e^{-t}.$$

This proves the claimed bound with $c_{15}=2C_6^2$. □

D.5.5 Proof of Lemma D.3.5

For any real-valued functions f, g on $\mathscr{X}$, note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\{Z_i(f)-Z_i(f_0)\}\{g(X_i)-g_0(X_i)\} \\
&=\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\{Z_i(f)-Z_i(f_0)\}\{g(X_i)-g_0(X_i)\}+\mathbb{E}\{Z_i(f)-Z_i(f_0)\}\{g(X_i)-g_0(X_i)\}.
\end{aligned}$$

STEP I. BOUND FOR THE EXPECTED VALUE TERM. We first establish a bound for the expectation $\mathbb{E}\{Z_i(f)-Z_i(f_0)\}\{g(X_i)-g_0(X_i)\}$. Let $\mathbb{E}_{X_i}$ be the conditional expectation given X_i and denoting $\Delta_f(X_i)=f_0(X_i)-f(X_i)$. Recall that $\varepsilon_i=Y_i-f_0(X_i)$ and $p_{\varepsilon_i|X_i}$ is the conditional density function of ε_i given X_i. By the definition of $Z_i(f)$, we have

$$\begin{aligned}
&Z_i(f)-Z_i(f_0) \\
&=\{Y_i-f(X_i)\}\mathbb{1}\{Y_i\le f(X_i)\}-\{Y_i-f_0(X_i)\}\mathbb{1}\{Y_i\le f_0(X_i)\}+\tau\{f(X_i)-f_0(X_i)\}, \qquad \text{(D.74)}
\end{aligned}$$

which implies

$$\begin{aligned}
&\mathbb{E}_{X_i}\{Z_i(f)-Z_i(f_0)\} \\
&=\mathbb{E}_{X_i}\big[\{\varepsilon_i+\Delta_f(X_i)\}\mathbb{1}\{\varepsilon_i\le-\Delta_f(X_i)\}-\varepsilon_i\mathbb{1}(\varepsilon_i\le 0)-\tau\Delta_f(X_i)\big] \\
&=\int_{-\infty}^{-\Delta_f(X_i)}\{u+\Delta_f(X_i)\}p_{\varepsilon_i|X_i}(u)\mathrm{d}u-\int_{-\infty}^{0}up_{\varepsilon_i|X_i}(u)\mathrm{d}u-\tau\Delta_f(X_i) \\
&=\int_{0}^{-\Delta_f(X_i)}up_{\varepsilon_i|X_i}(u)\mathrm{d}u+\Delta_f(X_i)\int_{-\infty}^{-\Delta_f(X_i)}p_{\varepsilon_i|X_i}(u)\mathrm{d}u-\Delta_f(X_i)\int_{-\infty}^{0}p_{\varepsilon_i|X_i}(u)\mathrm{d}u \\
&=\int_{0}^{-\Delta_f(X_i)}\{u+\Delta_f(X_i)\}p_{\varepsilon_i|X_i}(u)\mathrm{d}u,
\end{aligned}$$

where the third equality follows from the model assumption $\mathbb{P}(\varepsilon_i\le 0|X_i)=\tau$. By Condition

4.3.2 that $\sup_{u\in\mathbb{R}} p_{\varepsilon_i|X_i}(u) \le \bar{p}$ for some constant $\bar{p} > 0$, it holds

$$\left|\mathbb{E}_{X_i}\{Z_i(f) - Z_i(f_0)\}\right| \le \bar{p}\left|\int_0^{-\Delta_f(X_i)} \left\{u + \Delta_f(X_i)\right\} \mathrm{d}u\right| \le \frac{\bar{p}}{2}\left\{f(X_i) - f_0(X_i)\right\}^2.$$

Consequently,

$$\begin{aligned}
\left|\mathbb{E}\{Z_i(f) - Z_i(f_0)\}\{g_0(X_i) - g(X_i)\}\right| &\le \mathbb{E}\left[\left|\mathbb{E}_{X_i}\{Z_i(f) - Z_i(f_0)\}\right| \cdot \left|g(X_i) - g_0(X_i)\right|\right] \\
&\le \frac{\bar{p}}{2}\mathbb{E}\left[\{f(X_i) - f_0(X_i)\}^2 \left|g(X_i) - g_0(X_i)\right|\right] \\
&\le \frac{\bar{p}}{2}\|f - f_0\|_4^2 \|g - g_0\|_2, \qquad \text{(D.75)}
\end{aligned}$$

where the last line follows from Hölder's inequality.

STEP II. HIGH PROBABILITY BOUND FOR THE RANDOM FLUCTUATION TERM. In this step, we establish a high probability bound for the random fluctuation term by applying Talagrand's inequality. Denote $\Delta_f = f_0 - f$ for any $f \in \mathscr{H}$ and

$$\mathscr{F}(\delta_2) = \mathscr{F}(\delta_2, \lambda_q) = \{h \in \mathscr{H} : \|h\|_2^2 + \lambda_q \|h\|_{\mathscr{H}}^2 \le \delta_2^2\}.$$

It is evident that

$$\begin{aligned}
&\sup_{\Delta_f \in \mathscr{F}(\delta_2), g \in \mathscr{H}} \left| \frac{1}{n}\sum_{i=1}^n (1 - \mathbb{E}) \frac{\{Z_i(f) - Z_i(f_0)\}\{g(X_i) - g_0(X_i)\}}{\|g - g_0\|_2 + \lambda_e^{1/2}\|g - g_0\|_{\mathscr{H}}} \right| \\
&\qquad \le \underbrace{\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \sup_{\|h\|_{\mathscr{H}} \le \lambda_e^{-1/2}, \|h\|_2 \le 1} \left| \frac{1}{n}\sum_{i=1}^n (1 - \mathbb{E})\left\{Z_i(f) - Z_i(f_0)\right\} h(X_i) \right|}_{=:\Omega}. \qquad \text{(D.76)}
\end{aligned}$$

We first establish a relationship between $Z_i(f) - Z_i(f_0)$ and $f - f_0$. Suppose $\Delta_f(X_i) = f_0(X_i) -$

$f(X_i) \leq 0$. By (D.74), we have

$$
\begin{aligned}
&\big|Z_i(f) - Z_i(f_0)\big| \\
&= \big|\{Y_i - f(X_i)\}\mathbb{1}\{Y_i \leq f(X_i)\} - \{Y_i - f_0(X_i)\}\mathbb{1}\{Y_i \leq f_0(X_i)\} + \tau\{f(X_i) - f_0(X_i)\}\big| \\
&\leq \big|\Delta_f(X_i)\mathbb{1}\{Y_i \leq f_0(X_i)\} + \{Y_i - f(X_i)\}\mathbb{1}\{f_0(X_i) < Y_i \leq f(X_i)\} - \tau\Delta_f(X_i)\big| \\
&\leq \max(\tau, 1-\tau)\big|\Delta_f(X_i)\big| \leq \big|\Delta_f(X_i)\big|, \qquad \text{(D.77)}
\end{aligned}
$$

where the first inequality follows from the assumption $\Delta_f(X_i) \leq 0$, and the second inequality is obtained by the following inequality

$$
f_0(X_i) - f(X_i) \leq \{Y_i - f(X_i)\}\mathbb{1}\{f_0(X_i) \leq Y_i \leq f(X_i)\} \leq 0.
$$

By exchanging the roles of f and f_0 in (D.77), the same inequality holds when $\Delta_f(X_i) > 0$, leading to

$$
\big|Z_i(f) - Z_i(f_0)\big| \leq \big|f(X_i) - f_0(X_i)\big|. \qquad \text{(D.78)}
$$

Now, we claim for any $h \in \mathscr{H}$ and $\lambda > 0$ that

$$
\|h\|_\infty^2 \leq C_\phi^2 \mathfrak{D}_\lambda (\|h\|_2^2 + \lambda \|h\|_{\mathscr{H}}^2). \qquad \text{(D.79)}
$$

Write $h(\cdot) = \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot)$ and $\nu_j = \mu_j/(\mu_j + \lambda)$ for $j \geq 1$, we have

$$\begin{aligned}
\sup_{x \in \mathscr{X}} h^2(x) &= \sup_{x \in \mathscr{X}} \left\{ \sum_{j=1}^{\infty} \theta_j \phi_j(x) \right\}^2 \\
&= \sup_{x \in \mathscr{X}} \left\{ \sum_{j=1}^{\infty} \frac{\theta_j}{\sqrt{\nu_j}} \sqrt{\nu_j} \phi_j(x) \right\}^2 \\
&\leq \sup_{x \in \mathscr{X}} \sum_{j=1}^{\infty} \nu_j \phi_j^2(x) \cdot \sum_{j=1}^{\infty} \frac{\theta_j^2}{\nu_j} \\
&\leq C_\phi^2 \mathfrak{D}_\lambda \sum_{j=1}^{\infty} \frac{\theta_j^2}{\nu_j} = C_\phi^2 \mathfrak{D}_\lambda (\|h\|_2^2 + \lambda \|h\|_{\mathscr{H}}^2),
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality and the second inequality follows from Condition 4.3.3. This verifies (D.79). Thus, for any $\Delta_f \in \mathscr{F}(\delta_2)$ and $h \in \mathscr{H}$ with $\|h\|_{\mathscr{H}} \leq \lambda_e^{-1/2}$ and $\|h\|_2 \leq 1$, combining (D.78) and (D.79) with $\lambda = \lambda_q$ yields

$$\begin{aligned}
\big\| \{Z_i(f) - Z_i(f_0)\} h(X_i) \big\|_\infty &\leq \big\| Z_i(f) - Z_i(f_0) \big\|_\infty \cdot \big\| h(X_i) \big\|_\infty \\
&\leq \big\| f(X_i) - f_0(X_i) \big\|_\infty \cdot \big\| h(X_i) \big\|_\infty \leq C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \delta_2 \lambda_e^{-1/2}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathrm{Var}\big(\{Z_i(f) - Z_i(f_0)\} h(X_i) \big) &\leq \mathbb{E}\big[\{Z_i(f) - Z_i(f_0)\} h(X_i) \big]^2 \\
&\leq \|Z_i(f) - Z_i(f_0)\|_\infty^2 \mathbb{E}\{h^2(X_i)\} \\
&\leq C_\phi^2 \mathfrak{D}_{\lambda_q} \delta_2^2 \, \mathbb{E}\{h^2(X_i)\} \leq C_\phi^2 \mathfrak{D}_{\lambda_q} \delta_2^2.
\end{aligned}$$

By applying Theorem 7.3 in Bousquet (2003), the bound

$$\Omega \leq 2\mathbb{E}\Omega + C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \delta_2 \sqrt{\frac{2t}{n}} + C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \delta_2 \lambda_e^{-1/2} \frac{8t}{3n}$$

holds with probability at least $1-e^{-t}$ for any $t>0$. Since $\lambda_e \gtrsim t/n$, it follows that

$$\Omega \lesssim \mathbb{E}\Omega + C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \delta_2 \sqrt{\frac{t}{n}} \tag{D.80}$$

with the same probability.

STEP III. BOUND FOR THE EXPECTED VALUE OF THE RANDOM FLUCTUATION TERM. We next establish an upper bound for $\mathbb{E}\Omega$. By employing Rademacher symmetrization, we have

$$\mathbb{E}\Omega \le 2\mathbb{E}\left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \sup_{\|h\|_{\mathscr{H}} \le \lambda_e^{-1/2}, \|h\|_2 \le 1} \left| \frac{1}{n} \sum_{i=1}^n e_i \{ Z_i(f) - Z_i(f_0) \} h(X_i) \right| \right],$$

where $e_1, e_2, \ldots, e_n$ are independent Rademacher random variables. Now, we write $h(\cdot) = \sum_{j=1}^\infty h_j \phi_j(\cdot)$. Then, $\|h\|_{\mathscr{H}} \le \lambda_e^{-1}$ and $\|h\|_2 \le 1$ imply $\sum_{j\ge 1}(1+\lambda_e/\mu_j)h_j^2 = \sum_{j\ge 1} h_j^2/\nu_{j,e} \le 2$, where $\nu_{j,e} = \mu_j/(\mu_j+\lambda_e)$ for $j \ge 1$. Consequently, we have

$$\mathbb{E}\Omega \le 2\mathbb{E}\left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \sup_{\sum_{j\ge 1} h_j^2/\nu_{j,e} \le 2} \left| \frac{1}{n} \sum_{i=1}^n e_i \{ Z_i(f) - Z_i(f_0) \} \sum_{j=1}^\infty h_j \phi_j(X_i) \right| \right].$$

