Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Adapting to Noise Tails in Private Linear Regression

Jinyuan Chang, Lin Yang, Mengyue Zha & Wen-Xin Zhou

To cite this article: Jinyuan Chang, Lin Yang, Mengyue Zha & Wen-Xin Zhou (20 Mar 2026): Adapting to Noise Tails in Private Linear Regression, Journal of the American Statistical Association, DOI: 10.1080/01621459.2026.2644613

To link to this article: https://doi.org/10.1080/01621459.2026.2644613

Adapting to Noise Tails in Private Linear Regression

Jinyuan Chang[1,2,3], Lin Yang[1,*] , Mengyue Zha[4], and Wen-Xin Zhou[5]

[1]Joint Laboratory of Data Science and Business Intelligence, Institute of Statistical Interdisciplinary Research, Southwestern University of Finance and Economics, Chengdu, China

[2]State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

[3]China Center for Economic Research, Peking University, Beijing, China

[4]Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

[5]College of Business Administration, University of Illinois Chicago, Chicago, IL, USA

*Corresponding author. E-mail: yanglin@swufe.edu.cn

Abstract

While the traditional goal of statistics is to infer population parameters, modern practice increasingly demands protection of individual privacy. One way to address this need is to adapt classical statistical procedures into privacy-preserving algorithms. In this paper, we develop differentially private tail-robust methods for linear regression. The trade-off among bias, privacy, and robustness is controlled by a tunable robustification parameter in the Huber loss. We implement noisy clipped gradient descent for low-dimensional settings and noisy iterative hard thresholding for high-dimensional sparse models. Under sub-Gaussian errors, our method achieves near-optimal convergence rates while relaxing several assumptions required in earlier work. For heavy-tailed errors, we explicitly characterize how the non-asymptotic convergence rate depends on the moment index, privacy parameters, sample size, and intrinsic dimension. Our analysis shows how the moment index influences the choice of robustification parameters and, in turn, the resulting statistical error and privacy cost. By quantifying the interplay among bias, privacy, and robustness, we extend classical perspectives on privacy-preserving robust regression. The proposed methods are evaluated through simulations and two real datasets.

Keywords: Differential privacy, heavy-tailed error, Huber regression, iterative hard thresholding, linear model, sparsity.

1 Introduction

The increasing demand for privacy-preserving statistical methods has brought differential privacy (Dwork et al., 2006) to the forefront of statistical research. Informally, a differentially private (DP) algorithm ensures that an attacker cannot determine whether a particular data point is present in the dataset. Pioneering works, such as those by Hardt and Talwar (2010), Chaudhuri and Hsu (2012), Duchi et al. (2013), and Duchi et al. (2018), quantified the cost of privacy in a range of statistical estimation problems. More recently, the role of privacy in statistical inference has been rigorously investigated (Sheffet, 2017; Cai et al., 2017; Awan and Slavković, 2018; Karwa and Vadhan, 2018; Chang et al., 2024). Interestingly, even prior to these developments, Nissim et al. (2007) and Dwork and Lei (2009) recognized a fundamental connection between privacy and robustness (Hampel et al., 1986). In particular, Dwork and Lei (2009) introduced the propose-test-release (PTR) framework, which derives DP algorithms from principles of robust statistics. This philosophy of leveraging robustness to achieve privacy has since motivated a series of follow-up works, including Lei (2011), Smith (2011), Chaudhuri and Hsu (2012), Avella-Medina (2021), Liu et al. (2022), and Yu et al. (2024), to name a few. Specifically, Avella-Medina (2021) and Liu et al. (2022) focused on robustness against small fractions of contamination in the data, whereas Yu et al. (2024) considered robustness to heavy-tailed sampling distributions.

In this work, we focus on both linear and sparse linear regressions, which are among the most fundamental statistical problems and serve as building blocks for more advanced methodologies. Consider the linear model $y = \mathbf{x}^\top \boldsymbol{\beta}^* + \varepsilon$, where $y \in \mathbb{R}$ is the response variable, $\mathbf{x} \in \mathbb{R}^p$ denotes the (random) vector of covariates, $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the unknown vector of regression coefficients, and ε is a mean-zero error variable. Assuming that ε follows a normal or sub-Gaussian distribution, significant progress has been made in developing DP ordinary least squares (OLS) estimators for $\boldsymbol{\beta}^*$ since the works of Vu and Slavkovic (2009) and Kifer et al. (2012). When the sample size n is much larger than p, several studies have made notable advances in terms of accuracy, sample efficiency, and computational efficiency (Sheffet, 2017; Wang, 2018; Sheffet, 2019; Cai et al., 2021; Varshney et al., 2022; Brown et al., 2024). In the case where the covariates and noise are independent and both possess finite moments of polynomial order, Liu et al. (2022) established the existence of a DP algorithm. However, their proposed high-dimensional PTR algorithm is not computationally efficient. Furthermore, Kifer et al. (2012) and Cai et al. (2021) have extended DP OLS methods to high-dimensional settings.

In the presence of heavy-tailed errors, the finite-sample performance of the OLS estimator becomes sub-optimal, particularly in terms of how its error bounds depend on δ, which may scale logarithmically or polynomially at a confidence level of $1-\delta$ (Catoni, 2012). From a privacy-preserving perspective, the construction of DP procedures relies critically on the notion of sensitivity (Dwork et al., 2006), which is closely tied to the boundedness of the loss function's gradient (Bassily et al., 2014). This characteristic makes OLS-based methods inadequate, as their gradients exhibit heavy tails, which prevents them from being properly bounded. To achieve both robustness against heavy-tailed error distributions and strong privacy guarantees, we

consider employing loss functions with bounded derivatives, such as the well-known Huber loss (Huber, 1973). Many other loss functions, particularly smoothed variants of the Huber loss, share desirable properties such as Lipschitz continuity and local strong convexity. In this work, we focus on the Huber loss to more effectively illustrate the core ideas, without pursuing purely technical generalizations. When the noise variable is independent of the covariates and follows a symmetric distribution, using the Huber loss with a fixed parameter (independent of the data scale) is typically sufficient. In more general settings with heteroscedasticity or asymmetry, Fan et al. (2017) and Sun et al. (2020) introduced adaptive Huber regression, in which the robustification parameter τ is adjusted according to the sample size *n*, dimensionality *p*, and noise scale to balance bias and robustness effectively. We provide a brief review of Huber regression in Section 3.1 and Section B of the supplementary material. Building on these works, we aim to develop a unified DP tail-robust regression framework that applies to 'general' linear regression models in both low- and high-dimensional settings. Here, 'generality' refers to the minimal assumptions that $\mathbb{E}(\varepsilon|\mathbf{x})=0$ and $\mathbb{E}(\varepsilon^2|\mathbf{x})\le\sigma_0^2$, without requiring independence between ε and $\mathbf{x}$, or symmetry of the noise distribution. To elucidate the effect of tail behavior on the privacy cost, we explicitly characterize the choice of the robustification parameter when the noise variable exhibits either bounded higher-order moments or sub-Gaussian tails. This parameter offers new insight by effectively bridging the trinity of robustness (quantified via non-asymptotic tail bounds), bias (arising from skewness in the response or error distribution), and privacy.

