Bernoulli **30**(4), 2024, 3059–3088
https://doi.org/10.3150/23-BEJ1706

Gaussian differentially private robust mean estimation and inference

MYEONGHUN YU[1,a], ZHAO REN[2,b] and WEN-XIN ZHOU[3,c]

[1]*Department of Mathematics, University of California, San Diego, La Jolla, CA, 92093, USA,* [a]*myyu@ucsd.edu*
[2]*Department of Statistics, University of Pittsburgh, Pittsburgh, PA, 15260, USA,* [b]*zren@pitt.edu*
[3]*Department of Information and Decision Sciences, University of Illinois at Chicago, Chicago, IL, 60607, USA,* [c]*wenxinz@uic.edu*

In this paper, we propose differentially private algorithms for robust (multivariate) mean estimation and inference under heavy-tailed distributions, with a focus on Gaussian differential privacy. First, we provide a comprehensive analysis of the Huber mean estimator with increasing dimensions, including non-asymptotic deviation bound, Bahadur representation, and (uniform) Gaussian approximations. Secondly, we privatize the Huber mean estimator via noisy gradient descent, which is proven to achieve near-optimal statistical guarantees. The key is to characterize quantitatively the trade-off between statistical accuracy, degree of robustness and privacy level, governed by a carefully chosen robustification parameter. Finally, we construct private confidence intervals for the proposed estimator by incorporating a private and robust covariance estimator. Our findings are demonstrated by simulation studies.

Keywords: Confidence interval; differential privacy; heavy-tailed distribution; Huber loss; mean estimation

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, [6,11,12,15–17,28,30,38,40,41,44] for an unavoidably incomplete overview. For a more thorough review of robust mean estimation and beyond, we refer to the survey articles [18] and [39].

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 [11]: for any $z \geq 0$, $\|\bar{\boldsymbol{x}}_n - \boldsymbol{\mu}\|_2 \leq \sqrt{\operatorname{tr}(\Sigma)/n} + \sqrt{2\|\Sigma\|_2 \cdot z/n}$ 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 [11] 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 [38] and the trimmed mean estimator [40], 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 [28] and [15] 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 [41] 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 [41] demonstrates that

1350-7265 © 2024 ISI/BS

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 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 [22] 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) [19] 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, [8,32,35] considered optimal private mean estimation in terms of sample complexity under different differential privacy frameworks. Another work in statistics literature [10] 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, [34] 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, [37] 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. [29] 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 [11,41]. On the other hand, data privacy is typically guaranteed by randomly perturbing the output of non-private algorithms [22,42]. In particular, to privately release a non-private Huber-type robust estimator, inspired by [5,50], 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 [41] 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 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 [41] 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 [34], 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 [1–3,9,10,20,49,54,55]. For example, under the local differential privacy, a slightly stronger notion of DP, [55] revealed that existing private mechanisms lead to slower rates than the minimax rates, and [20,49] further derived new minimax rates and corresponding private algorithms for several models. [9,10] 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 [2], 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. 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 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 3. Section 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 are given in the Appendix. 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 [57].

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. 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 [41], 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 [11], and has also been employed in regression settings [25,58]. In Section 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 [41]. 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 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.

2.1. 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 $\mathbb{R}_+$. [41] provided a concentration study of the following M-estimator:

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

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 $\widehat{\mathcal{L}}_\tau(\cdot)$, 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, [41] considered three robust mean estimators that are determined by their corresponding score functions, which are

(i) (Huber's score) $\psi(u) = u\mathbb{1}(|u| \leq 1) + \mathrm{sign}(u)\mathbb{1}(|u| > 1)$;
(ii) (Catoni's score) $\psi(u) = \log(1 + u + u^2/2)\mathbb{1}(u \geq 0) - \log(1 - u + u^2/2)\mathbb{1}(u < 0)$;
(iii) (Polynomial score) For $p \geq 1$, $\psi(u) = \frac{u}{1+u^{1-1/p}}\mathbb{1}(u \geq 0) - \frac{u}{1+(-u)^{1-1/p}}\mathbb{1}(u < 0)$.

As demonstrated in [41], these three robust estimators exhibit similar theoretical and numerical performance. Therefore, we restrict attention to Huber's estimator [31]. 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 [26]. See, for example, Examples 1 and 2 in [2].

Theorem 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. Throughout the rest, we write

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

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

Theorem 2.1. *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 (1) with $\tau \gtrsim \sqrt{\mathrm{tr}(\Sigma)}$ satisfies the bound*

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

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

$$b_\tau := \left\| \mathbb{E}\left\{ \frac{\psi_\tau(\|\boldsymbol{x} - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x} - \boldsymbol{\mu}\|_2}(\boldsymbol{x} - \boldsymbol{\mu}) \right\} \right\|_2 \leq \frac{\sqrt{\bar{\lambda}}\mathrm{tr}(\Sigma)}{\tau}. \tag{3}$$

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) 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 $q \geq 2$, we define

$$\nu_q = \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \frac{\mathbb{E}|\langle \boldsymbol{x} - \boldsymbol{\mu}, \boldsymbol{u}\rangle|^q}{(\mathbb{E}\langle \boldsymbol{x} - \boldsymbol{\mu}, \boldsymbol{u}\rangle^2)^{q/2}} \quad \text{and} \quad \kappa_q = \max_{1 \leq k \leq d} \frac{\mathbb{E}|x_k - \mu_k|^q}{\{\mathbb{E}(x_k - \mu_k)^2\}^{q/2}}. \tag{4}$$

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

$$m_q = \mathbb{E}\left\{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^{q-2}\sum_{k=1}^{d}(x_k-\mu_k)^2\right\}$$
$$\le \sum_{k=1}^{d}(\mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^q)^{1-2/q}(\mathbb{E}|x_k-\mu_k|^q)^{2/q} \le m_q^{1-2/q}\cdot \kappa_q^{2/q}\operatorname{tr}(\Sigma), \tag{5}$$

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

Lemma 2.2. *Assume that there exists some $q \ge 2$ such that $m_q = \mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^q$ is finite. Then, the bias term b_τ satisfies*

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

Remark 1. By combining Lemma 2.2 and Theorem 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 $\tau \asymp \bar{\lambda}^{1/4}\operatorname{tr}(\Sigma)^{1/4}(n/z)^{1/2}$, which leads to the bound

$$\|\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\|_2 \lesssim \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \bar{\lambda}^{1/2}\mathrm{r}(\Sigma)^{1/4}\sqrt{\frac{z}{n}} \tag{6}$$

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 [41] with $q = 2$ and $\varepsilon_n = 0$. In detail, Proposition 2 of [41] establishes that the Huber estimator $\widehat{\boldsymbol{\mu}}_\tau$ with $\tau \asymp \operatorname{tr}(\Sigma)^{1/2}(n/z)^{1/2}$ satisfies

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

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 [41], refining it to $\mathrm{r}(\Sigma)^{1/4}$.