Note that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \sup_{\sum_{j\ge 1} h_j^2/\nu_{j,e} \le 2} \left| \sum_{i=1}^n e_i \{ Z_i(f) - Z_i(f_0) \} \sum_{j=1}^\infty h_j \phi_j(X_i) \right|^2 \right] \\
&= \mathbb{E}\left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \sup_{\sum_{j\ge 1} h_j^2/\nu_{j,e} \le 2} \left| \sum_{j=1}^\infty \frac{h_j}{\sqrt{\nu_{j,e}}} \sum_{i=1}^n e_i \{ Z_i(f) - Z_i(f_0) \} \sqrt{\nu_{j,e}} \phi_j(X_i) \right|^2 \right] \\
&\le 2\mathbb{E}\left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \sum_{j=1}^\infty \left| \sum_{i=1}^n e_i \{ Z_i(f) - Z_i(f_0) \} \sqrt{\nu_{j,e}} \phi_j(X_i) \right|^2 \right] \\
&\le 2\sum_{j=1}^\infty \nu_{j,e} \mathbb{E}\left\{ \left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \left| \sum_{i=1}^n e_i \{ Z_i(f) - Z_i(f_0) \} \phi_j(X_i) \right| \right]^2 \right\} \\
&= 2C_\phi^2 \sum_{j=1}^\infty \nu_{j,e} \mathbb{E}\left\{ \left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \left| \sum_{i=1}^n e_i \{ Z_i(f) - Z_i(f_0) \} \frac{\phi_j(X_i)}{C_\phi} \right| \right]^2 \right\},
\end{aligned}$$

where the first inequality comes from the Cauchy-Schwartz inequality, and the second inequality follows by the triangular inequality. Now, we fix $j \geq 1$. To employ Theorem 4.12 in Ledoux and Talagrand (1991), we take $F(u) = u^2$ for $u \geq 0$, which is convex and increasing for $u \geq 0$. Moreover, let $\mathbb{T}$ be a subset of $\mathbb{R}^n$ such that

$$\mathbb{T} := \big\{ \boldsymbol{t} = (t_1, t_2, \ldots, t_n) \in \mathbb{R}^n : t_i = f(X_i) - f_0(X_i), \;\; 1 \leq i \leq n \big\},$$

and define $\varphi_i : \mathbb{R} \to \mathbb{R}$ as

$$\varphi_i(a) = \frac{\phi_j(X_i)}{C_\phi} \big\{ (\varepsilon_i - a) \mathbb{1}(\varepsilon_i \leq a) + \tau f_0(X_i) + \tau a - Z_i(f_0) \big\},$$

so that $\varphi_i(f(X_i) - f_0(X_i)) = \{Z_i(f) - Z_i(f_0)\}\phi_j(X_i)/C_\phi$. Moreover, note that φ_i is a contraction for $1 \leq i \leq n$ by Condition 4.3.3 and (D.78). Then, by Theorem 4.12 in Ledoux and Talagrand (1991), we have

$$\begin{aligned}
\mathbb{E}\left\{ \left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \left| \sum_{i=1}^n e_i \big\{ Z_i(f) - Z_i(f_0) \big\} \frac{\phi_j(X_i)}{C_\phi} \right| \right]^2 \right\} &= \mathbb{E} F \left(\sup_{\boldsymbol{t} \in \mathbb{T}} \left| \sum_{i=1}^n e_i \varphi_i(t_i) \right| \right) \\
&\leq 4 \mathbb{E} F \left(\sup_{\boldsymbol{t} \in \mathbb{T}} \left| \sum_{i=1}^n e_i t_i \right| \right) \\
&= 4 \mathbb{E} \left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \left| \sum_{i=1}^n e_i \big\{ f(X_i) - f_0(X_i) \big\} \right|^2 \right].
\end{aligned}$$

After writing $f - f_0 = \sum_{k \geq 1} f_k \phi_k$, it follows that

$$\mathbb{E} \left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \left| \sum_{i=1}^n e_i \big\{ f(X_i) - f_0(X_i) \big\} \right|^2 \right] = \mathbb{E} \left[\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \left| \sum_{i=1}^n e_i \sum_{k=1}^\infty f_k \phi_k(X_i) \right|^2 \right].$$

Since $\Delta_f \in \mathscr{F}(\delta_2)$ is equivalent to $\sum_{k=1}^\infty f_k^2 / \nu_{k,q} \leq \delta_2^2$ with $\nu_{k,q} = \mu_k / (\mu_k + \lambda_q)$ for $k \geq 1$, we

have

$$\mathbb{E}\Big[\sup_{\Delta_f\in\mathscr{F}(\delta_2)}\Big|\sum_{i=1}^n e_i\sum_{k=1}^\infty f_k\phi_k(X_i)\Big|^2\Big] = \mathbb{E}\Big\{\sup_{\Delta_f\in\mathscr{F}(\delta_2)}\Big|\sum_{k=1}^\infty \frac{f_k}{\sqrt{\nu_{k,q}}}\sum_{i=1}^n e_i\sqrt{\nu_{k,q}}\phi_k(X_i)\Big|^2\Big\}$$
$$\le \delta_2^2\cdot\sum_{k=1}^\infty \mathbb{E}\Big\{\sum_{i=1}^n e_i\sqrt{\nu_{k,e}}\phi_k(X_i)\Big\}^2$$
$$= \delta_2^2\sum_{k=1}^\infty \mathbb{E}\Big\{\sum_{i=1}^n \nu_{k,q}\phi_k^2(X_i)\Big\} = n\delta_2^2\mathfrak{D}_{\lambda_q},$$

where the first inequality follows from the Cauchy-Schwartz inequality. Putting the pieces together and applying Jensen's inequality, we obtain

$$\mathbb{E}\Bigg[\sup_{\Delta_f\in\mathscr{F}(\delta_2)}\sup_{\sum_{j\ge1}h_j^2/\nu_{j,e}\le2}\Big|\frac{1}{n}\sum_{i=1}^n e_i\big\{Z_i(f)-Z_i(f_0)\big\}\sum_{j=1}^\infty h_j\phi_j(X_i)\Big|\Bigg]$$
$$\le C_\phi\sqrt{\frac{8}{n}}\delta_2\sqrt{\mathfrak{D}_{\lambda_q}}\Big(\sum_{j=1}^\infty \nu_{j,e}\Big)^{1/2} = C_\phi\delta_2\sqrt{\frac{8}{n}}\sqrt{\mathfrak{D}_{\lambda_q}\mathfrak{D}_{\lambda_e}},$$

which further implies

$$\mathbb{E}\Omega \le C_\phi\delta_2\sqrt{\frac{32\mathfrak{D}_{\lambda_q}\mathfrak{D}_{\lambda_e}}{n}}. \tag{D.81}$$

STEP IV. CONCLUSION OF THE PROOF. Putting the pieces together, (D.76), (D.80) and (D.81) yield

$$\sup_{\Delta_f\in\mathscr{F}(\delta_2),g\in\mathscr{H}}\Big|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\frac{\{Z_i(f)-Z_i(f_0)\}\{g(X_i)-g_0(X_i)\}}{\|g-g_0\|_2+\lambda_e^{1/2}\|g-g_0\|_{\mathscr{H}}}\Big| \le \Omega \lesssim C_\phi\delta_2\sqrt{\frac{\mathfrak{D}_{\lambda_q}(t+\mathfrak{D}_{\lambda_e})}{n}} \tag{D.82}$$

with probability at least $1-e^{-t}$. This implies

$$\sup_{\Delta_f \in \mathscr{F}(\delta_2), g \in \mathscr{H}} \left| \frac{\tau}{n} \sum_{i=1}^{n} (1-\mathbb{E})\{Z_i(f) - Z_i(f_0)\}\{g(X_i) - g_0(X_i)\} \right| \\ \leq \frac{\tau^2}{16}\|g-g_0\|_2^2 + \frac{\lambda_e \tau^2}{16}\|g-g_0\|_{\mathscr{H}}^2 + c_{16} C_\phi^2 \delta_2^2 \frac{\mathfrak{D}_{\lambda_q}(\mathfrak{D}_{\lambda_e}+t)}{n}$$

for a universal constant $c_{16} > 0$ with the same probability. Together, this bound and (D.75) give that with probability at least $1-e^{-t}$, the bound

$$\frac{\tau}{n}\sum_{i=1}^{n}\{Z_i(f) - Z_i(f_0)\}\{g(X_i) - g_0(X_i)\} \leq \frac{\tau^2}{16}\|g-g_0\|_2^2 + \frac{\lambda_e\tau^2}{16}\|g-g_0\|_{\mathscr{H}}^2 \\ + \frac{\bar{p}}{2}\tau\|f-f_0\|_4^2\|g-g_0\|_2 + c_{16}C_\phi^2\delta_2^2\frac{\mathfrak{D}_{\lambda_q}(\mathfrak{D}_{\lambda_e}+t)}{n}$$

holds uniformly for all $g \in \mathscr{H}$ and $f \in \mathscr{H}$ with $\|f-f_0\|_2^2 + \lambda_q\|f-f_0\|_{\mathscr{H}}^2 \leq \delta_2^2$, which completes the proof.

D.5.6 Proof of Lemma D.3.7

To begin with, we denote $E = \widehat{T}_\lambda - T_\lambda = \widehat{T} - T_K$. Since $\widehat{T}_\lambda \succeq \lambda I$ and $T_\lambda \succeq \lambda I$, $\widehat{T}_\lambda$ and T_λ are invertible. Note that $(I + T_\lambda^{-1}E)\widehat{T}_\lambda^{-1} = T_\lambda^{-1}$. To see this,

$$(I + T_\lambda^{-1}E)\widehat{T}_\lambda^{-1} = \widehat{T}_\lambda^{-1} + T_\lambda^{-1}E\widehat{T}_\lambda^{-1} = \widehat{T}_\lambda^{-1} + T_\lambda^{-1}(\widehat{T}_\lambda - T_\lambda)\widehat{T}_\lambda^{-1} = T_\lambda^{-1}. \tag{D.83}$$

Since $\|T_\lambda^{-1}E\|_{\mathrm{op},\lambda} \leq \zeta < 1$ by the assumption, the operator $(I + T_\lambda^{-1}E)$ is invertible and its inverse can be written as $(I+T_\lambda^{-1}E)^{-1} = \sum_{k=0}^{\infty}(-T_\lambda^{-1}E)^k$ (see, e.g. Corollary VII.2.3 in Conway (1990)). Therefore,

$$\|(I+T_\lambda^{-1}E)^{-1} - I\|_{\mathrm{op},\lambda} = \left\| \sum_{k=1}^{\infty}(-T_\lambda^{-1}E)^k \right\|_{\mathrm{op},\lambda} \leq \sum_{k=1}^{\infty}\left\|T_\lambda^{-1}E\right\|_{\mathrm{op},\lambda}^k \leq \sum_{k=1}^{\infty}\zeta^k = \frac{\zeta}{1-\zeta}.$$

Let $A := (I + T_\lambda^{-1}E)^{-1} - I$. By (D.83), we have

$$\widehat{T}_\lambda^{-1} - T_\lambda^{-1} = (I + T_\lambda^{-1}E)^{-1}T_\lambda^{-1} - T_\lambda^{-1} = AT_\lambda^{-1},$$

and $\|A\|_{\mathrm{op},\lambda} \le \zeta/(1-\zeta)$, thereby completing the proof. □

D.5.7 Proof of Lemma D.3.8

Proof of (D.22). Since $\lambda \gtrsim (t + \log \mathfrak{D}_\lambda)/n$, it follows from Lemma D.3.3 that the event

$$\mathscr{S} = \left\{ \frac{1}{2} \le \frac{\|h\|_n^2 + \lambda \|h\|_{\mathscr{H}}^2}{\|h\|_2^2 + \lambda \|h\|_{\mathscr{H}}^2} \le \frac{3}{2} \text{ for all } h \in \mathscr{H} \right\}$$

occurs with probability at least $1 - e^{-t}$. Conditioned on $\mathscr{S}$, $\|h\|_\lambda \le 1$ implies $\|h\|_n \le \sqrt{3/2}$ and $\|h\|_{\mathscr{H}} \le \lambda^{-1/2}$. Furthermore, if we write $h = \sum_{j=1}^\infty h_j \phi_j$,

$$\left\langle T_\lambda^{-1} K_{X_i}, h \right\rangle_\lambda = \left\langle \sum_{j=1}^\infty \frac{\mu_j}{\mu_j + \lambda} \phi_j(X_i)\phi_j, \sum_{j=1}^\infty h_j \phi_j \right\rangle_\lambda = \sum_{j=1}^\infty h_j \phi_j(X_i) = h(X_i). \tag{D.84}$$

Thus,

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \omega_i T_\lambda^{-1} K_{X_i} \right\|_\lambda &= \sup_{\|h\|_\lambda \le 1} \left\langle \frac{1}{n} \sum_{i=1}^n \omega_i T_\lambda^{-1} K_{X_i}, h \right\rangle_\lambda \\ &\le \sup_{\|h\|_n \le \sqrt{3/2}, \|h\|_{\mathscr{H}} \le \lambda^{-1/2}} \left| \frac{1}{n} \sum_{i=1}^n \omega_i h(X_i) \right|. \end{aligned}$$

Then, by employing the same argument which derives (D.73) in Lemma D.3.4, it can be shown that there exists an absolute constant $c_{17} > 0$ satisfying

$$\mathbb{P}\left[\left\{ \sup_{\|h\|_n \le \sqrt{3/2}, \|h\|_{\mathscr{H}} \le \lambda^{-1/2}} \left| \frac{1}{n} \sum_{i=1}^n \omega_i h(X_i) \right| \ge c_{17} \sigma_0 \sqrt{\frac{\mathfrak{D}_\lambda + t}{n}} \right\} \bigcap \mathscr{S} \right] \le 4e^{-t}.$$

Combining this probability bound with the fact that $\mathbb{P}(\mathscr{S}) \geq 1 - e^{-t}$ establishes the bound (D.22).