In low-dimensional settings where $n \gg p$, we propose a DP Huber regression estimator implemented via noisy clipped gradient descent, with noise carefully calibrated to ensure the desired privacy guarantee. Our method builds on the framework of private empirical risk minimization (ERM) via gradient perturbation (Bassily et al., 2014). This line of work, together with earlier studies based on objective function perturbation (Chaudhuri et al., 2011; Kifer et al., 2012), provides utility guarantees in terms of excess risk while preserving privacy, under the assumption that the loss function satisfies specific convexity and differentiability conditions. In particular, the objective function perturbation framework requires an objective composed of a convex loss with bounded derivative and a differentiable, strongly convex regularizer. The Huber loss, by contrast, has a derivative bounded in magnitude by τ, which is allowed to increase with the sample size, and neither the low- or high-dimensional Huber regressions employ a strongly convex penalty. As a result, the objective function perturbation method proposed by Chaudhuri et al. (2011) is not applicable to our setting. Under standard assumptions such as strong convexity of the loss function and boundedness of the parameter space, Wang et al. (2020) and Kamath et al. (2022) developed general theoretical frameworks for DP stochastic convex optimization problems with heavy-tailed data. Avella-Medina et al. (2023) incorporated robust statistics into noisy gradient descent to facilitate DP estimation and inference. Nevertheless, the presence of the robustification parameter τ introduces additional complexity. Existing techniques and theoretical results do not seamlessly extend to this setting, particularly under general noise distributions. To address this gap, we adapt the framework of Avella-Medina et al. (2023), explicitly quantifying the influence of τ on both the statistical error and the error induced by privacy constraints. We first show that the DP Huber estimator lies, with high probability, in a small neighborhood of its non-private counterpart. We then establish its convergence rate to the true parameter $\boldsymbol{\beta}^*$ in the ℓ_2-norm. Beyond standard differential privacy,

we further evaluate the performance of the DP Huber estimator under the framework of Gaussian differential privacy (GDP) (Dong et al., 2022). Compared to (ϵ, δ)-DP (see Definition 1 in Section 2), GDP has attracted growing attention in the statistics community due to its elegant interpretation of privacy through the lens of hypothesis testing (Wasserman and Zhou, 2010).

In high-dimensional settings where $\|\boldsymbol{\beta}^*\|_0 \ll \min\{n, p\}$, ensuring privacy becomes more challenging due to the sparse structure of $\boldsymbol{\beta}^*$. Wang and Gu (2019) and Cai et al. (2021) proposed a noisy variant of the iterative hard thresholding (HT) algorithm (Blumensath and Davies, 2009; Jain et al., 2014) for implementing a privatized sparse OLS estimator. They demonstrated that, under Gaussian errors and certain conditions on the covariates, their methods may achieve (near-)optimal statistical performance when applied to sparse linear models. In contrast, DP sparse regression that is robust to heavy-tailed errors has remained largely understudied until recent work by Liu et al. (2022) and Hu et al. (2022). Liu et al. (2022) adapted soft thresholding with the absolute loss and ℓ_1-regularization to achieve a convergence rate that scales with $\sqrt{p}$. Hu et al. (2022) proposed truncating the original data at a fixed threshold to mitigate the influence of heavy-tailed distributions. However, neither Liu et al. (2022) nor Hu et al. (2022) achieved the optimal convergence rate that accounts for sparsity. Moreover, the interaction among bias, privacy, and robustness remains insufficiently explored. To bridge this gap, we propose a sparse DP Huber estimator based on noisy iterative hard thresholding. Our analysis establishes the convergence rate of the sparse DP Huber estimator to the true parameter $\boldsymbol{\beta}^*$ in the ℓ_2-norm and examines the bias-privacy-robustness triad. Extending our approach to other robust *M*-estimators and iterative algorithms for ℓ_0-constrained *M*-estimation (e.g., Liu et al., 2019; She et al., 2023) is feasible but beyond the scope of the current work and left for future investigation.

Alongside statistical performance, we examine the interaction among estimation accuracy, privacy, and robustness. We demonstrate that the robustification parameter τ affects global sensitivity and, consequently, the privacy cost. In other words, τ governs the trade-off among bias, privacy, and robustness. Our theoretical results suggest choosing τ based on the effective sample size under privacy constraints. Similar to Barber and Duchi (2014) and Kamath et al. (2020), we study how the moment condition parameter $\iota \geq 0$ influences this trade-off when the noise variables have bounded $2+\iota$ moments. We find that ι affects the optimal choice of τ, which in turn impacts both the statistical error and the privacy cost. In high-dimensional settings, we further investigate how sparsity shapes the bias-privacy-robustness trade-off. By quantifying the effects of this trade-off on convergence rates and sample size requirements, we provide a new perspective on the relationship between privacy and robustness. This perspective complements prior work that explores the interplay between these two properties (Dwork and Lei, 2009; Avella-Medina, 2021; Liu et al., 2021; Georgiev and Hopkins, 2022; Asi et al., 2023; Liu et al., 2023; Hopkins et al., 2023).

The rest of the paper is organized as follows. Section 2 provides a brief review of (ϵ, δ)-DP and ϵ-GDP. Section 3 presents a noisy gradient descent algorithm for low-dimensional DP Huber regression and a noisy iterative hard thresholding method for sparse Huber regression in high dimensions. Section 4 provides theoretical guarantees for DP Huber regression, including

comprehensive non-asymptotic convergence results under either polynomial-moment or sub-Gaussian noise, followed by a discussion of the bias-privacy-robust trade-off. As a byproduct, we also establish the statistical convergence of the sparse Huber estimator (in the absence of privacy constraints), computed via iterative HT. Section 5 demonstrates the proposed methods on simulated datasets. The real data analysis, all proofs, and the construction of DP confidence intervals are provided in the supplementary material. The used real data and the code for implementing our proposed methods are available at the GitHub repository: https://github.com/JinyuanChang-Lab/DifferentiallyPrivateHuberRegression.

Notation. For every integer $k \geq 1$, denote by $\mathbb{R}^k$ the k-dimensional Euclidean space. Let $\|\mathbf{u}\|_1$, $\|\mathbf{u}\|_2$ and $\|\mathbf{u}\|_\infty$ denote the ℓ_1-norm, the ℓ_2-norm and the ℓ_∞-norm of the vector $\mathbf{u}$, respectively. We denote the (pseudo) ℓ_0-norm of $\mathbf{u} = (u_1, \ldots, u_k)^\top$ as $\|\mathbf{u}\|_0 = \sum_{i=1}^{k} 1(u_i \neq 0)$, where $1(\cdot)$ is the indicator function. Write $\mathbb{S}^{k-1} := \{\mathbf{u} \in \mathbb{R}^k : \|\mathbf{u}\|_2 = 1\}$, $[k] := \{1, \ldots, k\}$, and $\mathbb{B}^k(r) := \{\mathbf{u} \in \mathbb{R}^k : \|\mathbf{u}\|_2 \leq r\}$. For a subset $S \subseteq [k]$ with cardinality $|S|$ and a k-dimensional vector $\mathbf{u} \in \mathbb{R}^k$, we write $\mathbf{u}_S \in \mathbb{R}^k$ as the vector obtained by setting all entries in $\mathbf{u}$ to 0, except for those indexed by S. The inner product of any two vectors $\mathbf{u} = (u_1, \ldots, u_k)^\top$ and $\mathbf{v} = (v_1, \ldots, v_k)^\top$ is defined by $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{k} u_i v_i$. For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$, we denote the smallest and largest eigenvalues of $\mathbf{A}$ by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$, respectively. For two sequences of non-negative numbers $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, we say $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a constant $C > 0$ independent of n such that $a_n \leq C b_n$. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously, and $a_n \ll b_n$ or $b_n \gg a_n$ if $\lim_{n \to \infty} a_n / b_n = 0$. For a matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$, let $\|\mathbf{A}\|$ denote the spectral norm of $\mathbf{A}$. For any two matrices $\mathbf{A}$ and $\mathbf{B}$, we write $\mathbf{A} \succeq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is positive semi-definite. For any $x \in \mathbb{R}$, we write $\lceil x \rceil = \inf\{y \in \mathbb{Z} : y \geq x\}$ and $\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-u^2/2} \mathrm{d}u$.