Yet, the deviation bound (6) is still sub-optimal in terms of its dependence on $\bar{\lambda}$, $\operatorname{tr}(\Sigma)$ 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 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

$$\|\widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\|_2 \lesssim \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + m_q^{1/q}\left(\frac{z}{n}\right)^{1-1/q} \tag{7}$$

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 2.1 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 2.3. *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*

$$\left\| \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu} - \frac{1}{n}\sum_{i=1}^n \frac{\psi_\tau(\|\boldsymbol{x}_i-\boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i-\boldsymbol{\mu}\|_2}(\boldsymbol{x}_i-\boldsymbol{\mu}) \right\|_2 \lesssim \left\{ \bar{\lambda}^{1/2}\sqrt{\frac{\mathrm{r}(\Sigma)+z}{n}} + \frac{\tau z}{n} + b_\tau \right\} \left(\frac{m_q}{\tau^q} + \sqrt{\frac{z}{n}} \right) \tag{8}$$

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

Theorem 2.3 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 (8), in Section 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 [43,45].

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 $q \ge 3$.

Theorem 2.4 below provides a Berry-Esseen bound for all (deterministic) linear combinations of $\widehat{\boldsymbol{\mu}}_\tau$, from which the asymptotic normality immediately follows.

Theorem 2.4. *Assume that $m_q = \mathbb{E}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^q < \infty$ for some $q \ge 3$, 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*

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d,\, x\in\mathbb{R}} \left| \mathbb{P}\big(\sqrt{n}\langle \boldsymbol{u}/\|\boldsymbol{u}\|_\Sigma, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle \le x\big) - \Phi(x) \right| \lesssim \frac{m_q^{1/q}}{\underline{\lambda}^{1/2}}\frac{\log n}{\sqrt{n}} + v_q^{2/q}\left(\frac{\log n}{n}\right)^{(q-2)/(q-1)} + \frac{v_3}{\sqrt{n}}, \tag{9}$$

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. In Theorem 2.4, we require $\boldsymbol{x}$ to have at least the finite third moment so that the upper bound in (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 [27] for details. To achieve the optimal $n^{-1/2}$-rate, Theorem 2.4 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 (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 ($q \geq 3$) is finite, the two parameters ν_q and κ_q defined in (4) are essentially dimension-free. Using the inequality $m_q^{1/q} \leq \kappa_q^{1/q}\mathrm{tr}(\Sigma)^{1/2}$ from (5), we can substitute this bound into (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 (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 $\mathrm{r}(\Sigma) \leq d$, 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 [24] and [36]. Given a robustification parameter $\xi > 0$, the U-type covariance estimator $\widehat{\Sigma}_\xi$ is defined as

$$\widehat{\Sigma}_\xi = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \psi_\xi\left(\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\right) \frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}, \tag{10}$$

where $\psi_\xi(t) = \xi\psi(t/\xi)$. By choosing $\delta = e^{-z}$ in Theorem 3.2 of [36] 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 2.1 (Theorem 3.2 in [36]). *Assume $\boldsymbol{x} \in \mathbb{R}^d$ has bounded fourth moment, and write*

$$v_0^2 := \frac{1}{4}\|\mathbb{E}\{(\boldsymbol{x}_1 - \boldsymbol{x}_2)(\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\mathrm{T}}\}^2\|_2. \tag{11}$$

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 (10) *with $\xi = v_0\sqrt{n_0/\{\log(2d) + z\}}$ satisfies*

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

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

Remark 3. To compute $\widehat{\Sigma}_\xi$, the major barrier is due to the U-statistics structure of (10), in which the sum consists of $O(n^2)$ terms. [14] 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

$$\widetilde{\Sigma}_\xi = \frac{1}{n}\sum_{i=1}^{n} \frac{\psi_\xi(\|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}\|_2^2)}{\|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}\|_2^2}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^{\mathrm{T}}, \tag{12}$$

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 [56], 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

$$\|\widetilde{\Sigma}_\xi - \Sigma\|_2 \lesssim \sigma_0\sqrt{\frac{z + \log d}{n}} \tag{13}$$

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 [46]), a sufficient sample size requirement for (13) is $n \gtrsim \nu_4\,\mathrm{r}(\Sigma)(z + \log d)$. On the other hand, it follows from Theorem 2.1 and Lemma 2.2 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 (12).

The following result complements Theorem 2.4 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 2.5. *Assume $m_4 = \mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^4 < \infty$ and let the sample size satisfy $n \gtrsim \mathrm{r}(\Sigma) + \log n$. 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*

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d, x\in\mathbb{R}} \left|\mathbb{P}\left\{\sqrt{n}\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau - \boldsymbol{\mu}\rangle/(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_\xi\boldsymbol{u})^{1/2} \le x\right\} - \Phi(x)\right| \lesssim \nu_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)\log n}{n}}, \tag{14}$$

where $\widehat{\Sigma}_\xi$ is the U-type covariance estimator defined in (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 2.5 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 2.4.

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

$$\left[\langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau \rangle - z_{\alpha/2} \frac{(\boldsymbol{u}^{\mathrm{T}} \widehat{\Sigma}_\xi \boldsymbol{u})^{1/2}}{\sqrt{n}}, \langle \boldsymbol{u}, \widehat{\boldsymbol{\mu}}_\tau \rangle + z_{\alpha/2} \frac{(\boldsymbol{u}^{\mathrm{T}} \widehat{\Sigma}_\xi \boldsymbol{u})^{1/2}}{\sqrt{n}} \right], \tag{15}$$

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 2.6. *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*

$$\sup_{x \ge 0} \left| \mathbb{P}\left\{ \max_{1 \le k \le d} \left| \frac{\sqrt{n}(\widehat{\mu}_k - \mu_k)}{\sqrt{\widehat{\sigma}_{kk}}} \right| \le x \right\} - \mathbb{P}(\|\boldsymbol{G}\|_\infty \le x) \right| \lesssim v_4^{1/2} \frac{\bar{\lambda}}{\underline{\lambda}} \log^2(d) \log(n) \sqrt{\frac{d}{n}}, \tag{16}$$

where $\widehat{\sigma}_{kk}$ is the k-th diagonal element of $\widehat{\Sigma}_\xi$ defined in (10) and $\xi \asymp v_0\sqrt{n/\log(nd)}$.

Based on Theorem 2.6, we construct the confidence set

$$\times_{k=1}^{d} \left[\widehat{\mu}_k - \omega_\alpha \sqrt{\frac{\widehat{\sigma}_{kk}}{n}}, \widehat{\mu}_k + \omega_\alpha \sqrt{\frac{\widehat{\sigma}_{kk}}{n}} \right] \tag{17}$$

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 $\widehat{\boldsymbol{G}} \sim \mathcal{N}(\boldsymbol{0}, \mathrm{corr}(\widehat{\Sigma}_\xi))$ via Monte Carlo simulations. Its validity (consistency) is guaranteed by Proposition 2.2 below as long as the right-hand side of the inequality is $o(1)$.