□

Proof of (D.23). Define $\Phi_\lambda(x) = (\sqrt{\nu_1}\phi_1(x), \sqrt{\nu_2}\phi_2(x), \ldots)^{\mathrm{T}}$ for $x \in \mathscr{X}$, where $\nu_j = \mu_j/(\mu_j + \lambda)$ for $j \geq 1$. By Condition 4.3.3,

$$\|\Phi_\lambda(x)\|_{\ell_2}^2 = \sum_{j=1}^{\infty} \nu_j \phi_j^2(x) \leq C_\phi^2 \mathfrak{D}_\lambda < \infty,$$

so that $\Phi_\lambda(x) \in \ell_2$ for any $x \in \mathscr{X}$. Also, for any $h \in \mathscr{H}$, we can express $h(\cdot) = \langle \boldsymbol{\theta}_h, \Phi_\lambda(\cdot)\rangle_{\ell_2}$, where $\boldsymbol{\theta}_h = (\theta_{h,1}, \theta_{h,2}, \ldots)^{\mathrm{T}}$ with $\theta_{h,j} = \langle h, \phi_j\rangle_2/\sqrt{\nu_j}$. Then, $\|h\|_\lambda^2 = \|\boldsymbol{\theta}_h\|_{\ell_2}^2$. Since $\mathbb{E}\widehat{T} = T_K$, we can write

$$\|T_\lambda^{-1}(\widehat{T} - T_K)\|_{\mathrm{op},\lambda} = \sup_{\|h\|_\lambda \leq 1} \left| \left\langle \frac{1}{n}\sum_{i=1}^{n} (1 - \mathbb{E}) T_\lambda^{-1}(K_{X_i} \otimes K_{X_i})h, h \right\rangle_\lambda \right|.$$

For $h = \langle \boldsymbol{\theta}_h, \Phi_\lambda\rangle_{\ell_2}$ with $\boldsymbol{\theta}_h = (\theta_{h,1}, \theta_{h,2}, \ldots)^{\mathrm{T}}$, we have

$$\begin{aligned}
\langle T_\lambda^{-1}(K_{X_i} \otimes K_{X_i})h, h\rangle_\lambda &= \langle h(X_i) T_\lambda^{-1} K_{X_i}, h\rangle_\lambda \\
&= h(X_i) \cdot \left\langle \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda} \phi_j(X_i)\phi_j, \sum_{j=1}^{\infty} \theta_{h,j}\sqrt{\nu_j}\phi_j \right\rangle_\lambda \\
&= h(X_i) \sum_{j=1}^{\infty} \theta_{h,j}\sqrt{\nu_j}\phi_j(X_i) = \langle \boldsymbol{\theta}_h, \Phi_\lambda(X_i)\rangle_{\ell_2}^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
\|T_\lambda^{-1}(\widehat{T} - T_K)\|_{\mathrm{op},\lambda} &= \sup_{\|\boldsymbol{\theta}_h\|_{\ell_2} \leq 1} \left| \left\langle \frac{1}{n}\sum_{i=1}^{n} (1 - \mathbb{E}) \Phi_\lambda(X_i)\Phi_\lambda(X_i)^{\mathrm{T}}\boldsymbol{\theta}_h, \boldsymbol{\theta}_h \right\rangle_{\ell_2} \right| \\
&= \left\| \frac{1}{n}\sum_{i=1}^{n} (1 - \mathbb{E}) \Phi_\lambda(X_i)\Phi_\lambda(X_i)^{\mathrm{T}} \right\|_{\mathrm{op}}. \qquad \text{(D.85)}
\end{aligned}$$

Define $\mathbf{M}_i = (1-\mathbb{E})\Phi_\lambda(X_i)\Phi_\lambda(X_i)^{\mathrm{T}}$. By the definition, $\mathbf{M}_i$ is centered and satisfies

$$\big\|\mathbf{M}_i\big\|_{\mathrm{op}} \le 2\sup_{x\in\mathscr{X}}\big\|\Phi_\lambda(x)\Phi_\lambda(x)^{\mathrm{T}}\|_{\mathrm{op}} = 2\sup_{x\in\mathscr{X}}\big\|\Phi_\lambda(x)\big\|_{\ell_2}^2 \le 2C_\phi^2\mathfrak{D}_\lambda.$$

Moreover,

$$\mathbb{E}\mathbf{M}_i^2 \preceq \mathbb{E}\|\Phi_\lambda(X_i)\|_{\ell_2}^2\Phi_\lambda(X_i)\Phi_\lambda(X_i)^{\mathrm{T}} \preceq C_\phi^2\mathfrak{D}_\lambda\mathbb{E}\Phi_\lambda(X_i)\Phi_\lambda(X_i)^{\mathrm{T}} =: \mathbf{V}.$$

Since $\nu_j = \mu_j/(\mu_j+\lambda)$, we have

$$\|\mathbf{V}\|_{\mathrm{op}} = C_\phi^2\mathfrak{D}_\lambda\max_{j\ge1}\frac{\mu_j}{\mu_j+\lambda} \le C_\phi^2\mathfrak{D}_\lambda,\ \ \mathrm{tr}(\mathbf{V}) = C_\phi^2\mathfrak{D}_\lambda\sum_{j=1}^{\infty}\frac{\mu_j}{\mu_j+\lambda} = C_\phi^2\mathfrak{D}_\lambda^2.$$

Applying Lemma D.5.1 yields that for any $t>0$,

$$\mathbb{P}\bigg\{\bigg\|\frac{1}{n}\sum_{i=1}^n\mathbf{M}_i\bigg\|_{\mathrm{op}} \ge \frac{4}{3}C_\phi^2\frac{\mathfrak{D}_\lambda(t+\log n)}{n} + 2C_\phi\sqrt{\frac{\mathfrak{D}_\lambda(t+\log n)}{n}}\bigg\} \le 14\mathfrak{D}_\lambda e^{-t-\log n}, \tag{D.86}$$

which, together with (D.85), establishes the bound (D.23). □

Proof of (D.24). To begin with, observe that

$$\begin{aligned}\bigg\|\frac{1}{n}\sum_{i=1}^n\big\{Z_i(f)-Z_i(f_0)\big\}T_\lambda^{-1}K_{X_i}\bigg\|_\lambda &= \sup_{\|h\|_\lambda\le1}\frac{1}{n}\sum_{i=1}^n\big\{Z_i(f)-Z_i(f_0)\big\}\big\langle T_\lambda^{-1}K_{X_i},h\big\rangle_\lambda \\ &= \sup_{\|h\|_\lambda\le1}\frac{1}{n}\sum_{i=1}^n\big\{Z_i(f)-Z_i(f_0)\big\}h(X_i),\end{aligned}$$

where the last step follows from (D.84). This implies

$$\begin{aligned}\bigg\|\frac{1}{n}\sum_{i=1}^n\big\{Z_i(f)-Z_i(f_0)\big\}T_\lambda^{-1}K_{X_i}\bigg\|_\lambda &\le \sup_{\|h\|_\lambda\le1}\bigg|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\big\{Z_i(f)-Z_i(f_0)\big\}h(X_i)\bigg| \\ &\quad+\sup_{\|h\|_\lambda\le1}\mathbb{E}\big|\big\{Z_i(f)-Z_i(f_0)\big\}h(X_i)\big|. \end{aligned}\tag{D.87}$$

For the expectation bound, we remark that $h \in \mathscr{H}$ with $\|h\|_\lambda \leq 1$ satisfies $\|h\|_2 \leq 1$. Therefore, by (D.75), we have for any $f \in \mathscr{H}$ that

$$\sup_{\|h\|_\lambda \leq 1} \mathbb{E}\big|\{Z_i(f) - Z_i(f_0)\}h(X_i)\big| \leq \sup_{\|h\|_2 \leq 1} \frac{\bar{p}}{2}\|f - f_0\|_4^2 \|h\|_2 \leq \frac{\bar{p}}{2}\|f - f_0\|_4^2. \tag{D.88}$$

Turning to the random fluctuation term, we note that $h \in \mathscr{H}$ with $\|h\|_\lambda \leq 1$ satisfies $\|h\|_2 \leq 1$ and $\|h\|_{\mathscr{H}} \leq \lambda^{-1/2}$. Thus, if we denote $\Delta_f = f_0 - f$ for any $f \in \mathscr{H}$ and $\mathscr{F}(\delta_2) := \{h \in \mathscr{H} : \|h\|_2^2 + \lambda_q \|h\|_{\mathscr{H}}^2 \leq \delta_2^2\}$, we have

$$\begin{aligned}
&\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \sup_{\|h\|_\lambda \leq 1} \left| \frac{1}{n}\sum_{i=1}^n (1 - \mathbb{E})\big\{Z_i(f) - Z_i(f_0)\big\}h(X_i) \right| \\
&\qquad \leq \sup_{\Delta_f \in \mathscr{F}(\delta_2)} \sup_{\|h\|_{\mathscr{H}} \leq \lambda^{-1/2}, \|h\|_2 \leq 1} \left| \frac{1}{n}\sum_{i=1}^n \big\{Z_i(f) - Z_i(f_0)\big\}h(X_i) \right|.
\end{aligned}$$

Since the right-hand side of the above inequality is identical to the random variable Ω introduced in (D.76) within the proof of Lemma D.3.5, (D.82) implies that, with probability at least $1 - e^{-t}$,

$$\sup_{\Delta_f \in \mathscr{F}(\delta_2)} \sup_{\|h\|_{\mathscr{H}} \leq \lambda^{-1/2}, \|h\|_2 \leq 1} \left| \frac{1}{n}\sum_{i=1}^n (1 - \mathbb{E})\big\{Z_i(f) - Z_i(f_0)\big\}h(X_i) \right| \lesssim C_\phi \delta_2 \sqrt{\frac{\mathfrak{D}_{\lambda_q}(\mathfrak{D}_{\lambda_e} + t)}{n}}.$$

Combining this bound with (D.87) and (D.88) establishes the claim. □

D.5.8 Proof of Lemma D.3.9

In the proof, we need Bernstein's inequality for unbounded self-adjoint operators. The following lemma is an extension of Proposition 4.1 in Klochkov and Zhivotovskiy (2020) to self-adjoint operators in an infinite-dimensional separable Hilbert space. This extension follows a similar argument as the one presented in Section 3.2 of Minsker (2017), so we omit the proof details for brevity. For any $p \geq 1$, define $\|\cdot\|_{\psi_p}$ to be the ψ_p-Orlicz norm, that is, $\|R\|_{\psi_p} := \inf\{u > 0 : \mathbb{E}\psi_p(|R|/u) \leq 1\}$ for any random variable R, where $\psi_p : \mathbb{R} \to \mathbb{R}$ is defined as $\psi_p(x) = \exp(x^p) - 1$.

Lemma D.5.2. Let $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_n$ be n independent self-adjoint random operators on a separable Hilbert space. Assume that $\mathbb{E}\mathbf{M}_i = \mathbf{0}$ for $1 \le i \le n$ and $\| \max_{1\le i\le n} \|\mathbf{M}_i\|_{\mathrm{op}} \|_{\psi_1} \le B$ for some $B \ge 0$. Moreover, there exist positive trace-class operator $\mathbf{V}$ and $\sigma > 0$ such that $(1/n)\sum_{i=1}^n \mathbb{E}\mathbf{M}_i^2 \preceq \mathbf{V}$ and $\|\mathbf{V}\|_{\mathrm{op}} \le \sigma^2$. Then, there exists an absolute constant $C > 0$ such that for any $u \gtrsim B/n + \sqrt{\sigma^2/n}$,

$$\mathbb{P}\left(\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{M}_i \right\|_{\mathrm{op}} \ge u \right) \le 15 \frac{\mathrm{tr}(\mathbf{V})}{\sigma^2} \exp\left\{ -C\left(\frac{nu^2}{\sigma^2} + \frac{nu}{B} \right) \right\}.$$

Now, we are ready to prove the lemma.

Proof of Lemma D.3.9. To begin with, recall that

$$\delta_n = \lambda_q^{r_q+1/2}\|f^*\|_{\mathscr{H}} + \sqrt{\frac{\mathfrak{D}_{\lambda_q}+t}{n}} \quad \text{and} \quad \gamma_n = \tau\lambda_e^{r_e+1/2}\|g^*\|_{\mathscr{H}} + \sigma_0\sqrt{\frac{\mathfrak{D}_{\lambda_e}+t}{n}}.$$

Theorem 4.3.1 implies that the event

$$\mathscr{G}_1(t) := \left\{ \|\widehat{f} - f_0\|_2^2 + \lambda_q \|\widehat{f} - f_0\|_{\mathscr{H}}^2 \le C_1^2 \delta_n^2 \right\}$$

occurs with probability at least $1 - e^{-t}$, where $C_1 > 0$ is an absolute constant. Then, conditioned on $\mathscr{G}_1(t)$, (D.79) implies $\|\widehat{f} - f_0\|_4^2 \le \|\widehat{f} - f_0\|_\infty \|\widehat{f} - f_0\|_2 \le C_1^2 C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \delta_n^2$. Combining this with Theorem 4.3.2 and recalling the definition of δ_{s}, there exists an absolute constant $C_2 > 0$ such that the event

$$\mathscr{G}_2(t) := \left\{ \|\widehat{g} - g_0\|_2^2 + \lambda_e \|\widehat{g} - g_0\|_{\mathscr{H}}^2 \le C_2^2 r_n^2/\tau^2 := C_2^2(\gamma_n + \delta_{\mathrm{s}} + \bar{p} C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \delta_n^2)^2/\tau^2 \right\} \quad \text{(D.89)}$$

satisfies $\mathbb{P}\{\mathscr{G}_1(t) \cap \mathscr{G}_2(t)\} \ge 1 - 7e^{-t}$.