2 Background on Differential Privacy

Differential privacy was originally introduced to provide a formal framework for data privacy. It ensures that a randomized mechanism $\mathcal{M}$ produces similar output distributions for datasets $\mathbf{X}$ and $\mathbf{X}'$ that differ by only a single data point. Intuitively, this implies that an attacker cannot determine whether a particular data point x is included in the dataset $\mathbf{X}$ based on the output of the mechanism. The formal definition is given below.

Definition 1 (Dwork et al. (2006)).

A randomized mechanism $\mathcal{M}: \mathcal{X} \to \mathbb{R}^d$ is said to be (ϵ, δ)-DP if, for every pair of adjacent datasets $\mathbf{X}, \mathbf{X}' \in \mathcal{X}$ that differ in exactly one data point, and for every measurable subset $S \subseteq \mathbb{R}^d$, it holds that $\mathbb{P}\{\mathcal{M}(\mathbf{X}) \in S\} \leq e^{\epsilon} \cdot \mathbb{P}\{\mathcal{M}(\mathbf{X}') \in S\} + \delta$.

In addition to its privacy guarantees, differential privacy is valued for the simplicity and versatility in the design of private algorithms. Typically, such algorithms are constructed by adding random noise to the output of a non-private algorithm. Among the various types of noise, Gaussian and Laplace noise are the most commonly used; see Theorems 3.6 and 3.22 of Dwork and Roth (2014). Let $\mathcal{M}$ be an algorithm that maps a dataset $\mathbf{X}$ to $\mathbb{R}^d$. For any $q \geq 1$, the ℓ_q-sensitivity of $\mathcal{M}$ is defined as $\text{sens}_q(\mathcal{M}) = \sup_{\mathbf{X},\mathbf{X}'} \| \mathcal{M}(\mathbf{X}) - \mathcal{M}(\mathbf{X}') \|_q$, where the supremum is taken over all pairs of datasets $\mathbf{X}$ and $\mathbf{X}'$ that differ by a single data point.

Lemma 1 .

(Gaussian mechanism) Assume $\text{sens}_2(\mathcal{M}) \leq B$ for some $B > 0$. Then $\mathcal{M}(\mathbf{X}) + \mathbf{g}$, with $\mathbf{g} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ and $\sigma = B\epsilon^{-1}\sqrt{2\log(1.25/\delta)}$, is (ϵ, δ)-DP.

The construction of multi-step differential private algorithms is further enhanced by the post-processing property (Dwork and Roth, 2014) and the composition property (Dwork and Roth, 2014; Kairouz et al., 2017) of differential privacy. The composition property characterizes the evolution of privacy parameters under composition. Heuristically, the post-processing property ensures that no external entity can undo the privatization.

Lemma 2 .

(Post-processing property). Let $\mathcal{M}$ be an (ϵ, δ)-DP algorithm, and let *g* be an arbitrary deterministic mapping that takes the output of $\mathcal{M}$ as an input. Then $g(\mathcal{M}(\mathbf{X}))$ is also (ϵ, δ)-DP.

Lemma 3 .

(Standard composition property (Dwork and Roth, 2014)). Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be (ϵ_1, δ_1)-DP and (ϵ_2, δ_2)-DP algorithms, respectively. The composition $\mathcal{M}_1 \circ \mathcal{M}_2$ is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-DP.

Lemma 4 .

(Advanced composition property (Dwork and Roth, 2014)). If *T* algorithms are run sequentially, each satisfying (ϵ, δ)-DP, then for any $\delta' > 0$, the combined procedure is $(\epsilon\sqrt{2T\log(1/\delta')} + T\epsilon(e^{\epsilon} - 1), T\delta + \delta')$-DP.

By the standard composition property, if each step of an iterative algorithm is $(\epsilon/T, \delta/T)$-DP, then after *T* iterations, the resulting algorithm will be (ϵ, δ)-DP. Moreover, if each step of an iterative algorithm is $(\epsilon\sqrt{2/\{5T\log(2/\delta)\}}, \delta/(2T))$-DP, where

$$0 < \epsilon \leq 1 \text{ and } 0 < \delta \leq 0.01, \quad (1)$$

then, by the advanced composition property (with $\delta' = \delta/2$) and a straightforward computation, the combined algorithm after *T* iterations is also (ϵ, δ)-DP.

Wasserman and Zhou (2010) presented a statistical perspective that connects differential privacy to hypothesis testing. Briefly, consider the following hypothesis testing problem

$$H_0: \text{ the underlying data is } \mathbf{X} \quad \text{versus} \quad H_1: \text{ the underlying data is } \mathbf{X}'. \quad (2)$$

Suppose x_1 is the only data point present in $\mathbf{X}$ but not in $\mathbf{X}'$. Rejecting H_0 would reveal x_1's absence, while accepting H_0 would confirm its presence. When an (ϵ, δ)-DP algorithm $\mathcal{M}$ is used, the power of any test at significance level α is bounded by $\min\{e^{\epsilon}\alpha + \delta, 1 - e^{-\epsilon}(1 - \alpha - \delta)\}$. When both ϵ and δ are small, any α-level test becomes nearly powerless. To address this limitation, Dong et al. (2022) introduced the trade-off function to characterize the trade-off between Type I and Type II errors.

Definition 2 (Trade-off function).

For any two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on the same space, the trade-off function $T(\mathbb{P}, \mathbb{Q}): [0,1] \to [0,1]$ is defined as $T(\mathbb{P}, \mathbb{Q})(\alpha) = \inf\{\beta_\phi : \alpha_\phi \le \alpha\}$, where $\alpha_\phi = \mathbb{E}_{\mathbb{P}}(\phi)$ and $\beta_\phi = 1 - \mathbb{E}_{\mathbb{Q}}(\phi)$ represent the Type I and Type II errors associated with the test ϕ, respectively. The infimum is taken over all measurable tests.

The larger the trade-off function, the more difficult it becomes to distinguish between the two distributions via hypothesis testing. Building on trade-off functions, Dong et al. (2022) proposed a generalization of differential privacy, referred to as *f*-DP. With a slight abuse of notation, we identify $\mathcal{M}(\mathbf{X})$ and $\mathcal{M}(\mathbf{X}')$ with their corresponding probability distributions.

Definition 3 (*f*-DP and GDP).

(i) Let *f* be a trade-off function. A mechanism $\mathcal{M}$ is said to be *f*-DP if $T(\mathcal{M}(\mathbf{X}), \mathcal{M}(\mathbf{X}')) \ge f$ for all adjacent data sets $\mathbf{X}, \mathbf{X}' \in \mathcal{X}$ that differ by a single data point. (ii) Let $\epsilon > 0$. A mechanism $\mathcal{M}$ is said to be ϵ-Gaussian differentially private (ϵ-GDP) if it is G_ϵ-DP, where $G_\epsilon(\alpha) = \Phi(\Phi^{-1}(1-\alpha) - \epsilon)$.

A mechanism $\mathcal{M}$ is (ϵ, δ)-DP if and only if it is $f_{\epsilon,\delta}$-DP, where $f_{\epsilon,\delta}(\alpha) = \max\{0, 1 - e^{\epsilon}\alpha - \delta, e^{-\epsilon}(1 - \alpha - \delta)\}$ (Wasserman and Zhou, 2010). We refer to Figure 3 in Dong et al. (2022) for a visual comparison between $G_\epsilon(\cdot)$ introduced in Definition 3 and $f_{\epsilon,\delta}(\cdot)$. GDP is a core single-parameter family within the *f*-DP framework, and is defined via testing two shifted Gaussian distributions; see Definition 2.6 in Dong et al. (2022). Lemma 5 below extends Theorem 1 in Dong et al. (2022) from the univariate to the multivariate setting.

Lemma 5 .