Proposition 2.2. *Assume $m_4 = \mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^4 < \infty$, and let*

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

where $\widehat{\Sigma} = \widehat{\Sigma}_\xi$ is the U-type covariance estimator defined in (10) with $\xi \asymp v_0\sqrt{n/\log(nd)}$. Then, with probability at least $1 - 2n^{-1}$, we have

$$\sup_{t \ge 0} \left| \mathbb{P}\left(\max_{1 \le k \le d} |\widehat{G}_k| \le t \,\middle|\, x_1, \ldots, x_n \right) - \mathbb{P}\left(\max_{1 \le k \le d} |G_k| \le t \right) \right| \\ \lesssim v_4^{1/2} (\bar{\lambda}/\underline{\lambda})^2 \log(d) \log(n) \sqrt{\frac{\mathrm{r}(\Sigma)\log(nd)}{n}}. \tag{18}$$

Remark 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 [13,51], 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 $\mathcal{D}_n := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ 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

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

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

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

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

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

Practically, the conditional quantiles of $\widehat{\mathcal{L}}_\tau^{\mathrm{b}}(\widehat{\boldsymbol{\mu}}_\tau) - \widehat{\mathcal{L}}_\tau^{\mathrm{b}}(\widehat{\boldsymbol{\mu}}_\tau^{\mathrm{b}})$ 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 $\widehat{\mathcal{L}}_\tau(\boldsymbol{\mu}) - \widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\mu}}_\tau)$ and theoretical analysis of the bootstrap estimators, is still needed, we leave a rigorous theoretical investigation and validation of this approach to future work.

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 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 [5] 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.

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 3.1 ([21,22]). A dataset $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n) \in \mathcal{X}^n$ consist of n data from some space $\mathcal{X}$. We say two datasets $\boldsymbol{X}$ and $\boldsymbol{X}'$ are neighbors if they differ by one entry. A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is said to be (ϵ,δ)-differentially private ((ϵ,δ)-DP) for $\epsilon, \delta > 0$ if for any neighboring datasets $\boldsymbol{X}$ and $\boldsymbol{X}'$, and any measurable set $E \subseteq \mathcal{Y}$,

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

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

$$H_0 : \text{the underlying dataset is } \boldsymbol{X} \quad \text{vs} \quad H_1 : \text{the underlying dataset is } \boldsymbol{X}'. \tag{19}$$

As revealed by [55], 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 $e^{\epsilon}\alpha + \delta$. 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, [19] further proposed and advocated a notion of Gaussian differential privacy (GDP). GDP has an attractive interpretation to statisticians: the testing problem (19), e.g., identifying whether an individual is in a dataset, is at least as difficult as distinguishing between $\mathcal{N}(0,1)$ and $\mathcal{N}(\epsilon,1)$ based on a single draw for some $\epsilon > 0$. 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 3.2 ([19]). Let M be a randomized algorithm. We say M is ϵ-Gaussian differentially private (GDP) if any α-level test ϕ for (19) has a power function

$$\beta(\alpha) \le 1 - \Phi(\Phi^{-1}(1-\alpha) - \epsilon)$$

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 $\Phi(\Phi^{-1}(1-\alpha) - \epsilon)$ describes the supreme of the type II errors of all α-level tests for distinguishing $\mathcal{N}(0,1)$ and $\mathcal{N}(\epsilon,1)$ based on a single draw, which is achieved by the likelihood ratio test. For formal proof, we refer to Appendix A in [19] 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 [19]. 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 [48]. Furthermore, the GDP mechanism preserves a transparent hypothesis testing interpretation, while other relaxations of the (ϵ,δ)-DP mechanism, including concentrated differential privacy (CDP) [7,23] and Rényi differential privacy [47], 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 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

$$\operatorname{sens}(\boldsymbol{h}) = \sup_{\boldsymbol{X},\boldsymbol{X}'} \|\boldsymbol{h}(\boldsymbol{X}) - \boldsymbol{h}(\boldsymbol{X}')\|_2, \tag{20}$$

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 [19] but the extension to general $d \geq 1$ is straightforward.

Lemma 3.3 (Theorem 1 in [19]). *Define the Gaussian mechanism that operates on a statistic $\boldsymbol{h} \in \mathbb{R}^d$ as*

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

where $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. 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 [48], 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 3.4 (Corollary 2 in [19]). *Let $M_1 : \mathcal{X}^n \to \mathcal{Y}_1$ be the first mechanism and $M_t : \mathcal{X}^n \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{t-1} \to \mathcal{Y}_t$ be the t-th mechanism for $t = 2, \ldots, k$. We define the k-fold composed mechanism $M : \mathcal{X}^n \to \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_k$ 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 $y_1 \in \mathcal{Y}_1, \ldots, y_{t-1} \in \mathcal{Y}_{t-1}$, then the k-fold composed mechanism M is $\sqrt{\epsilon_1^2 + \ldots + \epsilon_k^2}$-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 3.5 (Proposition 4 in [19]). *Let $M : \mathcal{X}^n \to \mathcal{Y}$ be ϵ-GDP. Denote a post-processing (randomized) algorithm $Proc : \mathcal{Y} \to \mathcal{Z}$ that maps the input $M(\boldsymbol{X})$ to some space $\mathcal{Z}$. Then the post-processing $Proc \circ M : \mathcal{X}^n \to \mathcal{Z}$ is also ϵ-GDP.*

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 (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 [5]. For a predetermined number of iterations T, it computes

$$\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} + \frac{\eta_0}{n}\sum_{i=1}^n \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)}) + 2T^{1/2}\tau\frac{\eta_0}{\epsilon n}\boldsymbol{g}_t \tag{21}$$

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 $\epsilon > 0$ 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 3.3-3.4.

Proposition 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 (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

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

Therefore, the sensitivity of $\boldsymbol{h}$ is upper bounded by $2\tau\eta_0/n$. By Lemma 3.3, adding a Gaussian noise $2T^{1/2}\tau\eta_0(\epsilon n)^{-1}\boldsymbol{g}_0$ to the gradient update makes this step $(T^{-1/2}\epsilon)$-GDP. Consequently, $\boldsymbol{\mu}^{(1)}$ is $(T^{-1/2}\epsilon)$-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 3.4 that the two-fold composed (joint) mechanism $(\boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(2)})$ is $\sqrt{\epsilon^2/T + \epsilon^2/T}$-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

$$\mathcal{E}_1 = \mathcal{E}_1(r_0,\chi) = \{\widehat{\boldsymbol{\mu}} \in \Theta(r_0/2)\} \cap \{\nabla^2\widehat{\mathcal{L}}_\tau(\boldsymbol{\theta}) \succeq (1-\chi)\mathbf{I}_d,\ \forall\boldsymbol{\theta} \in \Theta(r_0)\}, \tag{22}$$

where

$$\Theta(r) := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \boldsymbol{\mu}\|_2 \le r\} \text{ for every } r > 0, \tag{23}$$

and $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_\tau$ is the non-private robust estimator defined in (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 $\mathcal{E}_1$.

Theorem 3.6. *Consider the private estimate $\boldsymbol{\mu}^{(T)}$ obtained from noisy gradient descent (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 $T \ge 1$. Define the optimization error r_{opt} and the privacy error r_{p} as*

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

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

$$n \gtrsim T^{1/2}\tau\frac{\sqrt{d} + \sqrt{\log T + z}}{(1-\chi)\epsilon r_0}. \tag{24}$$

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

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

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

Theorem 3.6 provides a concentration bound with two terms: optimization error $r_{\mathrm{opt}}(T)$ and privacy error $r_{\mathrm{p}}(T)$. 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) 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 $r_{\mathrm{opt}}(T)$ and $r_{\mathrm{p}}(T)$, while also choosing τ to balance bias, robustness and privacy error.