Now, note that $\omega_i = Z_i(f_0) - \tau g_0(X_i)$ and thus

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} U_i\big\{Z_i(\widehat{f}) - \tau\widehat{g}(X_i)\big\}\widehat{T}_{\lambda_e}^{-1}K_{X_i} - \frac{1}{n}\sum_{i=1}^{n} U_i\omega_i T_{\lambda_e}^{-1}K_{X_i} \\
&= \big(\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1}\big)\left[\frac{1}{n}\sum_{i=1}^{n} U_i\big\{Z_i(\widehat{f}) - \tau\widehat{g}(X_i)\big\}K_{X_i}\right] \\
&\quad + T_{\lambda_e}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} U_i\big\{Z_i(\widehat{f}) - Z_i(f_0) + \tau g_0(X_i) - \tau\widehat{g}(X_i)\big\}K_{X_i}\right] \\
&= \big(\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1}\big)\frac{1}{n}\sum_{i=1}^{n} U_i\omega_i K_{X_i} \\
&\quad + \big(\widehat{T}_{\lambda_e}^{-1} - T_{\lambda_e}^{-1}\big)\left[\frac{1}{n}\sum_{i=1}^{n} U_i\big\{Z_i(\widehat{f}) - Z_i(f_0) + \tau g_0(X_i) - \tau\widehat{g}(X_i)\big\}K_{X_i}\right] \\
&\quad + T_{\lambda_e}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} U_i\big\{Z_i(\widehat{f}) - Z_i(f_0) + \tau g_0(X_i) - \tau\widehat{g}(X_i)\big\}K_{X_i}\right]. \qquad \text{(D.90)}
\end{aligned}
$$

When $n \geq 64C_\phi^2\mathfrak{D}_{\lambda_e}(t+\log n)\log n$, $\zeta := 4C_\phi\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{(t+\log n)/n} \leq 1/2$ and Lemma D.3.8 implies that the event

$$
\mathscr{G}_3(t) := \big\{\big\|T_{\lambda_e}^{-1}(\widehat{T} - T_K)\big\|_{\mathrm{op},\lambda_e} \leq \zeta\big\}
$$

occurs with probability at least $1 - e^{-t}$. Denote

$$
\begin{aligned}
&D_1 := \left\|\frac{1}{n}\sum_{i=1}^{n} U_i\omega_i T_{\lambda_e}^{-1}K_{X_i}\right\|_{\lambda_e},\ D_2 := \sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)}\left\|\frac{1}{n}\sum_{i=1}^{n} U_i\big\{Z_i(f) - Z_i(f_0)\big\}T_{\lambda_e}^{-1}K_{X_i}\right\|_{\lambda_e}, \\
&\text{and } D_3 := \sup_{\|g-g_0\|_{\lambda_e} \leq C_2 r_n/\tau}\left\|\frac{1}{n}\sum_{i=1}^{n} U_i\tau\big\{g(X_i) - g_0(X_i)\big\}T_{\lambda_e}^{-1}K_{X_i}\right\|_{\lambda_e},
\end{aligned}
$$

where $\Delta_f = f_0 - f$ for $f \in \mathscr{H}$ and $\mathscr{F}(C_1\delta_n) := \{h \in \mathscr{H} : \|h\|_2^2 + \lambda_q\|h\|_{\mathscr{H}}^2 \leq C_1^2\delta_n^2\}$. By applying a similar argument as in the proof of Theorem 4.3.3, Lemma D.3.7 and the decomposition (D.90)

imply that, conditioned on $\mathscr{G}_1(t)\cap\mathscr{G}_2(t)\cap\mathscr{G}_3(t)$,

$$\left\|\frac{1}{n}\sum_{i=1}^n U_i\{Z_i(\widehat{f})-\tau\widehat{g}(X_i)\}\widehat{T}_{\lambda_e}^{-1}K_{X_i}-\frac{1}{n}\sum_{i=1}^n U_i\omega_i T_{\lambda_e}^{-1}K_{X_i}\right\|_{\lambda_e}\le 2\zeta D_1+(1+2\zeta)(D_2+D_3)$$
$$\le 2\zeta D_1+2(D_2+D_3). \qquad \text{(D.91)}$$

Therefore, it suffices to establish high-probability bounds for D_1, D_2 and D_3, respectively.

STEP I. BOUND FOR D_1. Note that $\langle T_{\lambda_e}^{-1}K_{X_i},h\rangle_{\lambda_e}=h(X_i)$ for any $h\in\mathscr{H}$ by (D.84). Thus,

$$D_1=\left\|\frac{1}{n}\sum_{i=1}^n U_i\omega_i T_{\lambda_e}^{-1}K_{X_i}\right\|_{\lambda_e}=\sup_{\|h\|_{\lambda_e}\le 1}\left\langle\frac{1}{n}\sum_{i=1}^n U_i\omega_i T_{\lambda_e}^{-1}K_{X_i},h\right\rangle_{\lambda_e}$$
$$=\sup_{\|h\|_{\lambda_e}\le 1}\frac{1}{n}\sum_{i=1}^n U_i\omega_i h(X_i).$$

Applying a similar argument as in the proof of Lemma D.3.4, combining conditional versions of the generic chaining bound (Theorem 8.5.5 in Vershynin (2018)) and Talagrand's majorizing measure theorem (Theorem 8.6.1 in Vershynin (2018)) yields that, with $\mathbb{P}^*$-probability at least $1-2e^{-t}$,

$$\sup_{\|h\|_{\lambda_e}\le 1}\frac{1}{n}\sum_{i=1}^n U_i\omega_i h(X_i)$$
$$\lesssim \sigma_W\left[\mathbb{E}^*\left\{\sup_{\|h\|_{\lambda_e}\le 1}\left|\frac{1}{n}\sum_{i=1}^n\xi_i\omega_i h(X_i)\right|\right\}+\sup_{\|h\|_{\lambda_e}\le 1}\left\{\frac{1}{n}\sum_{i=1}^n\omega_i^2h^2(X_i)\right\}^{1/2}\sqrt{\frac{t}{n}}\right], \qquad \text{(D.92)}$$

where $\xi_1,\ldots,\xi_n\sim\mathscr{N}(0,1)$ are mutually independent, and are also independent of $\mathscr{D}_n$.

We next establish bounds for the data-dependent quantities on the right-hand side of (D.92). By Jensen's inequality, we have

$$\mathbb{E}^*\left\{\sup_{\|h\|_{\lambda_e}\le 1}\left|\sum_{i=1}^n\xi_i\omega_i h(X_i)\right|\right\}\le\left[\mathbb{E}^*\left\{\sup_{\|h\|_{\lambda_e}\le 1}\left|\sum_{i=1}^n\xi_i\omega_i h(X_i)\right|^2\right\}\right]^{1/2}. \qquad \text{(D.93)}$$

If we write $h(\cdot)=\sum_{j=1}^\infty\theta_{h,j}\phi_j(\cdot)$ with $\|h\|_{\lambda_e}\le 1$, then $\{\theta_{h,j}\}_{j=1}^\infty$ satisfies $\sum_{j=1}^\infty\theta_{h,j}^2/\nu_{j,e}\le 1$,

where $\nu_{j,e} = \mu_j/(\mu_j + \lambda_e)$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}^*\left\{\sup_{\|h\|_{\lambda_e}\le 1}\left|\sum_{i=1}^n \xi_i\omega_i h(X_i)\right|^2\right\} &= \mathbb{E}^*\left\{\sup_{\sum_{j=1}^\infty \theta_{h,j}^2/\nu_{j,e}\le 1}\left|\sum_{j=1}^\infty\sum_{i=1}^n \xi_i\omega_i\theta_{h,j}\phi_j(X_i)\right|^2\right\}\\ &\le \mathbb{E}^*\sum_{j=1}^\infty\left|\sum_{i=1}^n \xi_i\omega_i\sqrt{\nu_{j,e}}\phi_j(X_i)\right|^2\\ &\le C_\phi^2\sum_{j=1}^\infty \nu_{j,e}\sum_{i=1}^n\omega_i^2 = C_\phi^2\mathfrak{D}_{\lambda_e}\sum_{i=1}^n\omega_i^2,\end{aligned}$$

To bound the data-dependent quantity $\sum_{i=1}^n \omega_i^2$, note that ω_i is sub-Gaussian, so Lemma 2.7.6 in Vershynin (2018) implies ω_i^2 is sub-exponential and $\|\omega_i^2\|_{\psi_1} = \|\omega_i\|_{\psi_2}^2 \lesssim \sigma_0^2$. Then, by Theorem 2.8.1 in Vershynin (2018), we have

$$\frac{1}{n}\sum_{i=1}^n\omega_i^2 = \mathbb{E}\omega_i^2 + \frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\omega_i^2 \lesssim \sigma_0^2 + \sigma_0^2\left(\sqrt{\frac{t}{n}}\vee\frac{t}{n}\right)\lesssim\sigma_0^2$$

with probability at least $1-2e^{-t}$ when $n \gtrsim t$. Thus, it follows that, with probability at least $1-2e^{-t}$,

$$\mathbb{E}^*\left\{\sup_{\|h\|_{\lambda_e}\le 1}\left|\sum_{i=1}^n \xi_i\omega_i h(X_i)\right|^2\right\}\lesssim C_\phi^2\sigma_0^2\mathfrak{D}_{\lambda_e}n,$$

which, combined with (D.93), further implies

$$\mathbb{E}^*\left\{\sup_{\|h\|_{\lambda_e}\le 1}\left|\frac{1}{n}\sum_{i=1}^n \xi_i\omega_i h(X_i)\right|\right\}\lesssim C_\phi\sigma_0\sqrt{\frac{\mathfrak{D}_{\lambda_e}}{n}}. \tag{D.94}$$

Turning to the second term in the right-hand side of (D.92), remark that

$$\begin{aligned}\sup_{\|h\|_{\lambda_e}\le 1}\frac{1}{n}\sum_{i=1}^n\omega_i^2h^2(X_i) &\le \sup_{\|h\|_{\lambda_e}\le 1}\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\omega_i^2h^2(X_i) + \sup_{\|h\|_{\lambda_e}\le 1}\mathbb{E}\omega_1^2h^2(X_1)\\ &\le \sup_{\|h\|_{\lambda_e}\le 1}\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\omega_i^2h^2(X_i) + 4\sigma_0^2,\end{aligned} \tag{D.95}$$

where the last inequality follows by (D.37) and the fact that $h \in \mathscr{H}$ with $\|h\|_{\lambda_e} \le 1$ satisfies $\|h\|_2 \le 1$. Now, denote $\Phi_{\lambda_e}(\cdot) = (\sqrt{\nu_{1,e}}\phi_1(\cdot), \sqrt{\nu_{2,e}}\phi_2(\cdot), \ldots)^{\mathrm{T}}$ with $\nu_{j,e} = \mu_j/(\mu_j + \lambda_e)$ for $j \ge 1$. Note that $\|\Phi_{\lambda_e}(\cdot)\|_{\ell_2}^2 = \sum_{j=1}^{\infty} \nu_{j,e}\phi_j(\cdot)^2 \le C_\phi^2 \mathfrak{D}_{\lambda_e}$. Moreover, if we write $h(\cdot) = \langle \boldsymbol{\theta}, \Phi_{\lambda_e}(\cdot)\rangle_{\ell_2}$ with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)^{\mathrm{T}}$, then $\|h\|_{\lambda_e}^2 = \sum_{j=1}^{\infty} \theta_j^2 = \|\boldsymbol{\theta}\|_{\ell_2}^2$. Therefore, we have

$$\begin{aligned}\sup_{\|h\|_{\lambda_e} \le 1} \frac{1}{n}\sum_{i=1}^{n} (1-\mathbb{E})\omega_i^2 h^2(X_i) &= \sup_{\|\boldsymbol{\theta}\|_{\ell_2} \le 1} \frac{1}{n}\sum_{i=1}^{n} (1-\mathbb{E})\omega_i^2 \langle \boldsymbol{\theta}, \Phi_{\lambda_e}(X_i)\rangle_{\ell_2}^2 \\ &= \left\| \frac{1}{n}\sum_{i=1}^{n} (1-\mathbb{E})\omega_i^2 \Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}} \right\|_{\mathrm{op}}.\end{aligned}$$

To apply Lemma D.5.2, remark that

$$\begin{aligned}\left\| \max_{1\le i\le n} \left\|(1-\mathbb{E})\omega_i^2 \Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\right\|_{\mathrm{op}} \right\|_{\psi_1} &\le 2 \left\| \max_{1\le i\le n} \left\|\omega_i^2 \Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\right\|_{\mathrm{op}} \right\|_{\psi_1} \\ &\le 2 \left\| \max_{1\le i\le n} \left\|\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\right\|_{\mathrm{op}} \right\|_{\infty} \left\| \max_{1\le i\le n} \omega_i^2 \right\|_{\psi_1} \\ &\lesssim \left\| \max_{1\le i\le n} \left\|\Phi_{\lambda_e}(X_i)\right\|_{\ell_2}^2 \right\|_{\infty} \log(n) \max_{1\le i\le n} \|\omega_i^2\|_{\psi_1} \\ &\lesssim C_\phi^2 \sigma_0^2 \mathfrak{D}_{\lambda_e} \log n,\end{aligned}$$

where the third inequality follows from Lemma 2.2.2 in van der Vaart and Wellner (1996) and the last inequality uses $\|\omega_i^2\|_{\psi_1} \lesssim \sigma_0^2$ again. Moreover,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\{(1-\mathbb{E})\omega_i^2 \Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\right\}^2 &\preceq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\omega_i^4 \left\{\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\right\}^2 \\ &\preceq 16\sigma_0^4 \sup_{x\in\mathscr{X}} \|\Phi_{\lambda_e}(x)\|_{\ell_2}^2 \mathbb{E}\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}} \\ &\preceq 16\sigma_0^4 C_\phi^2 \mathfrak{D}_{\lambda_e} \mathbb{E}\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}} =: \mathbf{V}_1.\end{aligned}$$