The Gaussian mechanism, $\mathcal{M}(\mathbf{X}) + \epsilon^{-1}\text{sens}_2(\mathcal{M})\cdot\mathbf{g}$ with $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, is ϵ-GDP.

Under the GDP framework, the composition of *T* mechanisms with parameters $\{\epsilon_t\}_{t=1}^T$ yields overall ϵ-GDP with $\epsilon = \sqrt{\epsilon_1^2 + \cdots + \epsilon_T^2}$. This implies that the individual privacy level ϵ_t can be set to $\epsilon / \sqrt{T}$.

3 Differentially Private Huber Regression

This section introduces algorithms for DP Huber regression in both low- and high-dimensional settings. We begin by reviewing non-private Huber regression, accompanied by the required notation and relevant background.

3.1 Huber regression

Suppose we observe *n* independent samples $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ satisfying the linear model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \varepsilon_i = \sum_{j=1}^p x_{i,j}\beta_j^* + \varepsilon_i, \quad i \in [n], \text{ (3)}$$

where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^\top$ is a *p*-dimensional feature vector with $x_{i,1} \equiv 1$, and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\top \in \mathbb{R}^p$ denotes the unknown true coefficient vector. In the random design setting, let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent copies of a random vector $\mathbf{x} \in \mathbb{R}^p$. The noise variable ε_i satisfies $\mathbb{E}(\varepsilon_i | \mathbf{x}_i) = 0$ and $\mathbb{E}(\varepsilon_i^2 | \mathbf{x}_i) \le \sigma_0^2 < \infty$. Under this moment condition, while the OLS estimator, obtained by minimizing $\boldsymbol{\beta} \mapsto \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$, exhibits desirable asymptotic properties as $n \to \infty$ with *p* fixed, its finite sample performance, particularly in terms of high-probability tail bounds, is suboptimal compared to the case when ε_i is sub-Gaussian (Catoni, 2012; Sun et al., 2020).

To obtain an estimator that is both asymptotically efficient and exhibits exponential-type concentration bounds in finite-sample settings, we employ the robust loss $\rho_\tau(u) = \tau^2 \rho(u/\tau)$, where $\rho : \mathbb{R} \to [0, \infty)$ is a continuously differentiable convex function, and $\tau > 0$ serves as a robustification parameter. Assume that $\psi(u) = \rho'(u)$ is Lipschitz continuous, concave, and differentiable almost everywhere. We consider the following *M*-estimator:

$$\boldsymbol{\beta}_\tau \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \{\mathcal{L}_\tau(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\}. \text{ (4)}$$

Define $\psi_\tau(u) = \rho_\tau'(u) = \tau\psi(u/\tau)$ for $u \in \mathbb{R}$. Due to the convexity of $\rho_\tau(\cdot)$ and, consequently, of $\mathcal{L}_\tau(\cdot)$, the *M*-estimator $\boldsymbol{\beta}_\tau$ can alternatively be characterized as the solution to the first-order condition $\sum_{i=1}^n \psi_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i = \mathbf{0}$. For simplicity, we focus on the Huber loss (Huber, 1964), defined as $\rho(u) = (u^2/2)1(|u| \le 1) + (|u| - 1/2)1(|u| > 1)$ for $u \in \mathbb{R}$. The associated score function is $\psi(u) = \text{sign}(u)\min\{|u|, 1\}$. To conserve space, we defer the details on Huber regression to Section B of the supplementary material. The next section demonstrates the construction of a DP Huber estimator in low dimensions.

3.2 Randomized estimator via noisy gradient descent

We begin by considering the low-dimensional setting where $n \gg p$. To compute the *M*-estimator $\boldsymbol{\beta}_\tau$ defined in (4), one of the most widely used algorithms is gradient descent, which generates iterates as follows:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0 \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^{(t)}) = \boldsymbol{\beta}^{(t)} + \frac{\eta_0}{n}\sum_{i=1}^n \psi_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(t)})\mathbf{x}_i, \quad t \ge 0,$$

where $\boldsymbol{\beta}^{(0)}$ is an initial estimator and $\eta_0 > 0$ is the learning rate. For each $t \ge 0$, given the previous iterate, the gradient descent update can be viewed as an algorithm that maps the dataset $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ to $\mathbb{R}^p$. However, in the case of random features, the sensitivity of this algorithm may not be bounded. To address this, we use a clipping function $w_\gamma(t) = \min\{\gamma/t, 1\}$, $t \ge 0$, enabling the privatization of gradient descent.

Motivated by the general concept of noisy gradient descent in the differential privacy literature (Avella-Medina et al., 2023), we consider a noisy clipped gradient descent method that incorporates the Gaussian mechanism. Let $T \ge 1$ be a prespecified number of iterations, $\gamma > 0$ a truncation parameter, and $\sigma > 0$ a noise level to be determined. Starting with an initial estimate $\boldsymbol{\beta}^{(0)}$, the noisy clipped gradient descent computes updates as follows:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \eta_0\{-\frac{1}{n}\sum_{i=1}^n \psi_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(t)})\mathbf{x}_i w_\gamma(\|\mathbf{x}_i\|_2) + \sigma \mathbf{g}_t\}, \quad t \in \{0\} \cup [T-1], \quad (5)$$

where $\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_{T-1} \in \mathbb{R}^p$ are i.i.d. standard normal random vectors.

Recall that $\psi_\tau(u) = \tau\,\text{sign}(u)\min\{|u/\tau|, 1\}$ is the derivative of the Huber loss, and it satisfies $\sup_{u \in \mathbb{R}} |\psi_\tau(u)| \le \tau$. Therefore, we have $\sup_{(y,\mathbf{x}) \in \mathbb{R}\times\mathbb{R}^p, \boldsymbol{\beta} \in \mathbb{R}^p} \|\psi_\tau(y - \mathbf{x}^\top \boldsymbol{\beta})\mathbf{x} w_\gamma(\|\mathbf{x}\|_2)\|_2 \le \gamma\tau$, which implies that the ℓ_2-sensitivity of each clipped gradient descent step is bounded by $2\eta_0\gamma\tau/n$. By Lemmas 1 and 5, each noisy clipped gradient descent step with noise scales

$$\sigma_{\mathrm{dp}} = \frac{2\gamma\tau}{n\epsilon} \cdot T\sqrt{2\log(\frac{1.25T}{\delta})} \quad \text{and} \quad \sigma_{\mathrm{gdp}} = \frac{2\gamma\tau}{n\epsilon} \cdot \sqrt{T} \quad (6)$$

is, respectively, $(\epsilon/T, \delta/T)$-DP and $(\epsilon/\sqrt{T})$-GDP. After *T* iterations, the standard composition property implies that the *T*-th iterate $\boldsymbol{\beta}^{(T)}$ is either (ϵ, δ)-DP or ϵ-GDP, depending on the choice of the noise scale. The complete algorithm is presented in Algorithm 1. Alternatively, a gradient descent step with noise scale

$$\sigma_{\mathrm{dp}} = \frac{2\gamma\tau}{n\epsilon}\sqrt{5T\log(\frac{2}{\delta})\log(\frac{5T}{2\delta})}$$

ensures $(\epsilon\sqrt{2/\{5T\log(2/\delta)\}}, \delta/(2T))$-DP per iteration. By Lemma 4, the final iterate $\boldsymbol{\beta}^{(T)}$ is (ϵ, δ)-DP after *T* iterations, provided that the privacy constraints in (1) are met. In practice, for a given triplet (T, ϵ, δ), we choose the configuration that yields the smaller value of σ_{dp} between the two alternatives.

Algorithm 1 DP Huber Regression

Input: Dataset $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, initial value $\boldsymbol{\beta}^{(0)}$, learning rate η_0, number of iterations *T*, truncation level γ, robustification parameter τ, and privacy parameters (ϵ, δ).