Before selecting appropriate parameters in Theorem 3.6 to consider the trade-off between different sources of error and make the event $\mathcal{E}_1$ occur with high probability, we provide a few remarks regarding the assumption on the initial iterate $\boldsymbol{\mu}^{(0)}$. In Theorem 3.6, the minimum sample size required and the event $\mathcal{E}_1$ 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 $\mathcal{E}_1(r_0,\chi)$, 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, $\|\boldsymbol{\mu}^{(T_0)} - \boldsymbol{\mu}\|_2 \le r_0$.

Proposition 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 $T_0 \ge R_0^2/(\eta_0\Delta)$ and the sample size satisfy*

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

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 (21). Then, $\boldsymbol{\mu}^{(T_0)}$ satisfies $\widehat{\mathcal{L}}_\tau(\boldsymbol{\mu}^{(T_0)}) - \widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\mu}}) \le \Delta$ with probability (over

$\{\boldsymbol{g}_t\}_{t=0}^{T_0-1}$) *at least* $1-e^{-z}$. *In particular, conditioning on* $\mathcal{E}_1(r_0,\chi)$ *and taking* $\Delta = (1-\chi)r_0^2/8$, *we have*

$$\|\boldsymbol{\mu}^{(T_0)} - \boldsymbol{\mu}\|_2 \le r_0 \tag{25}$$

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 $\mathcal{E}_1(r_0,\chi)$ occurs with high probability.

Proposition 3.3. *Assume the same conditions as in Theorem 2.1. Moreover, for a given* $z>0$, *let* (r_0,χ,τ) *and* n *satisfy*

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

Then, the event $\mathcal{E}_1(r_0,\chi)$ *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 3.3 with Theorem 3.6 yields the following result.

Corollary 3.1. *Let* $\epsilon>0$ *be a predetermined privacy parameter. For any* $z>1$, *let the sample size satisfy*

$$n \gtrsim \max\left\{\mathrm{r}(\Sigma)+z, T^{1/2}\frac{\sqrt{d}+\sqrt{\log T + z}}{\epsilon}\right\} \tag{26}$$

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* (21) *with* $\eta_0 = 1$ *and* $T \asymp \log(n/z)$ *satisfies the bounds*

$$\|\boldsymbol{\mu}^{(T)} - \widehat{\boldsymbol{\mu}}\|_2 \lesssim \tau\frac{z}{n} + (d+z)^{1/2}(\log n)^{1/2}\frac{\tau}{\epsilon n} \tag{27}$$

and

$$\|\boldsymbol{\mu}^{(T)} - \boldsymbol{\mu}\|_2 \lesssim \bar{\lambda}^{1/2}\sqrt{\frac{\mathrm{r}(\Sigma)+z}{n}} + \tau\frac{z}{n} + b_\tau + (d+z)^{1/2}(\log n)^{1/2}\frac{\tau}{\epsilon n} \tag{28}$$

with probability at least $1-5e^{-z}$, *where* b_τ *is the bias term defined in* (3).

Remark 5. Taking $r_0=\tau/2$ in Proposition 3.2, we observe that even when the initial iterate $\boldsymbol{\mu}^{(0)}$ fails to meet the assumption of Corollary 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 $\mathcal{E}_1(r_0,\chi)$, 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 3.6, 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 3.6. Note that we require $T \asymp \log n$ in Corollary 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 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 3.2 reduces to

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

Given that we have $T_0 = O(1)$, the sample size requirement of Corollary 3.1 implies that the above inequality holds as long as n is sufficiently large. Therefore, Corollary 3.1 and Proposition 3.2 together ensure that the accuracy of the initial estimator does not significantly impact the algorithm's convergence.

Remark 6. From Corollary 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 [19], as summarized in Lemma 3.3. 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 (5). Taking $z = \log n$ and

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

employing Lemma 2.2 yields

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

with probability exceeding $1 - 5n^{-1}$. Comparing this result with the bound (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

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

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, [10] showed that the minimax ℓ_2 risk of sub-Gaussian mean estimation with (ϵ,δ)-differential privacy is at least $O(\sqrt{\frac{d}{n}} + \frac{d\log^{1/2}(1/\delta)}{\epsilon n})$, explicitly demonstrating its dependence on ϵ and δ. By Corollary 1 in [19], an algorithm is ϵ-GDP if and only if $(\epsilon,\delta(\epsilon))$-DP, where $\delta(\epsilon) = \Phi(-1+\epsilon/2) - e^{\epsilon}\Phi(-1-\epsilon/2)$. Consequently, the cost of privacy of sub-Gaussian mean estimation with ϵ-GDP is thus at least $O(\frac{d}{\epsilon n})$, 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 [53]. In this case, it can be shown from Corollary 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 $O(\sqrt{d/n} + (\frac{d}{\epsilon n})^{1-1/q})$ with high probability, ignoring the $\log(n)$-factor. The slower term $(\frac{d}{\epsilon n})^{1-1/q}$ characterizes the impact of heavy-tailedness and privacy. For $q = 2$, we find that this matches the lower bound on the ℓ_2-risk [33]. 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 [34]. 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 [34] employs Markov's inequality, resulting in a bound with a polynomial decay.

Combining the deviation bound (27) with Theorem 2.3, we obtain a non-asymptotic Bahadur representation for the ϵ-GDP Huber estimator $\boldsymbol{\mu}^{(T)}$ as stated below.

Corollary 3.2. *For any $z > 0$, assume that all the conditions in Corollary 3.1 hold. Then, the ϵ-GDP Huber estimator $\boldsymbol{\mu}^{(T)}$ satisfies*

$$\left\| \boldsymbol{\mu}^{(T)} - \boldsymbol{\mu} - \frac{1}{n}\sum_{i=1}^{n} \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2}(\boldsymbol{x}_i - \boldsymbol{\mu}) \right\|_2 \lesssim \left\{ \bar{\lambda}^{1/2}\sqrt{\frac{\mathrm{r}(\Sigma) + z}{n}} + \frac{\tau z}{n} + b_\tau \right\}\left(\frac{m_q}{\tau^q} + \sqrt{\frac{z}{n}} \right) + (d+z)^{1/2}(\log n)^{1/2}\frac{\tau}{\epsilon n} \tag{29}$$

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

Corollary 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 (29), in Section 3.3 we obtain a Gaussian approximation result for $\boldsymbol{\mu}^{(T)}$ under a bounded third or fourth moment condition.

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 $\epsilon \leq 1$.