Then, it is obvious that $\mathrm{tr}(\mathbf{V}_1) = 16\sigma_0^4 C_\phi^2 \mathfrak{D}_{\lambda_e}^2$ and $\|\mathbf{V}_1\|_{\mathrm{op}} \le 16\sigma_0^4 C_\phi^2 \mathfrak{D}_{\lambda_e}$. Note that under the sample size requirement $n \ge 64 C_\phi^2 \mathfrak{D}_{\lambda_e}(t + \log n)\log n$, $15\mathfrak{D}_{\lambda_e} e^{-t-\log n} \le e^{-t}$. Thus, under this

sample size requirement, Lemma D.5.2 implies that, with probability at least $1-e^{-t}$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\omega_i^2\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\right\|_{\mathrm{op}} \lesssim C_\phi\sigma_0^2\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{t+\log n}{n}}+\frac{C_\phi^2\sigma_0^2\mathfrak{D}_{\lambda_e}\log(n)(t+\log n)}{n}$$
$$\lesssim \sigma_0^2.$$

Combining the above inequality with (D.92), (D.94) and (D.95), we conclude that there exists an event $\mathscr{G}_4(t)$ such that $\mathbb{P}\{\mathscr{G}_4(t)\}\geq 1-3e^{-t}$ and with $\mathbb{P}^*$-probability at least $1-2e^{-t}$ conditioned on $\mathscr{G}_4(t)$,

$$D_1=\sup_{\|h\|_{\lambda_e}\leq 1}\frac{1}{n}\sum_{i=1}^{n}U_i\omega_i h(X_i)\leq C_3\sqrt{\frac{\mathfrak{D}_{\lambda_e}+t}{n}}, \tag{D.96}$$

where $C_3=C_3(C_\phi,\sigma_0,\sigma_W)>0$.

STEP II. BOUND FOR D_2. Remark that

$$\begin{aligned}D_2&=\sup_{\Delta_f\in\mathscr{F}(C_1\delta_n)}\left\|\frac{1}{n}\sum_{i=1}^{n}U_i\{Z_i(f)-Z_i(f_0)\}T_{\lambda_e}^{-1}K_{X_i}\right\|_{\lambda_e}\\&=\sup_{\Delta_f\in\mathscr{F}(C_1\delta_n)}\sup_{\|h\|_{\lambda_e}\leq 1}\left\langle\frac{1}{n}\sum_{i=1}^{n}U_i\{Z_i(f)-Z_i(f_0)\}T_{\lambda_e}^{-1}K_{X_i},h\right\rangle_{\lambda_e}\\&=\sup_{\Delta_f\in\mathscr{F}(C_1\delta_n)}\sup_{\|h\|_{\lambda_e}\leq 1}\frac{1}{n}\sum_{i=1}^{n}U_i\{Z_i(f)-Z_i(f_0)\}h(X_i).\end{aligned}$$

By following a similar argument as (D.92) in Step I, combining conditional versions of the generic chaining bound and Talagrand's majorizing measure theorem gives that, with $\mathbb{P}^*$-probability at

least $1-2e^{-t}$,

$$\begin{aligned}
&\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)} \sup_{\|h\|_{\lambda_e}\le 1} \frac{1}{n}\sum_{i=1}^n U_i\big\{Z_i(f)-Z_i(f_0)\big\}h(X_i) \\
&\qquad \lesssim \sigma_W \mathbb{E}^*\bigg[\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)} \sup_{\|h\|_{\lambda_e}\le 1} \bigg|\frac{1}{n}\sum_{i=1}^n \xi_i\big\{Z_i(f)-Z_i(f_0)\big\}h(X_i)\bigg|\bigg] \\
&\qquad\quad + \sigma_W \sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)} \sup_{\|h\|_{\lambda_e}\le 1} \bigg[\frac{1}{n}\sum_{i=1}^n \big\{Z_i(f)-Z_i(f_0)\big\}^2 h^2(X_i)\bigg]^{1/2}\sqrt{\frac{t}{n}}, \qquad \text{(D.97)}
\end{aligned}$$

where $\xi_1,\ldots,\xi_n \sim \mathscr{N}(0,1)$ are mutually independent, and are also independent of $\mathscr{D}_n$.

We first establish a high probability bound for the first term on the right-hand side of (D.97). To apply a similar argument as in the derivation of (D.81), write $h(\cdot)=\sum_{j\ge 1} h_j\phi_j(\cdot)$. Then, $\|h\|_{\lambda_e}\le 1$ is equivalent to $\sum_{j\ge 1} h_j^2/\nu_{j,e}\le 1$, where $\nu_{j,e}=\mu_j/(\mu_j+\lambda_e)$ for $j\ge 1$. Applying the Cauchy-Schwartz inequality yields

$$\begin{aligned}
&\mathbb{E}^*\bigg[\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)} \sup_{\|h\|_{\lambda_e}\le 1} \bigg|\frac{1}{n}\sum_{i=1}^n \xi_i\big\{Z_i(f)-Z_i(f_0)\big\}h(X_i)\bigg|\bigg] \\
&= \mathbb{E}^*\bigg[\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)} \sup_{\|h\|_{\lambda_e}\le 1} \bigg|\frac{1}{n}\sum_{j=1}^\infty \frac{h_j}{\sqrt{\nu_{j,e}}}\sum_{i=1}^n \xi_i\big\{Z_i(f)-Z_i(f_0)\big\}\sqrt{\nu_{j,e}}\phi_j(X_i)\bigg|\bigg] \\
&\le \mathbb{E}^*\bigg\{\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)} \sum_{j=1}^\infty \nu_{j,e}\bigg[\frac{1}{n}\sum_{i=1}^n \xi_i\big\{Z_i(f)-Z_i(f_0)\big\}\phi_j(X_i)\bigg]^2\bigg\} \\
&\le C_\phi^2 \sum_{j=1}^\infty \nu_{j,e}\mathbb{E}^*\bigg\{\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)} \bigg[\frac{1}{n}\sum_{i=1}^n \xi_i\big\{Z_i(f)-Z_i(f_0)\big\}\frac{\phi_j(X_i)}{C_\phi}\bigg]^2\bigg\}.
\end{aligned}$$

Now, fix $j\ge 1$ and take $F(u)=u^2$ for $u\ge 0$. Moreover, let $\mathbb{T}$ be a subset of $\mathbb{R}^n$ such that

$$\mathbb{T} := \big\{\boldsymbol{t}=(t_1,t_2,\ldots,t_n)\in\mathbb{R}^n : t_i = f(X_i)-f_0(X_i),\ \ 1\le i\le n, \Delta_f\in\mathscr{F}(C_1\delta_n)\big\},$$

and define $\varphi_i:\mathbb{R}\to\mathbb{R}$ as

$$\varphi_i(a) = \frac{\phi_j(X_i)}{C_\phi}\big\{(\varepsilon_i-a)\mathbb{1}(\varepsilon_i\le a)+\tau f_0(X_i)+\tau a - Z_i(f_0)\big\},$$

so that $\varphi_i(f(X_i) - f_0(X_i)) = \{Z_i(f) - Z_i(f_0)\}\phi_j(X_i)/C_\phi$. Since φ_i is a contraction for $1 \le i \le n$ by Condition 4.3.3 and (D.78), applying Corollary 3.17 in Ledoux and Talagrand (1991) gives

$$\begin{aligned}
&\mathbb{E}^*\left\{\left[\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)}\left|\sum_{i=1}^n \xi_i\{Z_i(f) - Z_i(f_0)\}\frac{\phi_j(X_i)}{C_\phi}\right|\right]^2\right\} \\
&= \mathbb{E}^* F\left(\sup_{t\in\mathbb{T}}\left|\sum_{i=1}^n \xi_i \varphi_i(t_i)\right|\right) \\
&\le 16\mathbb{E}^* F\left(\sup_{t\in\mathbb{T}}\left|\sum_{i=1}^n \xi_i t_i\right|\right) \\
&= 16\mathbb{E}^*\left[\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)}\left|\sum_{i=1}^n \xi_i\{f(X_i) - f_0(X_i)\}\right|^2\right].
\end{aligned}$$

After writing $f - f_0 = \sum_{k\ge 1} f_k\phi_k$, it follows that

$$\mathbb{E}^*\left[\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)}\left|\sum_{i=1}^n \xi_i\{f(X_i) - f_0(X_i)\}\right|^2\right] = \mathbb{E}^*\left[\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)}\left|\sum_{i=1}^n \xi_i \sum_{k=1}^\infty f_k\phi_k(X_i)\right|^2\right].$$

Since $\Delta_f \in \mathscr{F}(C_1\delta_n)$ is equivalent to $\sum_{k=1}^\infty f_k^2/\nu_{k,q} \le C_1^2\delta_n^2$ with $\nu_{k,q} = \mu_k/(\mu_k + \lambda_q)$ for $k \ge 1$, we have

$$\begin{aligned}
\mathbb{E}^*\left[\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)}\left|\sum_{i=1}^n \xi_i \sum_{k=1}^\infty f_k\phi_k(X_i)\right|^2\right] &= \mathbb{E}^*\left\{\sup_{\Delta_f \in \mathscr{F}(C_1\delta_n)}\left|\sum_{k=1}^\infty \frac{f_k}{\sqrt{\nu_{k,q}}}\sum_{i=1}^n \xi_i\sqrt{\nu_{k,q}}\phi_k(X_i)\right|^2\right\} \\
&\le C_1^2\delta_n^2 \cdot \sum_{k=1}^\infty \mathbb{E}^*\left\{\sum_{i=1}^n \xi_i\sqrt{\nu_{k,e}}\phi_k(X_i)\right\}^2 \\
&= C_1^2\delta_n^2 \sum_{k=1}^\infty \mathbb{E}^*\left\{\sum_{i=1}^n \nu_{k,q}\phi_k^2(X_i)\right\} \le nC_1^2C_\phi^2\delta_n^2\mathfrak{D}_{\lambda_q},
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality. Putting the pieces

together and applying Jensen's inequality, we have

$$\mathbb{E}^*\Bigg[\sup_{\Delta_f\in\mathscr{F}(C_1\delta_n)}\sup_{\|h\|_{\lambda_e}\le 1}\bigg|\frac{1}{n}\sum_{i=1}^n\xi_i\big\{Z_i(f)-Z_i(f_0)\big\}h(X_i)\bigg|\Bigg]$$
$$\le 4C_1C_\phi^2\delta_n\frac{\mathfrak{D}_{\lambda_q}}{n}\bigg(\sum_{j=1}^\infty \nu_{j,e}\bigg)^{1/2}=4C_1C_\phi^2\delta_n\sqrt{\frac{\mathfrak{D}_{\lambda_q}\mathfrak{D}_{\lambda_e}}{n}}. \tag{D.98}$$

For the second term on the right-hand side of (D.97), recall that $\Delta_f = f_0 - f$ and $\mathscr{F}(C_1\delta_n) = \{h\in\mathscr{H}: \|h\|_2^2+\lambda_q\|h\|_{\mathscr{H}}^2\le C_1^2\delta_n^2\}$. Combining (D.78) and (D.79) with $\lambda=\lambda_q$ gives

$$\sup_{\Delta_f\in\mathscr{F}(C_1\delta_n)}\|Z_i(f)-Z_i(f_0)\|_\infty^2\le\sup_{\Delta_f\in\mathscr{F}(C_1\delta_n)}\|f-f_0\|_\infty^2\le C_1^2C_\phi^2\mathfrak{D}_{\lambda_q}\delta_n^2.$$

Thus, we have

$$\sup_{\Delta_f\in\mathscr{F}(C_1\delta_n)}\sup_{\|h\|_{\lambda_e}\le1}\frac{1}{n}\sum_{i=1}^n\big\{Z_i(f)-Z_i(f_0)\big\}^2h^2(X_i)\le C_1^2C_\phi^2\mathfrak{D}_{\lambda_q}\delta_n^2\cdot\sup_{\|h\|_{\lambda_e}\le1}\frac{1}{n}\sum_{i=1}^n h^2(X_i).$$

Moreover, if we write $h(\cdot)=\langle\boldsymbol{\theta},\Phi_{\lambda_e}(\cdot)\rangle_{\ell_2}$, where $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots)^{\mathrm{T}}$ with $\theta_j=\langle h,\phi_j\rangle_2/\sqrt{\nu_{j,e}}$, then we have $\|h\|_{\lambda_e}^2=\sum_{j=1}^\infty\theta_j^2=\|\boldsymbol{\theta}\|_{\ell_2}^2$. Therefore,

$$\sup_{\|h\|_{\lambda_e}\le1}\frac{1}{n}\sum_{i=1}^n h^2(X_i)=\sup_{\|\boldsymbol{\theta}\|_{\ell_2}\le1}\frac{1}{n}\sum_{i=1}^n\langle\boldsymbol{\theta},\Phi_{\lambda_e}(X_i)\rangle_{\ell_2}^2=\bigg\|\frac{1}{n}\sum_{i=1}^n\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\bigg\|_{\mathrm{op}}.$$