1: **for** $t = 0, \ldots, T-1$ **do**

2: Generate $\mathbf{g}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and compute

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \eta_0\{\frac{1}{n}\sum_{i=1}^{n}\psi_\tau(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}^{(t)})\mathbf{x}_i w_\gamma(\|\mathbf{x}_i\|_2) + \sigma\mathbf{g}_t\},$$

where $\sigma > 0$ is set to either σ_{dp} or σ_{gdp} specified in (6);

3: **end for**

Output: $\boldsymbol{\beta}^{(T)}$.

Remark 1 .

Algorithm 1 extends and strengthens the (ϵ, δ)-DP least squares algorithm (Algorithm 4.1 in Cai et al. (2021)) in several important aspects. First, it permits the noise variable in (3) to follow a general distribution, which may be non-Gaussian and may exhibit heavy tails or asymmetry; see Theorem 1 in Section 4.1. This added flexibility is enabled by the careful choice of τ, which governs the trade-off among robustness, bias, and privacy. Second, by combining the Huber loss with an adaptively selected τ and incorporating covariate clipping, the proposed algorithm accommodates a much broader range of design and parameter settings, while requiring fewer

tuning parameters tied to the underlying data-generating process. This stands in contrast to the bounded design and parameter assumptions (D1) and (P1) in Cai et al. (2021). Specifically, these assumptions require $\|\mathbf{x}\|_2 \le c_{\mathbf{x}}$ almost surely and $\|\boldsymbol{\beta}^*\|_2 \le c_0$, where the constants $c_{\mathbf{x}}, c_0 > 0$ are not only essential for the theoretical analysis but also appear explicitly in the algorithm, introducing nontrivial practical constraints. Moreover, the covariate vector $\mathbf{x}$ must have zero mean with covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ satisfying $(pL)^{-1} \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le Lp^{-1}$ for some constant $L > 1$, which likewise enters the algorithm.

Remark 2 .

Algorithm 1 fits within the general framework of DP *M*-estimation considered in Avella-Medina et al. (2023). The key distinction is that, when applied to linear models with heavy-tailed and asymmetric errors, the robustification parameter τ is no longer a fixed constant. Instead, it is chosen as a function of the sample size, dimensionality, privacy level, and noise scale to balance bias, tail-robustness and privacy. In Section 4.1, we demonstrate that under certain assumptions, achieving an optimal trade-off among bias, robustness, and privacy requires selecting τ to be of order $\sigma_0(n\epsilon / p)^{1/(2+t)}$ in the case of ϵ-GDP.

Remark 3 .

Algorithm 1 has the potential to save privacy budget by leveraging privacy amplification techniques (Kasiviswanathan et al., 2008; Bassily et al., 2014; Abadi et al., 2016; Feldman et al., 2018; Wang et al., 2019). However, to ensure a fair and consistent comparison with existing methods under similar settings (Cai et al., 2021; Avella-Medina et al., 2023), we have not incorporated privacy amplification into the theoretical analysis presented in this work. We leave the incorporation of privacy amplification into our theoretical framework as a direction for future research. Moreover, we note that the theoretical framework in Feldman et al. (2018) relies on the assumption that the underlying feasible set is convex. This assumption is violated in the high-dimensional setting when applying NoisyHT (see Algorithm 2 in Section 3.3), as the feasible set defined by sparsity constraints is inherently non-convex. Developing both theoretical and empirical justifications for privacy amplification in non-convex settings remains an open challenge and may require fundamentally different analytical techniques.

Define $\boldsymbol{\Xi} := (\Xi^{jk})_{j,k\in[p]} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}$ with $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)$ and $\boldsymbol{\Omega} = \mathbb{E}(\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i^\top)$. We will show in Theorem C.3 in the supplementary material that, under certain conditions, the DP Huber estimator is asymptotically normal:

$$\sqrt{n}(\Xi^{jj})^{-1/2}(\beta_j^{(T)} - \beta_j^*) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1) \text{ as } n \to \infty.$$

To construct confidence intervals, for some $\tau_1 > 0$ and $\gamma_1 > 0$, we define the private estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$, respectively, as $\boldsymbol{\Sigma}_{\gamma_1,\epsilon} = \boldsymbol{\Sigma}_{\gamma_1} + \varsigma_1\mathbf{E}$ and $\boldsymbol{\Omega}_{\tau_1,\gamma_1,\epsilon} = \boldsymbol{\Omega}_{\tau_1,\gamma_1}(\boldsymbol{\beta}^{(T)}) + \varsigma_2\mathbf{E}$, where ς_1 and ς_2 are two noise scale parameters depending on ϵ or (ϵ,δ), and $\mathbf{E} \in \mathbb{R}^{p\times p}$ is a symmetric random matrix whose upper-triangular and diagonal entries are i.i.d. $\mathcal{N}(0,1)$. Here,

$\boldsymbol{\Sigma}_{\gamma_1} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top w_{\gamma_1}^2(\|\mathbf{x}_i\|_2)$, and $\boldsymbol{\Omega}_{\tau_1,\gamma_1}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\psi_{\tau_1}^2(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i^\top w_{\gamma_1}^2(\|\mathbf{x}_i\|_2)$, and their sensitivities are given in Section C.4 of the supplementary material. To ensure positive definiteness, we further project both $\boldsymbol{\Sigma}_{\gamma_1,\epsilon}$ and $\boldsymbol{\Omega}_{\tau_1,\gamma_1,\epsilon}$ onto the cone of positive definite matrices $\{\mathbf{H}:\mathbf{H}\succeq\zeta\mathbf{I}\}$. Specifically, we obtain $\boldsymbol{\Sigma}_{\gamma_1,\epsilon}^{+} = \arg\min_{\mathbf{H}\succeq\zeta\mathbf{I}}\|\mathbf{H}-\boldsymbol{\Sigma}_{\gamma_1,\epsilon}\|$ and $\boldsymbol{\Omega}_{\tau_1,\gamma_1,\epsilon}^{+} = \arg\min_{\mathbf{H}\succeq\zeta\mathbf{I}}\|\mathbf{H}-\boldsymbol{\Omega}_{\tau_1,\gamma_1,\epsilon}\|$ for a sufficiently small constant $\zeta>0$. Consequently, we take

$$\boldsymbol{\Xi}_{\tau_1,\gamma_1,\epsilon} := (\tilde{\Xi}^{jk}_{\tau_1,\gamma_1,\epsilon})_{j,k\in[p]} = (\boldsymbol{\Sigma}_{\gamma_1,\epsilon}^{+})^{-1}\boldsymbol{\Omega}_{\tau_1,\gamma_1,\epsilon}^{+}(\boldsymbol{\Sigma}_{\gamma_1,\epsilon}^{+})^{-1} \quad (7)$$

as a private estimator of $\boldsymbol{\Xi}$. The privacy guarantee and statistical consistency of $\boldsymbol{\Xi}_{\tau_1,\gamma_1,\epsilon}$, as well as the theoretical requirements for τ_1 and γ_1, are established in Section C.4 of the supplementary material. For any $\alpha\in(0,1)$, we construct the $100(1-\alpha)\%$ confidence interval of β_j^* as $\beta_j^{(T)} \pm z_{\alpha/2}(\tilde{\Xi}^{jj}_{\tau_1,\gamma_1,\epsilon})^{1/2}/\sqrt{n}$, where $z_{\alpha/2}$ denotes the $(1-\alpha/2)$-th quantile of $\mathcal{N}(0,1)$.