Theorem 3.7. *Assume $m_q = \mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^q < \infty$ for some $q \geq 3$. Let the sample size satisfy (26) and $n \gtrsim \sqrt{(d + \log n)\log n}/\epsilon$ with $z = \log n$ and $\tau \asymp m_q^{1/q}\{\epsilon n/\sqrt{(d+\log n)\log n}\}^{1/q}$. 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*

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d,\, x\in\mathbb{R}} \left| \mathbb{P}(\sqrt{n}\langle \boldsymbol{u}/\|\boldsymbol{u}\|_\Sigma, \boldsymbol{\mu}^{(T)} - \boldsymbol{\mu}\rangle \leq x) - \Phi(x) \right| \lesssim \frac{m_q^{1/q}}{\underline{\lambda}^{1/2}}\left\{ \frac{\sqrt{(d+\log n)\log n}}{\epsilon} \right\}^{1-1/q}\left(\frac{1}{n}\right)^{1/2-1/q} + v_q^{2/q}\left\{ \frac{\sqrt{(d+\log n)\log n}}{\epsilon n} \right\}^{1-2/q}, \tag{30}$$

where v_q is defined in (4).

Remark 7. Since $m_q^{1/q} \le \kappa_q^{1/q} \mathrm{tr}(\Sigma)^{1/2}$, the first term on the right-hand side of (30) is further bounded, up to constants, by

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

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 $\epsilon^{-1}\sqrt{\mathrm{r}(\Sigma)(d+\log n)\log(n)/n}$ (as if $q=\infty$). Comparing this result with Theorem 2.4 for non-private robust estimator $\widehat{\boldsymbol{\mu}}$, the different choice of $\tau \asymp m_q^{1/q}\{\epsilon n/\sqrt{(d+\log n)\log n}\}^{1/q}$ 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 2.4, 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 cannot be directly applied. In the following, we introduce a differentially private counterpart of the robust covariance estimator $\widehat{\Sigma}_\xi$ given in (10).

Proposition 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* $\widehat{\Sigma}_\xi + \frac{4\xi}{\epsilon n}\mathbf{E}$ *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

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

By Lemma 3.3, $\boldsymbol{h}(\boldsymbol{X}_n) + \frac{4\xi}{\epsilon n}\boldsymbol{g}$ with $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$ is ϵ-GDP. Then it follows from Lemma 3.5 that $\widehat{\Sigma}_\xi + \frac{4\xi}{\epsilon n}\mathbf{E}$ is also ϵ-GDP. □

Remark 8. Based on Remark 3, we further consider a differentially private counterpart of the truncated covariance estimator $\widetilde{\Sigma}_\xi$ given in (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 3.1 and 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 $\widetilde{\Sigma}_\xi + \frac{2\xi}{\epsilon n}\mathbf{E}$ is $\sqrt{2}\epsilon$-GDP.

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

$$\widehat{\Sigma}_{\xi,\epsilon} = \operatorname*{argmin}_{\mathbf{H}\geq\zeta\mathbf{I}} \left\| \mathbf{H} - \left(\widehat{\Sigma}_\xi + \frac{4\xi}{\epsilon n}\mathbf{E}\right) \right\|_2, \tag{31}$$

where $\zeta > 0$ is sufficiently small. By Lemma 3.5, $\widehat{\Sigma}_{\xi,\epsilon}$ 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 $\widehat{\Sigma}_{\xi,\epsilon}$.

Proposition 3.5. *Assume $\boldsymbol{x}$ has the finite fourth moment so that v_0^2 given in (11) is well-defined. Let $n_0 = \lfloor n/2 \rfloor$ be the largest integer not exceeding $n/2$. Then, the private covariance estimator $\widehat{\Sigma}_{\xi,\epsilon}$ defined in (31) with $\xi = v_0\sqrt{n_0/\log(2nd)}$ satisfies*

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

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

Similarly to Theorem 2.5, we establish below a Berry-Esseen-type bound for the studentized private statistic $\sqrt{n}\langle \boldsymbol{u}, \boldsymbol{\mu}^{(T)} - \boldsymbol{\mu}\rangle / (\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_{\xi,\epsilon}\boldsymbol{u})^{1/2}$ for any $\boldsymbol{u} \in \mathbb{R}^d$.

Corollary 3.3. *Under the same conditions as in Theorem 3.7 with $q \geq 4$, we have*

$$\begin{aligned} &\sup_{\boldsymbol{u}\in\mathbb{R}^d, x\in\mathbb{R}} \left|\mathbb{P}\{\sqrt{n}\langle \boldsymbol{u}, \boldsymbol{\mu}^{(T)} - \boldsymbol{\mu}\rangle/(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_{\xi,\epsilon}\boldsymbol{u})^{1/2} \leq x\} - \Phi(x)\right| \\ &\lesssim \frac{m_q^{1/q}}{\underline{\lambda}^{1/2}}\left\{\frac{\sqrt{d+\log n)\log n}}{\epsilon}\right\}^{1-1/q}\left(\frac{1}{n}\right)^{1/2-1/q} + v_4^{1/2}\frac{\bar{\lambda}}{\underline{\lambda}}\sqrt{\mathrm{r}(\Sigma)\log n}\left(\sqrt{\frac{\log n}{n}} + \frac{1}{\epsilon}\sqrt{\frac{d}{n}}\right), \end{aligned} \tag{32}$$

where $\widehat{\Sigma}_{\xi,\epsilon}$ is the differentially private covariance estimator defined in (31) with $\xi = v_0\sqrt{n/\log(2nd)}$.

Recall from Theorem 2.5 that $v_0^2 \leq 2v_4\bar{\lambda}\,\mathrm{tr}(\Sigma)$. Based on Theorem 3.7 and Proposition 3.5, the proof of (32) is almost identical to that of Theorem 2.5, 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 (32) is

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

which essentially matches the upper bound in (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 3.3, for any $\alpha \in (0,1)$ and deterministic vector $\boldsymbol{u} \in \mathbb{R}^d$, we construct the following $(\sqrt{2}\epsilon)$-GDP (approximate) $100(1-\alpha)\%$ confidence interval of $\langle \boldsymbol{u}, \boldsymbol{\mu}\rangle$:

$$\left[\langle \boldsymbol{u}, \boldsymbol{\mu}^{(T)}\rangle - z_{\alpha/2}\frac{(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_{\xi,\epsilon}\boldsymbol{u})^{1/2}}{\sqrt{n}}, \langle \boldsymbol{u}, \boldsymbol{\mu}^{(T)}\rangle + z_{\alpha/2}\frac{(\boldsymbol{u}^{\mathrm{T}}\widehat{\Sigma}_{\xi,\epsilon}\boldsymbol{u})^{1/2}}{\sqrt{n}}\right], \tag{33}$$

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

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 2.4 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 (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

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

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 [4] 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 [57].

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) [44] under the following three distributions, the multivariate normal (lighted-tailed and symmetric), multivariate t (heavy-tailed and symmetric) and Pareto (heavy-tailed and asymmetric).

(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 the 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 [41] 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 1 depicts the ℓ_2-error versus sample size for the three methods, averaged over 500 repetitions. The Huber estimator is almost

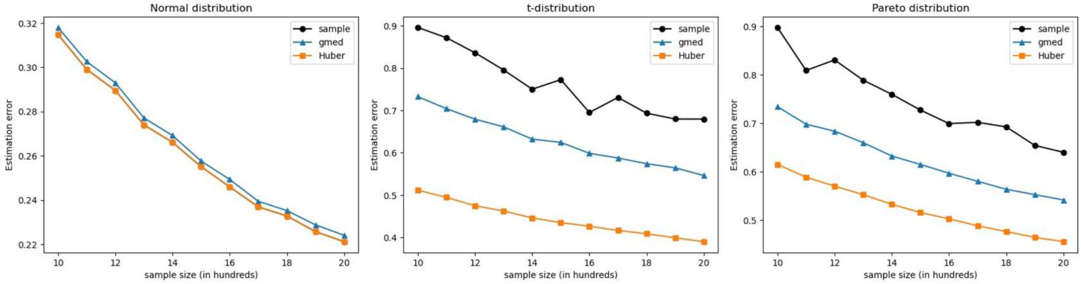

Figure 1. Plots of estimation error (under ℓ_2-norm) versus sample size based on 500 repetitions when $d = 100$.