To derive a high probability bound for the operator norm, we apply the bound (D.86) for $\lambda=\lambda_e$, leading to

$$\mathbb{P}\bigg\{\bigg\|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\bigg\|_{\mathrm{op}}\ge\frac{4}{3}C_\phi^2\frac{\mathfrak{D}_{\lambda_e}(t+\log n)}{n}+2C_\phi\sqrt{\frac{\mathfrak{D}_{\lambda_e}(t+\log n)}{n}}\bigg\}$$
$$\le 14\mathfrak{D}_{\lambda_e}e^{-t-\log n}.$$

Under the sample size requirement $n\ge 64C_\phi^2\mathfrak{D}_{\lambda_e}(t+\log n)$, the above probability bound implies

that the event $\mathscr{G}_5(t)$, defined as

$$\mathscr{G}_5(t) := \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} (1-\mathbb{E}) \Phi_{\lambda_e}(X_i) \Phi_{\lambda_e}(X_i)^{\mathrm{T}} \right\|_{\mathrm{op}} \leq 1 \right\} \tag{D.99}$$

satisfies $\mathbb{P}\{\mathscr{G}_5(t)\} \geq 1 - e^{-t}$. Then, since $\left\| \mathbb{E} \Phi_{\lambda_e}(X_1) \Phi_{\lambda_e}(X_1)^{\mathrm{T}} \right\|_{\mathrm{op}} \leq 1$, conditioned on $\mathscr{G}_5(t)$, we have

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=1}^{n} \Phi_{\lambda_e}(X_i) \Phi_{\lambda_e}(X_i)^{\mathrm{T}} \right\|_{\mathrm{op}} \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^{n} (1-\mathbb{E}) \Phi_{\lambda_e}(X_i) \Phi_{\lambda_e}(X_i)^{\mathrm{T}} \right\|_{\mathrm{op}} + \left\| \mathbb{E} \Phi_{\lambda_e}(X_i) \Phi_{\lambda_e}(X_i)^{\mathrm{T}} \right\|_{\mathrm{op}} \leq 2, \end{aligned} \tag{D.100}$$

which implies

$$\sup_{\Delta_f \in \mathscr{F}(C_1 \delta_n)} \sup_{\|h\|_{\lambda_e} \leq 1} \frac{1}{n} \sum_{i=1}^{n} \left\{ Z_i(f) - Z_i(f_0) \right\}^2 h^2(X_i) \leq 2 C_1^2 C_\phi^2 \mathfrak{D}_{\lambda_q} \delta_n^2.$$

By combining the above bound, (D.97) and (D.98), it follows that

$$\begin{aligned} D_2 &\lesssim \sigma_W C_\phi^2 \delta_n \sqrt{\frac{\mathfrak{D}_{\lambda_q} \mathfrak{D}_{\lambda_e}}{n}} + \sigma_W C_\phi \mathfrak{D}_{\lambda_q}^{1/2} \delta_n \sqrt{\frac{t}{n}} \\ &\leq C_4 \delta_n \mathfrak{D}_{\lambda_q}^{1/2} \sqrt{\frac{\mathfrak{D}_{\lambda_e} + t}{n}} \end{aligned} \tag{D.101}$$

with $\mathbb{P}^*$-probability at least $1 - 2e^{-t}$ conditioned on $\mathscr{G}_5(t)$, where $C_4 = C_4(C_\phi, \sigma_W)$.

STEP III. BOUND FOR D_3. Note that

$$\begin{aligned}
D_3 &= \sup_{\|g-g_0\|_{\lambda_e}\le C_2 r_n/\tau} \left\| \frac{1}{n}\sum_{i=1}^n U_i \tau\big\{g(X_i)-g_0(X_i)\big\} T_{\lambda_e}^{-1} K_{X_i} \right\|_{\lambda_e} \\
&= \sup_{\|g-g_0\|_{\lambda_e}\le C_2 r_n/\tau} \sup_{\|h\|_{\lambda_e}\le 1} \left\langle \frac{1}{n}\sum_{i=1}^n U_i \tau\big\{g(X_i)-g_0(X_i)\big\} T_{\lambda_e}^{-1} K_{X_i}, h \right\rangle_{\lambda_e} \\
&= \sup_{\|g-g_0\|_{\lambda_e}\le C_2 r_n/\tau} \sup_{\|h\|_{\lambda_e}\le 1} \frac{1}{n}\sum_{i=1}^n U_i \tau\big\{g(X_i)-g_0(X_i)\big\} h(X_i) \\
&= C_2 r_n \sup_{\|h_1\|_{\lambda_e}\le 1} \sup_{\|h_2\|_{\lambda_e}\le 1} \frac{1}{n}\sum_{i=1}^n U_i h_1(X_i) h_2(X_i). \qquad \text{(D.102)}
\end{aligned}$$

Recall that $\Phi_{\lambda_e}(\cdot) = (\sqrt{\nu_{1,e}}\phi_1(\cdot), \sqrt{\nu_{2,e}}\phi_2(\cdot), \ldots)^{\mathrm{T}}$ with $\nu_{j,e} = \mu_j/(\mu_j+\lambda_e)$ and when $h(\cdot) = \langle \boldsymbol{\theta}, \Phi_{\lambda_e}(\cdot)\rangle_{\ell_2}$, $\|h\|_{\lambda_e}^2 \le 1$ is equivalent to $\|\boldsymbol{\theta}\|_{\ell_2}^2 \le 1$. Thus, we have

$$\begin{aligned}
&\sup_{\|h_1\|_{\lambda_e}\le 1} \sup_{\|h_2\|_{\lambda_e}\le 1} \frac{1}{n}\sum_{i=1}^n U_i h_1(X_i) h_2(X_i) \\
&= \sup_{\|\boldsymbol{\theta}_1\|_{\ell_2}\le 1} \sup_{\|\boldsymbol{\theta}_2\|_{\ell_2}\le 1} \frac{1}{n}\sum_{i=1}^n U_i \langle \boldsymbol{\theta}_1, \Phi_{\lambda_e}(X_i)\rangle_{\ell_2} \langle \boldsymbol{\theta}_2, \Phi_{\lambda_e}(X_i)\rangle_{\ell_2} \\
&= \left\| \frac{1}{n}\sum_{i=1}^n U_i \Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}} \right\|_{\mathrm{op}}. \qquad \text{(D.103)}
\end{aligned}$$

To apply Lemma D.5.2, note that $\|\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\|_{\mathrm{op}} \le \|\Phi_{\lambda_e}(X_i)\|_{\ell_2}^2 \le C_\phi^2 \mathfrak{D}_{\lambda_e}$. Therefore,

$$\begin{aligned}
\left\| \max_{1\le i\le n} \left\| U_i \Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}} \right\|_{\mathrm{op}} \right\|_{\psi_1} &\le \left\| \max_{1\le i\le n} \left\| \Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}} \right\|_{\mathrm{op}} \right\|_{\infty} \cdot \left\| \max_{1\le i\le n} |U_i| \right\|_{\psi_1} \\
&\lesssim C_\phi^2 \mathfrak{D}_{\lambda_e} \left\| \max_{1\le i\le n} |U_i| \right\|_{\psi_2} \\
&\lesssim C_\phi^2 \mathfrak{D}_{\lambda_e} \sqrt{\log n} \max_{1\le i\le n} \big\| |U_i| \big\|_{\psi_2} \\
&\lesssim C_\phi^2 \mathfrak{D}_{\lambda_e} \sqrt{\log n}\, \sigma_W,
\end{aligned}$$

where the third inequality is obtained by Lemma 2.2.2 in van der Vaart and Wellner (1996) and

the last inequality is due to Condition 4.3.4. Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}^*\big\{U_i\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\big\}^2 = \frac{1}{n}\sum_{i=1}^{n}\|\Phi_{\lambda_e}(X_i)\|_{\ell_2}^2\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}} \preceq \frac{C_\phi^2\mathfrak{D}_{\lambda_e}}{n}\sum_{i=1}^{n}\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}} =: \mathbf{V}_3.$$

It is obvious that $\mathrm{tr}(\mathbf{V}_3) = C_\phi^2\mathfrak{D}_{\lambda_e}\sum_{i=1}^{n}\|\Phi_{\lambda_e}(X_i)\|_{\ell_2}^2/n \le C_\phi^4\mathfrak{D}_{\lambda_e}^2$. Moreover, conditioned on $\mathscr{G}_5(t)$ defined in (D.99), $\|\mathbf{V}_3\|_{\mathrm{op}} \le 2C_\phi^2\mathfrak{D}_{\lambda_e}$ by (D.100). Note that $15\mathrm{tr}(\mathbf{V}_3)e^{-t-\log n}/(2C_\phi^2\mathfrak{D}_{\lambda_e}) \le e^{-t}$ by the sample size requirement. Thus, applying Lemma D.5.2 yields that, conditioned on $\mathscr{G}_5(t)$, the bound

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}U_i\Phi_{\lambda_e}(X_i)\Phi_{\lambda_e}(X_i)^{\mathrm{T}}\right\|_{\mathrm{op}} &\lesssim C_\phi\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{t+\log n}{n}} + C_\phi^2\mathfrak{D}_{\lambda_e}\sigma_W\frac{(\log n)^{1/2}(t+\log n)}{n}. \\ &\lesssim \max(C_\phi,\sigma_W)\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{t+\log n}{n}}\end{aligned}$$

holds with $\mathbb{P}^*$-probability at least $1-e^{-t}$, where the last inequality follows from the sample size requirement. Combining this bound with (D.102) and (D.103) yields that, conditioned on $\mathscr{G}_5(t)$, it holds with $\mathbb{P}^*$-probability at least $1-e^{-t}$ that

$$D_3 \le C_5 r_n\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{t+\log n}{n}}, \tag{D.104}$$

where $C_5 = C_5(C_\phi,\sigma_W)$.

STEP IV. CONCLUDING THE PROOF. Let $\mathscr{G}(t) = \mathscr{G}_1(t)\cap\mathscr{G}_2(t)\cap\mathscr{G}_3(t)\cap\mathscr{G}_4(t)\cap\mathscr{G}_5(t)$, which satisfies $\mathbb{P}\{\mathscr{G}(t)\} \ge 1-12e^{-t}$. Recall that $\zeta = 4C_\phi\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{(t+\log n)/n} \le 1/2$ under the sample size requirement. Combining the decomposition (D.91) with the bounds (D.96), (D.101) and (D.104), and recalling the definition of r_n in (D.89), it follows that, conditioned on the event

$\mathscr{G}(t)$, the bounds

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} U_i\{Z_i(\widehat{f}) - \tau\widehat{g}(X_i)\}\widehat{T}_{\lambda_e}^{-1}K_{X_i} - \frac{1}{n}\sum_{i=1}^{n} U_i\omega_i T_{\lambda_e}^{-1}K_{X_i} \right\|_{\lambda_e} \\
&\lesssim \mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{t+\log n}{n}}\sqrt{\frac{\mathfrak{D}_{\lambda_e}+t}{n}} + \delta_n\mathfrak{D}_{\lambda_q}^{1/2}\sqrt{\frac{t+\mathfrak{D}_{\lambda_e}}{n}} \\
&\quad + (\gamma_n + \delta_{\mathrm{s}} + \mathfrak{D}_{\lambda_q}^{1/2}\delta_n^2)\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{t+\log n}{n}} \\
&\leq C_6\left(\gamma_n\mathfrak{D}_{\lambda_e}^{1/2}\sqrt{\frac{t+\log n}{n}} + \delta_{\mathrm{s}} + \mathfrak{D}_{\lambda_q}^{1/2}\delta_n^2\right)
\end{aligned}$$

hold with $\mathbb{P}^*$-probability at least $1-5e^{-t}$, where $C_6 = C_6(C_\phi, \sigma_0, \sigma_W) > 0$ and the last inequality follows by the sample size requirement $n \geq 64C_\phi^2\mathfrak{D}_{\lambda_e}(t+\log n)$. This establishes the claim. $\square$

Bibliography

ACERBI, C. and TASCHE, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance* **26** 1487–1503.

ANTHONY, M. and BARTLETT, P. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge Univ. Press, Cambridge.

ARGYRIOU, A., EVGENIOU, T. and PONTIL, M. (2008). Convex multi-task feature learning. *Mach. Learn.* **73** 243–272.

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404.

ARTZNER, P., DELBAEN, F., EBER, J.-M. and HEATH, D. (1997). Thinking coherently. *RISK* **10** 68–71.

AVELLA-MEDINA, M. (2021). Privacy-preserving parametric inference: A case for robust statistics. *J. Amer. Statist. Assoc.* **116** 969–983.

AVELLA-MEDINA, M., BRADSHAW, C. and LOH, P.-L. (2023). Differentially private inference via noisy optimization. *Ann. Statist.*, in press. *arXiv preprint arXiv:2103.11003*.

BARBER, R. F. and DUCHI, J. (2014). Privacy: A few definitional aspects and consequences for minimax mean-squared error. In *53rd IEEE Conference on Decision and Control* 1365–1369.

BARENDSE, S. (2020). Efficiently weighted estimation of the tail and interquantile expectations. *Prepreint*. http://dx.doi.org/10.2139/ssrn.2937665.