3.3 Noisy iterative hard thresholding for sparse Huber regression

In this section, we consider the high-dimensional setting, where $\boldsymbol{\beta}^*\in\mathbb{R}^p$ in (3) is sparse with $\|\boldsymbol{\beta}^*\|_0 \ll \min\{n,p\}$. A common approach to induce sparsity is ℓ_1-penalization, which is well-known for its computational efficiency and desirable theoretical properties. For sparse least absolute deviation (LAD) regression, that is, quantile regression with $\tau=1/2$, Liu et al. (2024) proposed a private estimator by reformulating the sparse LAD problem as a penalized least squares estimation and adopting a three-stage noise injection mechanism to ensure (ϵ,δ)-DP. However, the convergence rate of this private estimator is suboptimal, as it scales with $\sqrt{p}$.

From a different perspective, Cai et al. (2021) proposed a noisy variant of the iterative HT algorithm (Blumensath and Davies, 2009; Jain et al., 2014) for least squares regression and established its statistical (near-)optimality for sparse linear models with Gaussian errors. Instead of relying on soft thresholding as in Liu et al. (2024), which is closely associated with ℓ_1-penalization, a key step in Cai et al. (2021) is the adoption of the 'peeling' procedure from Dwork et al. (2021); see Algorithm 2 below. To emphasize its connection to HT, we refer to it as NoisyHT throughout the remainder of this paper. This method involves adding independent Laplace random variables to the absolute values of the entries in a given vector and then selecting the top s largest coordinates from the resulting vector. As demonstrated by Lemma D.1 in the supplementary material, when $0<\epsilon\le 0.5$, $0<\delta\le 0.011$ and $s\ge 10$, Algorithm 2 is an (ϵ,δ)-DP algorithm if the involved parameter λ satisfies $\|\mathbf{v}(\mathbf{Z})-\mathbf{v}(\mathbf{Z}')\|_\infty<\lambda$ for every pair of adjacent datasets $\mathbf{Z}$ and $\mathbf{Z}'$.

Algorithm 2 NoisyHT $(\mathbf{v}, \mathbf{Z}, s, \epsilon, \delta, \lambda)$ Algorithm (Dwork et al., 2021)

Input: Dataset $\mathbf{Z}$, vector-valued function $\mathbf{v} = \mathbf{v}(\mathbf{Z}) = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$, sparsity level *s*, privacy parameters (ϵ, δ), and noise scale λ.

1: Initialize $S = \varnothing$;

2: **for** $i \in [s]$ **do**

3: Generate $\mathbf{w}_i = (w_{i,1}, \ldots, w_{i,p})^\top \in \mathbb{R}^p$ with $w_{i,1}, \ldots, w_{i,p} \overset{\text{i.i.d.}}{\sim}$ Laplace $(2\epsilon^{-1}\lambda\sqrt{5s\log(1/\delta)})$;

4: Append $j^* = \arg\max_{j \in [p] \setminus S}(|v_j| + w_{i,j})$ to *S*;

5: **end for**

6: Set $\tilde{P}_s(\mathbf{v}) = \mathbf{v}_S$;

7: Generate $\mathbf{w} = (\tilde{w}_1, \ldots, \tilde{w}_p)^\top \in \mathbb{R}^p$ with $\tilde{w}_1, \ldots, \tilde{w}_p \overset{\text{i.i.d.}}{\sim}$ Laplace $(2\epsilon^{-1}\lambda\sqrt{5s\log(1/\delta)})$;

Output: $\tilde{P}_s(\mathbf{v}) + \mathbf{w}_S \in \mathbb{R}^p$.

In high-dimensional settings, the noisy clipped gradient decent step involves calculating the intermediate update and the noisy update sequentially. The intermediate update involving clipped gradients is defined as

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \frac{\eta_0}{n}\sum_{i=1}^{n}\psi_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(t)})\mathbf{x}_i w_\gamma(\|\mathbf{x}_i\|_\infty), \quad t \in \{0\} \cup [T-1]. \quad (8)$$

The noisy update $\boldsymbol{\beta}^{(t+1)}$ is then obtained by using $\boldsymbol{\beta}^{(t+1)}$ from (8) as the input to Algorithm 2. Due to $\sup_{u \in \mathbb{R}} |\psi_\tau(u)| \le \tau$, we know $\sup_{(y,\mathbf{x}) \in \mathbb{R} \times \mathbb{R}^p, \boldsymbol{\beta} \in \mathbb{R}^p} \|\psi_\tau(y - \mathbf{x}^\top \boldsymbol{\beta})\mathbf{x} w_\gamma(\|\mathbf{x}\|_\infty)\|_\infty \le \gamma\tau$, which implies the ℓ_∞-sensitivity of the procedure for calculating the intermediate update is bounded by $2\eta_0\gamma\tau/n$. By setting the privacy parameters to $(\epsilon/T, \delta/T)$ and choosing the noise scale as $\lambda = 2\eta_0\gamma\tau/n$ in Algorithm 2, Lemma D.1 in the supplementary material indicates that each noisy clipped gradient descent step is $(\epsilon/T, \delta/T)$-DP. After *T* iterations, the standard composition property (Lemma 3) implies that the *T*-th iterate $\boldsymbol{\beta}^{(T)}$ is (ϵ, δ)-DP. The complete algorithm is presented in Algorithm 3. Alternatively, by setting the privacy parameters to $(\epsilon\sqrt{2/\{5T\log(2/\delta)\}}, \delta/(2T))$ and choosing the noise scale as $\lambda = 2\eta_0\gamma\tau/n$ in Algorithm 2, Lemma D.1 and the advanced composition property (Lemma 4) imply that the *T*-th iterate $\boldsymbol{\beta}^{(T)}$ is also (ϵ, δ)-DP, provided that the privacy constraints (1) hold. In the practical implementation, the choice of privacy parameters is guided by which configuration results in a smaller scale of the Laplace random variables injected in NoisyHT.

Algorithm 3 DP Sparse Huber Regression

Input: Dataset $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, learning rate η_0, number of iterations *T*, truncation parameter γ, robustification parameter τ, privacy parameters (ϵ, δ), sparsity level *s*, and initial value $\boldsymbol{\beta}^{(0)}$.

1: **for** $t = 0, \ldots, T-1$ **do**

2: Compute

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \frac{\eta_0}{n} \sum_{i=1}^{n} \psi_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(t)}) \mathbf{x}_i w_\gamma(\|\mathbf{x}_i\|_\infty),$$

$$\boldsymbol{\beta}^{(t+1)} = \text{NoisyHT}(\boldsymbol{\beta}^{(t+1)}, \{(y_i, \mathbf{x}_i)\}_{i=1}^n, s, \epsilon/T, \delta/T, 2\eta_0\gamma\tau/n).$$

3: **end for**

Output: $\boldsymbol{\beta}^{(T)}$.

Remark 4 .

Algorithm 3 integrates a noisy gradient-based descent method with Huber regression to simultaneously ensure differential privacy and robustness against heavy-tailed errors, provided that τ is properly tuned, as discussed in Section 4. It is applicable to a broader class of random features, including those following a normal distribution. Since the Huber loss has a bounded derivative, the residuals are truncated by the function $\psi_\tau(\cdot)$, eliminating the need to truncate the response variables. In practice, truncating the response variable can be problematic, particularly when it is positive and follows a right-skewed distribution, such as wages or prices. In contrast, residuals are expected to fluctuate around zero, making their truncation more reasonable.

4 Statistical Analysis of Private Huber Regression

For the theoretical analysis, we impose the following assumptions on the distributions of the random covariates and errors under the linear model (3).

Assumption 1 .

For each $i \in [n]$, the random covariate vector $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^\top$ with $x_{i,1} \equiv 1$ satisfies: (i) $\mathbb{E}(x_{i,j}) = 0$ for $j \in [p] \setminus \{1\}$, and (ii) $\mathbb{P}(|\langle \mathbf{u}, \boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i \rangle| \geq \upsilon_1 z) \leq 2e^{-z^2/2}$ for all $\mathbf{u} \in \mathbb{S}^{p-1}$ and $z \geq 0$, where $\upsilon_1 \geq 1$ is a dimension-free constant and $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{x}_i \mathbf{x}_i^\top)$. Moreover, there exist constants $\lambda_1 \geq \lambda_p > 0$ such that $\lambda_p \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq \lambda_1$.