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. 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 (15) but with $\widehat{\Sigma}_\xi$ replaced by $\widetilde{\Sigma}_\xi$ in (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 1. 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 (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. 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.

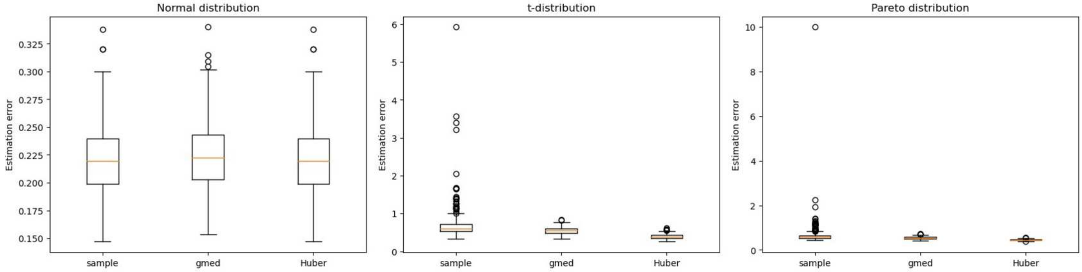

Figure 2. Boxplots of estimation error (under ℓ_2-norm) based on 500 repetitions when $(n,d) = (2000,100)$.

	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 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 $(n, d) = (3000, 100)$.

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 Huber estimator under the following two scenarios.

(i) Fix $d = 64$, let n increase from 10000 to 50000, and set $\epsilon \in \{0.3, 0.5, 0.9, \infty\}$, the privacy parameter. Here "$\epsilon = \infty$" corresponds to the non-private Huber estimator.
(ii) Fix $\epsilon = 0.5$, 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 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 3.6.

Next, we proceed to assess the performance of the proposed robust GDP CIs based on the private robust estimator. We fix the parameters $(n, d) = (50000, 32), \epsilon = 0.5$, and randomly 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 $(\sqrt{2}\epsilon)$-GDP robust 95% CI for $\langle \boldsymbol{u}, \boldsymbol{\mu} \rangle$ following the formulation outlined in (33). However, we replace $\widehat{\Sigma}_{\xi, \epsilon}$ with the perturbed plug-in covariance estimator outlined in Remark 8 to reduce computational cost. The empirical coverage probabilities, averaged over 500 Monte Carlo simulations, are presented in Table 3. The result demonstrates that private confidence intervals achieve nominal

	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

Table 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 $(n, d) = (3000, 100)$.

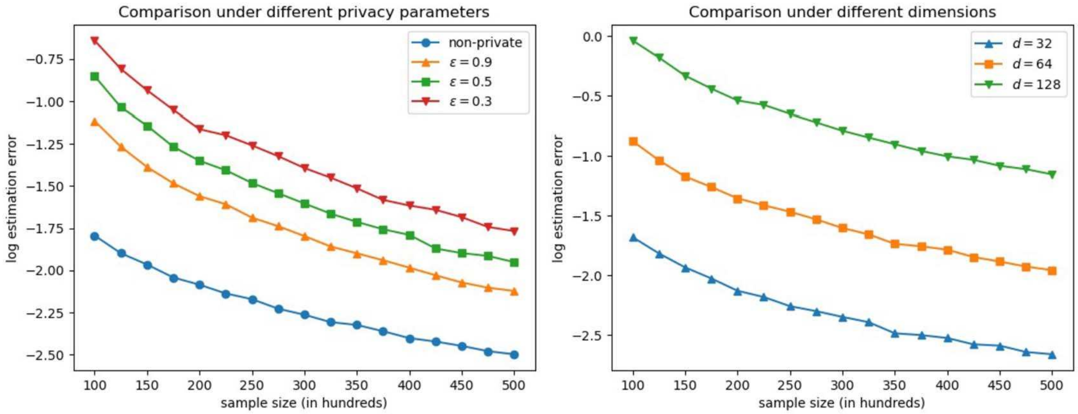

Figure 3. Plots of logarithmic ℓ_2-error versus sample size, averaged over 100 repetitions, for the private Huber mean estimator under the $t_{2.1}$ sampling distribution.

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 $\delta = \Phi(-1+\epsilon/2) - e^{\epsilon}\Phi(-1-\epsilon/2)$ (see Algorithm 3.1 in [10]) 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 $d = 50$, $\epsilon = 0.5$ (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 [10] involves a truncation tuning parameter R. For normal distributions, we use the theoretically optimal choice $R = 4\sqrt{\log n}$ as suggested in [10]; 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 4 and 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 4.1 and 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.

	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

Table 3. Empirical coverage probabilities of normal-based $100(1-\alpha)\%$ $(\sqrt{2}\epsilon)$-GDP CIs for $\boldsymbol{\mu}$ using the private Huber estimator with $\alpha \in \{0.1, 0.05\}$ and $\epsilon = 0.5$. The results are based on 500 Monte Carlo simulations when $(n,d) = (50000, 32)$.

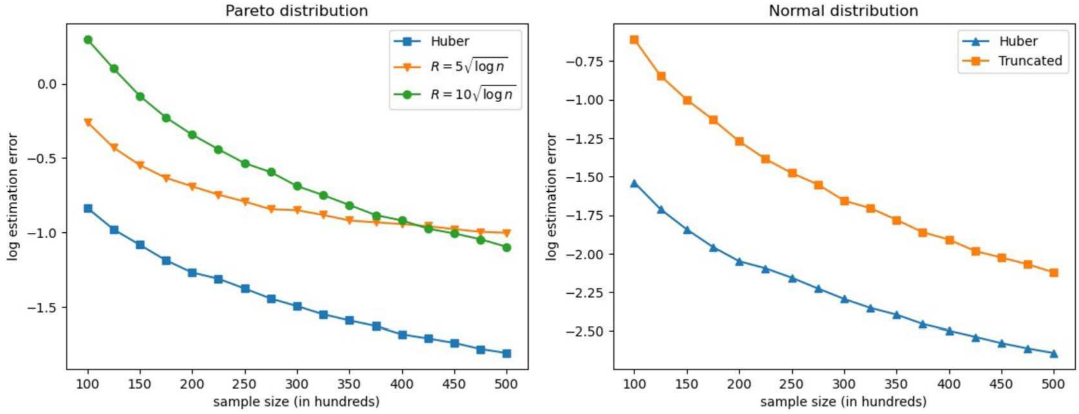

Figure 4. Plots of logarithmic ℓ_2-error versus sample size, averaged over 100 repetitions, for the ϵ-GDP Huber estimator and (ϵ, δ)-DP truncated mean estimator [10] when $d = 50$.