BARTLETT, P. L., HARVEY, N., LIAW, C. and MEHRABIAN, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.* **20** 2285–2301.

BARZILAI, J. and BORWEIN, J. M. (1988). Two-point step size gradient methods. *IMA J. Numer. Anal.* **8** 141–148.

BASEL COMMITTEE. (2019). *Minimum Capital Requirements for Market Risk. Technical Report*.

Bank for International Settlements.

BASSILY, R., SMITH, A. and THAKURTA, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science* 464–473.

BAUER, B. and KOHLER, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.* **47** 2261–2285.

BECKER, E. J., BERBERY, E. H. and HIGGINS, R. W. (2009). Understanding the characteristics of daily precipitation over the united states using the North American Regional Reanalysis. *J. Climate* **22** 6268–6286.

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and FERNÁNDEZ-VAL, I. (2019). Conditional quantile processes based on series or many regressors. *J. Econometrics* **213** 4–29.

BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford: Oxford Univ. Press.

BOUSQUET, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic Inequalities and Applications. Progress in Probability* **56** 213–247. Birkhäuser, Basel.

BUBECK, S., CESA-BIANCHI, N. and LUGOSI, G. (2013). Bandits with heavy tail. *IEEE Trans. Inform. Theory* **59** 7711–7717.

BUN, M. and STEINKE, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference* 635–658.

BUN, M. and STEINKE, T. (2019). Average-case averages: Private algorithms for smooth sensitivity and mean estimation. In *Advances in Neural Information Processing Systems* 181–191.

BURER, S. and MONTEIRO, R. D. (2003). A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Program.* **93** 329–357.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

CAI, J.-F., CANDÈS, E. J. and SHEN, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20** 1956–1982.

CAI, T. T., WANG, Y. and ZHANG, L. (2020). The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. *arXiv preprint arXiv:2011.03900.*

CAI, T. T., WANG, Y. and ZHANG, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *Ann. Statist.* **49** 2825–2850.

CAI, Z. and WANG, X. (2008). Nonparametric estimation of conditional VaR and expected shortfall. *J. Econom.* **147** 120–130.

CANDÈS, E. J. and PLAN, Y. (2009). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

CANDÈS, E. J. and PLAN, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory* **57** 2342–2359.

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772.

CAPONNETTO, A. and DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7** 331–368.

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185.

CAYAN, D. R., REDMOND, K. T. and RIDDLE, L, G. (1999). ENSO and hydrologic extremes in the western United States. *J. Climate* **12** 2881–2893.

CHAUDHURI, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Ann. Statist.* **19** 760–777.

CHEN, J., LIU, D. and LI, X. (2020). Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Trans. Inform. Theory* **66** 5806–5841.

CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960.

CHEN, X. and ZHOU, W.-X. (2020). Robust inference via multiplier bootstrap. *Ann. Statist.* **48** 1665–1691.

CHEN, Y., JALALI, A., SANGHAVI, S. and CARAMANIS, C. (2013). Low-rank matrix recovery from errors and erasures. *IEEE Trans. Inform. Theory* **59** 4324–4337.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural

parameters. *Econom. J.* **21** C1–C68.

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597.

CHERNOZHUKOV, V., CHETVERIKOV, D. and KOIKE, Y. (2023). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *Ann. Appl. Probab.* **33** 2374–2425.

CHETVERIKOV, D., LIU, Y. and TSYVINSKI, A. (2022). Weighted-average quantile regression. *arXiv preprint arXiv:2203.03032*.

CLÉMENÇON, S., BELLET, A. and COLIN, I. (2016). Scaling-up empirical risk minimization: Optimization of incomplete U-statistics. *J. Mach. Learn. Res.* **17** 1–36.

CONWAY, J. B. (1990). *A Course in Functional Analysis, 2nd ed.* Springer, New York.

CORONESE, M., LAMPERTI, F., KELLER, K., CHIAROMONTE, F. and ROVENTINI, A. (2019). Evidence for sharp increase in the economic damages of extreme natural disasters. In *Proceedings of the National Academy of Sciences* **116** 21450–21455.

DALALYAN, A. and THOMPSON, P. (2019). Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. In *Advances in Neural Information Processing Systems* **32** 13188–13198.

DEPERSIN, J. and LECUÉ, G. (2022a). Robust sub-Gaussian estimation of a mean vector in nearly linear time. *Ann. Statist.* **50** 511–536.

DEPERSIN, J. and LECUÉ, G. (2022b). Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. *Probab. Theory Relat. Fields* **183** 997–1025.

DETTINGER, M. D., CAYAN, D. R., DIAZ, H. F. and MEKO, D. M. (1998). Precipitation patterns in western North America on interannual-to-decadal timescales. *J. Climate* **11** 3095–3111.

DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725.

DIAKONIKOLAS, I. and KANE, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

DIMITRIADIS, T. and BAYER, S. (2019). A joint quantile and expected shortfall regression framework. *Electron. J. Stat.* **13** 1823–1871.

DONG, J., ROTH, A. and SU, W. J. (2022). Gaussian differential privacy. *J. R. Stat. Soc. Series*

B. **84** 3–54.

DUCHI, J. C., JORDAN, M. I. and WAINWRIGHT, M. J. (2018). Minimax optimal procedures for locally private estimation. *J. Amer. Statist. Assoc.* **113** 182–201.

DWORK, C., KENTHAPADI, K., MCSHEERY, F., MIRONOV, I. and NAOR, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques* 486–503.

DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference* 265–284.

DWORK, C. and ROTH, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9** 211–407.

DWORK, C. and ROTHBLUM, G. N. (2016). Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.

ELSENER, A. and VAN DE GEER, S. (2018). Robust low-rank matrix estimation. *Ann. Statist.* **46** 3481–3509.

ENGL, H. W., HANKE, M. and NEUBAUER, A. (1996). *Regularization of Inverse Problems* Kluwer Academic, Dordrecht.

FAN, J., GU, Y. and ZHOU, W.-X. (2022). How do noise tails impact on deep ReLU networks? *arXiv preprint arXiv:2203.10418*.

FAN, J., KE, Y., SUN, Q. and ZHOU, W.-X. (2019). FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. *J. Amer. Statist. Assoc.* **114** 1880–1893.

FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265.

FAN, J., LIU, H., SUN, Q. and ZHANG, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Ann. Statist.* **46** 1348–1360.

FAN, J., WANG, W. and ZHU, Z. (2021). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Ann. Statist.* **49** 1239–1266.

FANG, X. and KOIKE, Y. (2021). High-dimensional central limit theorems by Stein's method. *Ann. Appl. Probab.* **31** 1660–1686.

FARRELL, M. H., LIANG, T. and MISRA, S. (2021). Deep neural networks for estimation and

inference. *Econometrica* **89** 181–213.

FISSLER, T. and ZIEGEL, J. F. (2016). Higher order elicitability and Osband's principle. *Ann. Statist.* **44** 1680–1707.

FISSLER, T., MERZ, M. and WÜTHRICH, M. V. (2023). Deep quantile and deep composite triplet regression. *Insurance Math. Econom.* **109** 94–112

FOSTER, D. J. and SYRGKANIS, V. (2023). Orthogonal statistical learning. *Ann. Statist.* **51** 879–908.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762.

GOLDBERG, D., NICHOLS, D., OKI, B. M. and TERRY, D. (1992). Using collaborative filtering to weave an information tapestry. *Comm. ACM* **35** 61–70.

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. MIT Press, Cambridge.

GROSS, D., LIU, Y.-K., FLAMMIA, S. T., BECKER, S. and EISERT, J. (2010). Quantum state tomography via compressed sensing. *Phys. Rev. Lett.* **105** 150401.

GU, C. (2013). *Smoothing Spline ANOVA Models, 2nd Ed.* Springer, New York.

GUILLEN, M., BERMÚDEZ, L. and PITARUQE, A. (2021). Joint generalized quantile and conditional tail expectation regression for insurance risk analysis. *Insurance Math. Econom.* **99** 1–8.

GYÖRFI, L., KOHLER, M., KRZYZAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York.

HALL, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20** 675–694.

HAMPEL, F., HENNIG, C. and RONCHETTI, E. (2011). A smoothing principle for the Huber and other location M-estimators. *Comput. Stat. Data Anal.* **55** 324–337.

HAN, Q. and WELLNER, J. A. (2018). Robustness of shape-restricted regression estimators: An envelope perspective. *arXiv preprint arXiv:1805.02542*.

HAN, Q. and WELLNER, J. A. (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *Ann. Statist.* **47** 2286–2319.

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178.

HE, X., HSU, Y.-H. and HU, M. (2010). Detection of treatment effects by covariate-adjusted expected shortfall. *Ann. Appl. Stat.* **4** 2114–2125.

HE, X., PAN, X., TAN, K. M. and ZHOU, W.-X. (2023). Smoothed quantile regression with large-scale inference. *J. Econom.* **232** 367–388.

HE, X., TAN, K. M. and ZHOU, W.-X. (2023). Robust estimation and inference for expected shortfall regression with many regressors. *J. Roy. Statist. Soc. Ser. B* **85** 1223–1246.

HEYDE, C. (1967). On the influence of moments on the rate of convergence to the normal distribution. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **8** 12–18.

HOPKINS, S. B. (2022). Mean estimation with sub-Gaussian rates in polynomial time. *Ann. Statist.* **48** 1193–1213.

HOPKINS, S. B., LI, J. and ZHANG, F. (2020). Robust and heavy-tailed mean estimation made simple, via regret minimization. In *Advances in Neural Information Processing Systems* **33** 11902–11912.

HOPKINS, S. B., KAMATH, G. and MAJID, M. (2022). Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing* 1406–1417.

HSU, D., KAKADE, S. M. and ZHANG, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **52** 1–6.

HUANG, A., VEGA-WESTHOFF, B. and SRIVER, R. L. (2019). Analyzing El Niño-Southern Oscillation predictability using long-short-term-memory models. *Earth and Space Science* **6** 212–221.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821.

IMAIZUMI, M. (2023). Sup-norm convergence of deep neural network estimator for nonparametric regression by adversarial training. *arXiv preprint arXiv:2307.04042*.

IZENMAN, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.* **5** 248–264.

JIAO, Y., SHEN, G., LIN, Y. and HUANG, J. (2023). Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *Ann. Statist.* **51** 691–716.

JOSE, D. M., VINCENT, A. M. and DWARAKISH, G. S. (2022). Improving multiple model ensemble predictions of daily precipitation and temperature through machine learning techniques. *Sci. Rep.* **12** 4678.

KAHYA, E. and DRACUP, J. A. (1993). U.S. streamflow patterns in relation to the El Niño/Southern Oscillation. *Water Resour. Res.* **29** 2491–2503.

KAMATH, G., LI, J., SINGHAL, V. and ULLMAN, J. (2019). Privately learning high-dimensional distributions. In *Conference on Learning Theory* **99** 1853–1902.

KAMATH, G., MOUZAKIS, A. and SINGHAL, V. (2022). New lower bounds for private estimation and a generalized fingerprinting lemma. In *Advances in Neural Information Processing Systems* **35** 24405–24418.

KAMATH, G., SINGHAL, V. and ULLMA, J. (2020). Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory* **125** 2204–2235.

KARWA, V. and VADHAN, S. (2018). Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference* **44** 1–9.

KARWA, V. and VADHAN, S. (2017). Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908.*

KATO, K. (2012). Weighted Nadaraya–Watson estimation of conditional expected shortfall. *Journal of Financial Econometrics* **10** 265–291.

KE, Y., MINSKER, S., REN, Z., SUN, Q. and ZHOU, W.-X. (2019). User-friendly covariance estimation for heavy-tailed distributions. *Statist. Sci.* **34** 454–471.

KIMDELDORF, G. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95.

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

KLOCHKOV, Y. and ZHIVOTOVSKIY, N. (2020). Uniform Hanson-Wright type concentration inequalities for unbounded entries via the entropy method. *Electron. J. Probab.* **25** 1–30.

KLOPP, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20** 282–303.

KLOPP, O., LOUNICI, K. and TSYBAKOV, A. B. (2017). Robust matrix completion. *Probab. Theory Relat. Fields* **169** 523–564.

KNIGHT, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist.* **26** 755–770.

KOENKER, R. (2005). *Quantile Regression.* Cambridge University Press, Cambridge.

KOENKER, R. and BASSETT, G. (1978). Regression quantiles. *Econometrica* **46** 33–50.

KOENKER, R., CHERNOZHUKOV, V., HE, X. and PENG, L. (2017). *Handbook of Quantile Regression.* CRC Press, New York.

KOHLER, M. and LANGER, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *Ann. Statist.* **49** 2231–2249.

KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329.

KOLTCHINSKII, V. and YUAN, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38** 3660–3695.

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60** 84–90.

KUCHIBHOTLA, A. K. and PATRA, R. K. (2022). On least squares estimation under heteroscedastic and heavy-tailed errors. *Ann. Statist.* **50** 277–302.

LANTZ, B. (2013). *Machine Learning with R.* Packt Publishing, U.K.

LECUÉ, G. and MENDELSON, S. (2013). Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*.

LEDOUX, M. and TALAGRAND M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes. Classics in Mathematics.* Berlin: Springer.

LI, X. (2013). Compressed sensing and matrix completion with constant proportion of corruptions. *Constr. Approx.* **37** 73–99.

LI, Y., LIU, Y. and ZHU, J. (2007). Quantile regression in reproducing kernel Hilbert spaces. *J. Amer. Statist. Assoc.* **102** 255–268.