Assumption 2 .

The regression errors $\{\varepsilon_i\}_{i=1}^n$ are independent random variables satisfying $\mathbb{E}(\varepsilon_i|\mathbf{x}_i)=0$ and $\mathbb{E}(\varepsilon_i^2|\mathbf{x}_i)\le\sigma_0^2$ almost surely. Moreover, $\mathbb{E}(|\varepsilon_i|^{2+\iota}|\mathbf{x}_i)\le\sigma_\iota^{2+\iota}$ almost surely for some $\iota\ge 0$ and $\sigma_0\le\sigma_\iota<\infty$.

Assumption 1 imposes a sub-Gaussian condition on the random covariates, generalizing the standard one-dimensional sub-Gaussian assumption to random vectors. This condition also complements the bounded design assumptions (D1) and (D1′) in Cai et al. (2021). Various types of random vectors fulfill this assumption, for example: (i) Gaussian and Bernoulli random vectors, (ii) random vectors uniformly distributed on the Euclidean sphere or ball centered at the origin with radius $\sqrt{p}$, and (iii) random vectors uniformly distributed on the unit cube $[-1,1]^p$. For more detailed discussions of high-dimensional sub-Gaussian distributions, including discrete cases, we refer to Chapter 3.4 in Vershynin (2018). Let $\kappa_4=\sup_{\mathbf{u}\in\mathbb{S}^{p-1}}\mathbb{E}(\langle\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i,\mathbf{u}\rangle^4)$, which serves as an upper bound on the kurtoses of all marginal projections of the normalized covariate vectors. It can be shown that $\kappa_4\le C\upsilon_1^4$ for some absolute constant $C>1$. Assumption 2 relaxes the Gaussian error condition (4.1) in Cai et al. (2021) by allowing for heavy-tailed error distributions, such as the t_ν-distribution with $\nu>2$, the centered lognormal distribution, and the centered Pareto distribution with shape parameter greater than 2. Our theoretical results show that finite conditional second moment condition (i.e., $\iota=0$) suffices for coefficient estimation (see Theorems 1 and 2), while the stronger moment condition $\iota>0$ is required only for inference (see Theorem C.3 in the supplementary material). Given a dataset $\{(y_i,\mathbf{x}_i)\}_{i=1}^n$ satisfying Assumptions 1 and 2, recall

$$\mathcal{L}_\tau(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^n\rho_\tau(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}) \quad \text{with } \rho_\tau(u)=\frac{u^2}{2}1(|u|\le\tau)+(\tau|u|-\frac{\tau^2}{2})1(|u|>\tau)$$

denotes the empirical Huber loss. In the following analysis, all constants depending on $(\upsilon_1,\lambda_1,\lambda_p,\kappa_4)$ are absorbed into the notation $\lesssim$, $\gtrsim$, and $\asymp$.

4.1 Low-dimensional setting

In this section, we establish the statistical properties of the DP Huber estimator $\boldsymbol{\beta}^{(T)}$ as defined in Algorithm 1. As a benchmark, let $\boldsymbol{\beta}_\tau\in\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\mathcal{L}_\tau(\boldsymbol{\beta})$ denote the non-private Huber estimator in the low-dimensional setting. Given an initial estimate $\boldsymbol{\beta}^{(0)}$, the statistical analysis of the noisy clipped gradient descent iterates $\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T$ relies heavily on the properties of the empirical Huber loss $\mathcal{L}_\tau(\cdot)$, including its local strong convexity and global smoothness. The key observation is that the Huber loss $\rho_\tau(u)$ is strongly convex only when $|u|<\tau$, with its second-order derivative $\rho_\tau''(u)=1$ in this region. Our analysis consists of two parts. First, we establish

that, by conditioning on a series of 'good events' associated with the empirical Huber loss, the noisy clipped gradient descent iterates exhibit favorable convergence properties. Second, we prove that these good events occur with high probability under Assumptions 1 and 2. Due to space limitations, intermediate results that hold conditioned on good events are provided in the supplementary material.

Given that the initial value $\boldsymbol{\beta}^{(0)}$ lies within a neighborhood of the non-private Huber estimator $\boldsymbol{\beta}_\tau$, Theorem 1 below establishes a high-probability bound for $\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2$.

Theorem 1 .

Let Assumptions 1 and 2 hold. Assume that $\boldsymbol{\beta}^{(0)} \in \boldsymbol{\beta}_\tau + \mathbb{B}^p(r_0)$ for some $r_0 \asymp \tau$, and that the learning rate satisfies $\eta_0 = \eta/(2\lambda_1)$ for some $\eta \in (0,1]$. Moreover, let $\gamma \asymp \sqrt{p + \log n}$, $T \asymp \log\{r_0 n\epsilon(\sigma_0 p)^{-1}\}$, and $\tau \asymp \tau_0\{n\epsilon(p+\log n)^{-1}\}^{1/(2+\iota)}$ for some $\tau_0 \geq \sigma_0$. Then, if the noise scale σ defined in (6) satisfies $\sigma\sqrt{p + \log T + \log n} \lesssim r_0$, the DP Huber estimator $\boldsymbol{\beta}^{(T)}$ obtained in Algorithm 1 satisfies

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \underbrace{\frac{\sigma_0 p}{n\epsilon} + \sigma\sqrt{\frac{p+\log n}{n\epsilon}}}_{\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}_\tau\|_2} + \underbrace{\max\{\sigma_\iota^{2+\iota}\tau_0^{-1-\iota}, \tau_0\}\left(\frac{p+\log n}{n\epsilon}\right)^{(1+\iota)/(2+\iota)} + \sigma_0\sqrt{\frac{p+\log n}{n}}}_{\|\boldsymbol{\beta}_\tau - \boldsymbol{\beta}^*\|_2}$$

with probability at least $1 - Cn^{-1}$, provided that $n\epsilon \gtrsim C_{\tau_0,\sigma_\iota}(p + \log n)$, where C_{τ_0,σ_ι} is a positive constant depending only on (τ_0, σ_ι).

The selection of robustification parameter τ in Theorem 1 is intended to balance bias and robustness, while accounting for the effective sample size $n\epsilon$. See the proof of Proposition C.1 in Section C.3.4 of the supplementary material for details. The initial condition $\boldsymbol{\beta}^{(0)} \in \boldsymbol{\beta}_\tau + \mathbb{B}^p(r_0)$ in Theorem 1 is not restrictive. As shown in Theorem C.2 of the supplementary material, for any initial point $\boldsymbol{\beta}^{(0)}$ satisfying $\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}_\tau\|_2 \geq r_0$, we have $\|\boldsymbol{\beta}^{(T_0)} - \boldsymbol{\beta}_\tau\|_2 \leq r_0$ with high probability for some $T_0 \in \mathbb{N}_+$, provided that the sample size n is sufficiently large. Moreover, since $T \asymp \log(n\epsilon)$ and each iteration in Algorithm 1 incurs a computational cost of $O(np)$, the total complexity of Algorithm 1 is $O\{np\log(n\epsilon)\}$.