Appendix: Proofs of main results

A.1. Proof of Theorem 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 $\widehat{\mathcal{L}}_\tau(\cdot)$, the first-order condition holds, that is, $\nabla\widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\mu}}) = \mathbf{0}$. Further, applying Lemma C.1 in the supplementary material of [52] implies

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

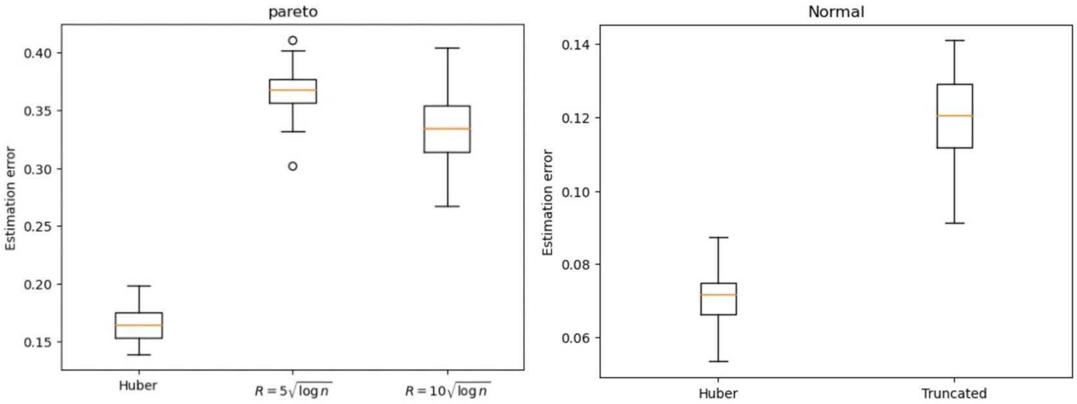

Figure 5. Boxplots of logarithmic ℓ_2 error based on 100 repetitions for the ϵ-GDP Huber estimator and (ϵ, δ)-DP truncated mean estimator [10] when $(n, d) = (50000, 50)$.

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

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

where $\lambda_{\min}(\nabla^2 \widehat{\mathcal{L}}_\tau(\boldsymbol{\theta}))$ is the smallest eigenvalue of $\nabla^2 \widehat{\mathcal{L}}_\tau(\boldsymbol{\theta})$. For any $z > 0$ and $r < \tau$, Lemma D.1 in [57] implies that, with probability at least $1 - e^{-z}$,

$$1 - \mathbb{P}\big(\|\boldsymbol{x} - \boldsymbol{\mu}\|_2 > \gamma\big) - \sqrt{\frac{z}{2n}} \le \boldsymbol{u}^{\mathrm{T}} \nabla^2 \widehat{\mathcal{L}}_\tau(\boldsymbol{\theta}) \boldsymbol{u} \le 1 \tag{34}$$

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 (23). Furthermore, by Lemma D.2 in [57], we have

$$\|\nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\mu})\|_2 \le 2\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \sqrt{\frac{2\|\Sigma\|_2 z}{n}} + \frac{4\tau z}{3n} + b_\tau \tag{35}$$

with probability at least $1 - e^{-z}$. Therefore, denoting $\mathcal{G}_z$ to be the event that (34) and (35) hold, $\mathcal{G}_z$ occurs with probability at least $1 - 2e^{-z}$. By Markov's inequality, $\mathbb{P}(\|\boldsymbol{x} - \boldsymbol{\mu}\|_2 > \gamma) \le \gamma^{-2} \operatorname{tr}(\Sigma)$. Then, conditioned on $\mathcal{G}_z$, the above upper and lower bounds yield

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

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

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

To conclude the proof, note from Lemma D.2 in [57] that $b_\tau \le \tau^{-1}\sqrt{\|\Sigma\|_2 \operatorname{tr}(\Sigma)}$. Taking $r = \gamma = \tau/2$, and let (n, τ) satisfy $n \gtrsim \mathrm{r}(\Sigma) + z$ and $\gamma \gtrsim \sqrt{\operatorname{tr}(\Sigma)}$, 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 $\widehat{\boldsymbol{\mu}}$. □

A.2. Proof of Theorem 2.3

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

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

Hence, for any $r > 0$, we have $\sup_{\|\boldsymbol{h}\|_2 \le r} \|\Delta(\boldsymbol{h})\|_2 \le \sup_{\boldsymbol{\theta} \in \Theta(r)} \|\nabla^2 \widehat{\mathcal{L}}_\tau(\boldsymbol{\theta}) - \mathbf{I}_d\|_2 \cdot r$. This together with Lemma D.1 in [57] implies that, with probability at least $1 - e^{-z}$,

$$\sup_{\|\boldsymbol{h}\|_2 \le r} \big\|\nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\mu} + \boldsymbol{h}) - \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\mu}) - \boldsymbol{h}\big\|_2 \le r\left(\gamma^{-q}\mathbb{E}\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^q + \sqrt{\frac{z}{2n}}\right), \tag{36}$$

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 2.1 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 $\widehat{\mathcal{L}}_\tau(\cdot)$ is given by

$$\nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\psi_\tau(\|\boldsymbol{x}_i - \boldsymbol{\theta}\|_2)}{\|\boldsymbol{x}_i - \boldsymbol{\theta}\|_2} (\boldsymbol{x}_i - \boldsymbol{\theta}) \tag{37}$$

for $\boldsymbol{\theta} \in \mathbb{R}^d$. Taking $r = r_0$, the claimed bound (8) follows from (36), (37) and the fact that $\nabla \widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\mu}}) = \mathbf{0}$. □

Acknowledgments

We thank the Editor, an Associate Editor, and three anonymous reviewers for their constructive comments and valuable suggestions, which have significantly helped us improve the quality of this work.

Funding

MY and WZ are supported in part by the NSF Grants DMS-2113409 and DMS-2401268. ZR is supported in part by the NSF Grant DMS-2113568 and the Central Research Development Fund at the University of Pittsburgh.

Supplementary Material

Supplement to "Gaussian differentially private robust mean estimation and inference" (DOI: 10.3150/23-BEJ1706SUPP; .pdf). The supplementary material [57] contains an extension of our construction of private robust estimators to other notions of differential privacy, a detailed description of the numerical algorithm of private robust estimators, and remaining proofs for the theoretical results in Section 2 and Section 3.

References

[1] Avella-Medina, M. (2021). Privacy-preserving parametric inference: A case for robust statistics. *J. Amer. Statist. Assoc.* **116** 969–983. MR4270037 https://doi.org/10.1080/01621459.2019.1700130

[2] Avella-Medina, M., Bradshaw, C. and Loh, P.-L. (2023). Differentially private inference via noisy optimization. *Ann. Statist.* **51** 2067–2092. MR4678796 https://doi.org/10.1214/23-aos2321

[3] 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.