LI, Y., MA, T. and ZHANG, H. (2018). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory* 2–47.

LIAN, H. (2022). Distributed learning of conditional quantiles in the reproducing kernel Hilbert space. In *Advances in Neural Information Processing Systems* **35** 11686–11696.

LIN, Y. and BROWN, L. D. (2004). Statistical properties of the method of regularization with periodic Gaussian reproducing kernel. *Ann. Statist.* **32** 1723–1743.

LINTON, O. and XIAO, Z. (2013). Estimation and inference about the expected shortfall for time series with infinite variance. *Econometric Theory* **29** 771–807.

LIU, X., KONG, W., KAKADE, S. and OH, S. (2021). Robust and differentially private mean estimation. In *Advances in Neural Information Processing Systems* **34** 3887–3901.

LIU, X., JAIN, P., KONG, W., OH, S. and SUGGALA, A. S. (2023). Near optimal private and robust linear regression. *arXiv preprint arXiv:2301.13273.*

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204.

LUAN, X., FANG, B., LIU, L., YANG, W.. and QIAN, J. (2014). Extracting sparse error of robust PCA for face recognition in the presence of varying illumination and occlusion. *Pattern Recognition* **47** 495–508.

LUGOSI, G. and MENDELSON, S. (2019a). Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** 783–794.

LUGOSI, G. and MENDELSON, S. (2019b). Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.* **19** 1145–1190.

LUGOSI, G. and MENDELSON, S. (2021). Robust multivariate mean estimation: The optimality of trimmed mean. *Ann. Statist.* **49** 393–410.

MA, C., WANG, K., CHI, Y. and CHEN, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Found. Comput. Math.* **20** 451–632.

MA, J. and FATTAHI, S. (2023). Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *J. Mach. Learn.*

Res. **24**(96): 1–84.

MARTINS-FILHO, C., YAO, F. and TORERO, M. (2018). Nonparametric estimation of conditional value-at-risk and expected shortfall based on extreme value theory. *Econometric Theory* **34** 23–67.

MATHIEU, T. (2022). Concentration study of M-estimators using the influence function. *Electron. J. Stat.* **16** 3695–3750.

MCNEIL, A. J., FREY, R. and EMBRECHTS, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools. 2nd Ed.* Princeton University Press, Princeton.

MCSHERRY, F. and TALWAR, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science* 94–103.

MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999.

MENDELSON, S. and NEEMAN, J. (2010). Regularization in kernel learning. *Ann. Statist.* **38** 526–565.

MENDELSON, S. and ZHIVOTOVSKIY, N. (2020). Robust covariance estimation under L_4–L_2 norm equivalence. *Ann. Statist.* **48** 1648–1664.

MINH, H. Q. (2010). Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constr. Approx.* **32** 307–338.

MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335.

MINSKER, S. (2017). On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters* **127** 111–119.

MINSKER, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903.

MINSKER, S. and WEI, X. (2020). Robust modifications of U-statistics and applications to covariance estimation problems. *Bernoulli* **26** 694–727.

MIRONOV, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)* 263–275.

MO, K. C. and HIGGINS, R. W. (1998). Tropical influences on California precipitation. *J. Climate* **11** 412–430.

MURTAGH, J. and VADHAN, S. (2016). The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference* 157–175.

NAZAROV, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis. Lecture Notes in Math.* **1807** 169–187. Berlin: Springer.

NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097.

NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13** 1665–1697.

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statisti. Sci.* **27** 538–557.

OLMA, T. (2021). Nonparametric estimation of truncated conditional expectation functions. *arXiv preprint arXiv:2109.06150.*

OTTER, D. W., MEDINA, J. R. and KALITA, J. K. (2021). A survey of the usages of deep learning for natural language processing. In *IEEE Transactions on Neural Networks and Learning Systems* **32** 604–624.

PADILLA, O, H. M. and CHATTERJEE, S. (2022). Risk bounds for quantile trend filtering. *Biometrika* **109** 751–768.

PADILLA, O. H. M., TANSEY, W. and CHEN, Y. (2022). Quantile regression with ReLU Networks: Estimators and minimax rates. *J. Mach. Learn. Res.* **23** 1–42.

PAN, X. and ZHOU, W.-X. (2021). Multiplier bootstrap for quantile regression: non-asymptotic theory under random design. *Information and Inference: A Journal of the IMA* **10** 813–861.

PATTON, A. J., ZIEGEL, J. F. and CHEN, R. (2019). Dynamic semiparametric models for expected shortfall (and Value-at-Risk). *J. Econometrics.* **211** 388–413.

PENG, X. and WANG, H. J. (2023). Inference for joint quantile and expected shortfall regression. *Stat* **12** e619.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13** 389–427.

RECHT, B., FAZEL, M. and PARRILO, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52** 471–501.

ROCKAFELLAR, R. T. and ROYSET, J. O. (2010). On buffered failure probability in design and optimization of structures. *Reliability Engineering & System Safety* **95** 499–510.

ROCKAFELLAR, R. T., URYASEV, S. and ZABARANKIN, M. (2008). Risk tuning with generalized linear regression. *Math. Oper. Res.* **33** 712–729.

ROCKAFELLAR, R. T. and URYASEV, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk* **2** 21–42.

ROCKAFELLAR, R. T. and URYASEV, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance* **26** 1443–1471.

ROHDE, A. and STEINBERGER, L. (2020). Geometrizing rates of convergence under local differential privacy constraints. *Ann. Statist.* **48** 2646–2670.

ROHDE, A. and TSYBAKOV, A. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930.

ROPELEWSKI, C. F. and HALPERT, M. S. (1986). North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Mon. Wea. Rev.* **114** 2352–2362.

ROPELEWSKI, C. F. and HALPERT, M. S. (1996). Quantifying southern oscillation-precipitation relationships. *J. Climate* **9** 1043–1059.

SAUNDERS, C., GAMMERMAN, A. and VOVK, V. (1998). Ridge regression learning algorithm in dual variables. In *ICML '98* 515–521.

SCAILLET, O. (2005). Nonparametric estimation of conditional expected shortfall. *Revue Assurances et Gestion des Risques/Insurance and Risk Management Journal* **74** 639–660.

SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.* **48** 1875–1897.

SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

SHANG, Z. and CHENG, G. (2013). Local and global asymptotic inference in smoothing spline models. *Ann. Statist.* **41** 2608–2636.

SHE, Y. and CHEN, K. (2017). Robust reduced-rank regression. *Biometrika* **104** 633–647.

SHEN, G., JIAO, Y., LIN, Y., HOROWITZ, J. L. and HUANG, J. (2021). Deep quantile regression: Mitigating the curse of dimensionality through composition. *arXiv preprint arXiv:2107.04907*.

SHEN, Y., LI, J., CAI, J. and XIA, D. (2022). Computationally efficient and statistically optimal robust low-rank matrix estimation. *arXiv:2203.009533*.

SHEVTSOVA, I. G. (2014). On the absolute constants in the Berry-Esseen-type inequalities. *Dokl. Math.* **89** 378–381.

SILVERMAN, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Ann. Statist.* **12** 898–916.

SINGH, R. and VIJAYKUMAR, V. (2023). Kernel ridge regression inference. *arXiv preprint arXiv:2302.06578*.

SLIVINSKI, L. C., COMPO, G. P., WHITAKER, J. S., SARDESHMUKH, P. D., GIESE, B. S., MCCOLL, C., ALLAN, R., YIN, X., VOSE, R. and TITCHNER, H. (2019). Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Q. J. R. Meteorol. Soc.* **145** 2876–2908.

SMOLA, A. J., SCHÖLKOPF, B. and MÜLLER, K. (1998) The connection between regularization operators and support vector kernels. *Neural Networks* **11** 637–649

SONG, S., CHAUDHURI, K. and SARWATE, A. D. (2013). Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing* 245–248.

SPOKOINY, V. and ZHILOVA, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.* **43** 2653–2675.

SREBRO, N. and SALAKHUTDINOV, R. R. (2010). Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems* **23** 1329–1336.

STEINWART, I. and CHRISTMANN, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli* **17** 211–225.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

SUN, Q., ZHOU, W.-X. and FAN, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265.

SUZUKI, T. and SUGIYAMA, M. (2013). Fast learning rate of multiple kernel learning: Trade-off

between sparsity and smoothness. *Ann. Statist.* **41** 1381–1405.

TAKEUCHI, I., LE, Q. V., SEARS, T. D. and SMOLA, A. J. (2006). Nonparametric quantile estimation. *J. Mach. Learn. Res.* **7** 1231–1264.

TAN, K. M., SUN, Q. and WITTEN, D. (2023). Sparse reduced rank Huber regression in high dimensions. *J. Amer. Statist. Assoc.*, **118** 2383–2393.

TAYLOR, J. W. (2019). Forecasting Value at Risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *J. Bus. Econom. Statist.* **37** 121–133.

TERTI, G., RUIN, I., GOURLEY, J. J., KIRSTETTER, P., FLAMIG, Z., BLANCHET, J., ARTHUR, A. and Anquetin, S. (2019). Toward probabilistic prediction of flash flood human impacts. *Risk Analysis* **39** 140–161.

THOMPSON, P. (2020). Outlier-robust sparse/low-rank least-squares regression and robust matrix completion. *arXiv:2012.06750*.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288.

TONG, T., MA, C. and CHI, Y. (2021). Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *J. Mach. Learn. Res.* **22** 1–63.

TREFETHEN, L. N. and BAU III, D. (1997). *Numerical Linear Algebra.* Philadelphia: Society for Industrial and Applied Mathematics.

TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434.

TROPP, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.* **8** 1–230.

TYURIN, I. S. (2011). On the convergence rate in Lyapunov's theorem. *Theory Probab. Appl.* **55** 253–270.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

VERSHYNIN, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge: Cambridge Univ. Press.

WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cam-*

bridge Series in Statistical and Probabilistic Mathematics **48**. Cambridge: Cambridge Univ. Press.

WANG, B. and FAN, J. (2022). Robust matrix completion with heavy-tailed noise. *arXiv:2206.04276*.

WANG, L, AMMONS, S., HUR, V. M., SRIVER, R. L and ZHAO, Z. (2023). Convolutional GRU network for seasonal prediction of the El Niño-Southern Oscillation. *arXiv preprint arXiv:2306.10443*.

WANG, Y., KIFER, D. and LEE, J. (2019). Differentially private confidence intervals for empirical risk minimization. *J. Priv. Confid.* **9** 1–36.

WASSERMAN, L. and ZHOU, S. (2010). A statistical framework for differential privacy. *J. Amer. Statist. Assoc.* **105** 375–389.

WEI, K., CAI, J.-F., CHAN, T. F. and LEUNG, S. (2016). Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM J. Matrix Anal. Appl.* **37** 1198–1222.

WEI, X. and MINSKER, S. (2017). Estimation of the covariance structure of heavy-tailed distributions. In *Advances in Neural Information Processing Systems* **30** 2855–2864.

WILLIAMS, C. and SEEGER, M. (2000). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems* **13** 682–688.

YANG, Y., PILANCI, M. and WAINWRIGHT, M. J. (2017). Randomized sketches for kernels: Fast and optimal nonparametric regression. *Ann. Statist.* **45** 991–1023.

YU, M., REN, Z. and ZHOU, W.-X. (2023). Supplement to "Gaussian differentially private robust mean estimation and inference".

YU, M., SUN, Q. and ZHOU, W.-X. (2023). Supplement to "Low-rank matrix recovery under heavy-tailed errors".

YUN, K.-S., LEE, J.-Y., TIMMERMANN, A., STEIN, K., STUECKER, M. F., FYFE, J. C., and CHUNG, E.-S. (2021). Increasing ENSO–rainfall variability due to changes in future tropical temperature–rainfall relationship. *Commun. Earth Environ.* **2** 43.

ZHANG, C., LIU, Y. and WU, Y. (2016). On quantile regression in reproducing kernel Hilbert spaces with the data sparsity constraint. *J. Mach. Learn. Res.* **17**(40): 1–45.

ZHANG, J., FATTAHI, S. and ZHANG, R. (2021). Preconditioned gradient descent for over-parameterized nonconvex matrix factorization. In *Advances in Neural Information Processing Systems* **34** 5985–5996.

ZHANG, T. (2002). Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems* **15** 454–461.

ZHANG, Y., DUCHI, J. and WAINWRIGHT, M. J. (2013). Divide and conquer kernel ridge regression. In *Conference on Learning Theory* 592–617.

ZHAO, T., CHENG, G. and LIU, H. (2016). A partially linear framework for massive heterogeneous data. *Ann. Statist.* **44** 1400–1437.

ZHAO, S., LIU, R. and SHANG, Z. (2021). Inference on panel data models: A kernel ridge regression method. *J. Bus. Econom. Statist.* **39** 325–337.

ZHIVOTOVSKIY, N. (2024). Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electron. J. Probab.* **29** 1–28.

ZHONG, Q., MUELLER, J. and WANG, J.-L. (2022). Deep learning for the partially linear Cox model. *Ann. Statist.* **50** 1348–1375.

ZHONG, Q. and WANG, J.-L. (2023). Neural networks for partially linear quantile regression. *J. Bus. Econom. Statist.*, in press. https://doi.org/10.1080/07350015.2023.2208183.

ZHOU, W.-X., BOSE, K., FAN, J. and LIU, H. (2018). A new perspective on robust *M*-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.* **46** 1904–1931.