[4] Barzilai, J. and Borwein, J.M. (1988). Two-point step size gradient methods. *IMA J. Numer. Anal.* **8** 141–148. MR0967848 https://doi.org/10.1093/imanum/8.1.141

[5] Bassily, R., Smith, A. and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2014* 464–473. Los Alamitos, CA: IEEE Computer Soc. MR3344896 https://doi.org/10.1109/FOCS.2014.56

[6] Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Trans. Inf. Theory* **59** 7711–7717. MR3124669 https://doi.org/10.1109/TIT.2013.2277869

[7] Bun, M. and Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography. Part I. Lecture Notes in Computer Science* **9985** 635–658. Berlin: Springer. MR3591832 https://doi.org/10.1007/978-3-662-53641-4_24

[8] 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.

[9] Cai, T.T., Wang, Y. and Zhang, L. (2020). The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. ArXiv preprint. Available at arXiv:2011.03900.

[10] 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. MR4338894 https://doi.org/10.1214/21-aos2058

[11] Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

[12] Chen, M., Gao, C. and Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607

[13] Chen, X. and Zhou, W.-X. (2020). Robust inference via multiplier bootstrap. *Ann. Statist.* **48** 1665–1691. MR4124339 https://doi.org/10.1214/19-AOS1863

[14] Clémençon, S., Colin, I. and Bellet, A. (2016). Scaling-up empirical risk minimization: Optimization of incomplete U-statistics. *J. Mach. Learn. Res.* **17** Paper No. 76. MR3517099

[15] Depersin, J. and Lecué, G. (2022). Robust sub-Gaussian estimation of a mean vector in nearly linear time. *Ann. Statist.* **50** 511–536. MR4382026 https://doi.org/10.1214/21-aos2118

[16] Depersin, J. and Lecué, G. (2022). Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. *Probab. Theory Related Fields* **183** 997–1025. MR4453320 https://doi.org/10.1007/s00440-022-01127-y

[17] Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R.I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558 https://doi.org/10.1214/16-AOS1440

[18] Diakonikolas, I. and Kane, D.M. (2019). Recent advances in algorithmic high-dimensional robust statistics. ArXiv preprint. Available at arXiv:1911.05911.

[19] Dong, J., Roth, A. and Su, W.J. (2022). Gaussian differential privacy. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 3–54. MR4400389

[20] 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. MR3803452 https://doi.org/10.1080/01621459.2017.1389735

[21] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I. and Naor, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology—EUROCRYPT 2006*. *Lecture Notes in Computer Science* **4004** 486–503. Berlin: Springer. MR2423560 https://doi.org/10.1007/11761679_29

[22] Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*. *Lecture Notes in Computer Science* **3876** 265–284. Berlin: Springer. MR2241676 https://doi.org/10.1007/11681878_14

[23] Dwork, C. and Rothblum, G.N. (2016). Concentrated differential privacy. ArXiv preprint. Available at arXiv:1603.01887.

[24] 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. MR4047307 https://doi.org/10.1080/01621459.2018.1527700

[25] 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. MR3597972 https://doi.org/10.1111/rssb.12166

[26] Hampel, F., Hennig, C. and Ronchetti, E. (2011). A smoothing principle for the Huber and other location M-estimators. *Comput. Statist. Data Anal.* **55** 324–337. MR2736558 https://doi.org/10.1016/j.csda.2010.05.001

[27] Heyde, C.C. (1967). On the influence of moments on the rate of convergence to the normal distribution. *Z. Wahrsch. Verw. Gebiete* **8** 12–18. MR0215344 https://doi.org/10.1007/BF00533941

[28] Hopkins, S.B. (2020). Mean estimation with sub-Gaussian rates in polynomial time. *Ann. Statist.* **48** 1193–1213. MR4102693 https://doi.org/10.1214/19-AOS1843

[29] Hopkins, S.B., Kamath, G. and Majid, M. (2022). Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *STOC '22—Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing* 1406–1417. New York: ACM. MR4490088

[30] 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.

[31] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

[32] Kamath, G., Li, J., Singhal, V. and Ullman, J. (2019). Privately learning high-dimensional distributions. In *Conference on Learning Theory* **99** 1853–1902.

[33] 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.

[34] Kamath, G., Singhal, V. and Ullma, J. (2020). Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory* **125** 2204–2235.

[35] Karwa, V. and Vadhan, S. (2018). Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science*. *LIPIcs. Leibniz Int. Proc. Inform.* **94** Art. No. 44. Wadern: Schloss Dagstuhl. Leibniz-Zent. Inform. MR3761780

[36] 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. MR4017523 https://doi.org/10.1214/19-STS711

[37] 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.

[38] Lugosi, G. and Mendelson, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** 783–794. MR3909950 https://doi.org/10.1214/17-AOS1639

[39] Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.* **19** 1145–1190. MR4017683 https://doi.org/10.1007/s10208-019-09427-x

[40] Lugosi, G. and Mendelson, S. (2021). Robust multivariate mean estimation: The optimality of trimmed mean. *Ann. Statist.* **49** 393–410. MR4206683 https://doi.org/10.1214/20-AOS1961

[41] Mathieu, T. (2022). Concentration study of M-estimators using the influence function. *Electron. J. Stat.* **16** 3695–3750. MR4444667 https://doi.org/10.1214/22-ejs2030

[42] McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science* 94–103.

[43] Mendelson, S. and Zhivotovskiy, N. (2020). Robust covariance estimation under L_4-L_2 norm equivalence. *Ann. Statist.* **48** 1648–1664. MR4124338 https://doi.org/10.1214/19-AOS1862

[44] Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468 https://doi.org/10.3150/14-BEJ645

[45] Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758 https://doi.org/10.1214/17-AOS1642

[46] Minsker, S. and Wei, X. (2020). Robust modifications of U-statistics and applications to covariance estimation problems. *Bernoulli* **26** 694–727. MR4036049 https://doi.org/10.3150/19-BEJ1149

[47] Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)* 263–275.

[48] Murtagh, J. and Vadhan, S. (2016). The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography. Part I. Lecture Notes in Computer Science* **9562** 157–175. Berlin: Springer. MR3487659 https://doi.org/10.1007/978-3-662-49096-9_7

[49] Rohde, A. and Steinberger, L. (2020). Geometrizing rates of convergence under local differential privacy constraints. *Ann. Statist.* **48** 2646–2670. MR4152116 https://doi.org/10.1214/19-AOS1901

[50] 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.

[51] Spokoiny, V. and Zhilova, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.* **43** 2653–2675. MR3405607 https://doi.org/10.1214/15-AOS1355

[52] Sun, Q., Zhou, W.-X. and Fan, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265. MR4078461 https://doi.org/10.1080/01621459.2018.1543124

[53] 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. MR3837109 https://doi.org/10.1017/9781108231596

[54] Wang, Y., Kifer, D. and Lee, J. (2019). Differentially private confidence intervals for empirical risk minimization. *J. Priv. Confid.* **9** 1–36.

[55] Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *J. Amer. Statist. Assoc.* **105** 375–389. MR2656057 https://doi.org/10.1198/jasa.2009.tm08651

[56] 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.

[57] Yu, M., Ren, Z. and Zhou, W.-X. (2024). Supplement to "Gaussian differentially private robust mean estimation and inference." https://doi.org/10.3150/23-BEJ1706SUPP

[58] 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. MR3845005 https://doi.org/10.1214/17-AOS1606

Received April 2023 and revised December 2023