Let $\tau_0 \asymp \sigma_0$ in Theorem 1. The robustification parameter τ is then required to satisfy $\tau \asymp \sigma_0\{n\epsilon(p+\log n)^{-1}\}^{1/(2+\iota)}$. Based on Theorem 1, we now present the final convergence

guarantees for the DP Huber estimator $\boldsymbol{\beta}^{(T)}$ obtained from Algorithm 1, under different choices of the noise scale σ as specified in (6). The results are summarized as follows:

(i) (ϵ-GDP) As long as $n\epsilon \gtrsim \sqrt{T}(p+\log n)$, the ϵ-GDP Huber estimator $\boldsymbol{\beta}^{(T)}$ obtained in Algorithm 1 by setting $\sigma = \sigma_{\text{gdp}}$ satisfies, with probability at least $1-Cn^{-1}$, that

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \sigma_0\{(\frac{\sigma_\iota}{\sigma_0})^{2+\iota} + \sqrt{T}\}(\frac{p+\log n}{n\epsilon})^{(1+\iota)/(2+\iota)} + \sigma_0\sqrt{\frac{p+\log n}{n}}. \quad (9)$$

(ii) ((ϵ,δ)-DP) As long as $n\epsilon \gtrsim T(p+\log n)\sqrt{\log(T/\delta)}$, the (ϵ,δ)-DP Huber estimator $\boldsymbol{\beta}^{(T)}$ obtained in Algorithm 1 by setting $\sigma = \sigma_{\text{dp}}$ satisfies, with probability at least $1-Cn^{-1}$,

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \sigma_0\{(\frac{\sigma_\iota}{\sigma_0})^{2+\iota} + T\sqrt{\log(\frac{T}{\delta})}\}(\frac{p+\log n}{n\epsilon})^{(1+\iota)/(2+\iota)} + \sigma_0\sqrt{\frac{p+\log n}{n}}. \quad (10)$$

In the error bounds (9) and (10), the second term reflects the convergence of the non-private Huber estimator $\boldsymbol{\beta}_\tau$, while the first term represents the privacy cost. Notably, $n\epsilon$ can be interpreted as the effective sample size under privacy constraints, and the choice of τ is influenced by this effective sample size. For normally distributed errors, Cai et al. (2021) established a minimax lower bound for (ϵ,δ)-DP estimation of $\boldsymbol{\beta}^*$ under ℓ_2-risk. The lower bound is of order $\sigma_0\{\sqrt{pn^{-1}} + p(n\epsilon)^{-1}\}$ when $\epsilon \in (0,1)$ and $\delta < n^{-1-\omega}$ for some fixed constant $\omega > 0$. In comparison, the slower term $\{p(n\epsilon)^{-1}\}^{1-1/(2+\iota)}$, ignoring logarithmic factors, in (9) and (10) explicitly captures the combined influence of heavy-tailedness and privacy.

Remark 5 .

Another notable implication of our results is that DP Huber regression achieves a near-optimal convergence rate for sub-Gaussian errors. In addition to Assumption 1, assume that the regression error ε_i satisfies $\mathbb{E}(e^{\lambda\varepsilon_i}|\mathbf{x}_i) \le e^{\sigma_0^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$. In this case, we instead choose $\tau \asymp \sigma_0\sqrt{\log\{n\epsilon(p+\log n)^{-1}\}}$. The ℓ_2-error of the resulting ϵ-GDP Huber estimator is, up to constant factors, upper bounded by

$$\frac{p+\log n}{n\epsilon}\cdot\sigma_0\sqrt{T\log(\frac{n\epsilon}{p+\log n})} + \sigma_0\sqrt{\frac{p+\log n}{n}}$$

with probability at least $1-Cn^{-1}$ as long as $n\epsilon \gtrsim \sqrt{T}(p+\log n)$. Compared to Theorem 4.2 in Cai et al. (2021), the above results hold without requiring (i) $\|\boldsymbol{\beta}^*\|_2 < c_0$ for some constant $c_0 > 0$ that appears in Algorithm 4.1 and Theorem 4.2 of Cai et al. (2021), and (ii) $\|\mathbf{x}_i\|_2 < c_{\mathbf{x}}$ with

probability one for some dimension-free constant $c_{\mathbf{x}}$. Moreover, the sample size requirement is relaxed from $n\epsilon \gtrsim p^{3/2}$ to $n\epsilon \gtrsim p$, ignoring logarithmic factors.

4.2 High-dimensional setting

In this section, we establish the statistical properties of the sparse DP Huber estimator $\boldsymbol{\beta}^{(T)}$ as defined in Algorithm 3. Compared to the low-dimensional setting, the sparsity level *s* is prespecified to perform HT on the high-dimensional gradient vector after the injection of Laplace noises. Analogous to the low-dimensional case, our analysis proceeds in two parts. First, we establish that by conditioning on a series of 'good events' associated with the empirical Huber loss, the noisy clipped gradient descent iterates exhibit favorable convergence properties. Second, we prove that these good events occur with high probability under Assumptions 1 and 2. All intermediate results that hold conditionally on the good events are provided in the supplementary material.

Let $\mathbb{H}(s) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 \leq s\}$ denote the set of *s*-sparse vectors in $\mathbb{R}^p$, and let $\Theta(r) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r\}$ be the ball of radius *r* centered at $\boldsymbol{\beta}^*$. Theorem 2 below provides a high-probability bound for $\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2$, given that the initial value $\boldsymbol{\beta}^{(0)}$ lies within some neighborhood of $\boldsymbol{\beta}^*$.

Theorem 2 .

Let Assumptions 1 and 2 hold. Assume that $\boldsymbol{\beta}^{(0)} \in \mathbb{H}(s) \cap \Theta(r_0)$ for some $r_0 \asymp \tau \geq 16\sigma_0$, and the learning rate satisfies $\eta_0 = \eta/(2\lambda_1)$ for some $\eta \in (0,1)$. Moreover, let $\gamma \asymp \sqrt{\log(pn)}$ and $T \asymp \log\{r_0 n\epsilon(\sigma_0 \log p)^{-1}\}$. Write $s^* := \|\boldsymbol{\beta}^*\|_0$. Then, the sparse DP Huber estimator $\boldsymbol{\beta}^{(T)}$ obtained from Algorithm 3 satisfies

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \underbrace{\frac{\sigma_t^{2+t}}{\tau^{1+t}} + \sigma_0\sqrt{\frac{s\log(ep/s)+\log n}{n}} + \tau\frac{s\log(ep/s)+\log n}{n}}_{\tilde{r}} + \frac{\sigma_0 \log p}{n\epsilon} + C_{\eta,1}\frac{sT\tau\{\log(pn)\}^{3/2}}{n\epsilon}\sqrt{\log(\frac{T}{\delta})}$$

with probability at least $1 - Cn^{-1}$, provided that $s \geq s^* \max\{192(1+\eta^{-2})(8\lambda_1/\lambda_p)^2, 16\eta/(1-\eta)\}$, $\tilde{r} \lesssim r_0$, and $n\epsilon \gtrsim C_{\eta,2} sT\{\log(pn)\}^{3/2}\sqrt{\log(T/\delta)}$, where $C_{\eta,1}$ and $C_{\eta,2}$ are two positive constants depending only on η.

By setting $\tau \asymp \sigma_0(n\epsilon)^{1/(2+t)}\{s\log(ep/s)+\log n\}^{-1/(2+t)}$, Theorem 2 implies that as long as $n\epsilon \gtrsim sT\{\log(pn)\}^{3/2}\sqrt{\log(T/\delta)}$, the sparse DP Huber estimator $\boldsymbol{\beta}^{(T)}$ satisfies, with probability at least $1 - Cn^{-1}$, that

$$\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^{*}\|_2 \lesssim \frac{\sigma_t^{2+t}}{\sigma_0^{1+t}}\{\frac{s\log(ep/s)+\log n}{n\epsilon}\}^{(1+t)/(2+t)}\{\log(pn)\}^{3/2}T\sqrt{\log(\frac{T}{\delta})} + \sigma_0\sqrt{\frac{s\log(ep/s)+\log n}{n}}. \quad (11)$$

In the estimation error bound given by (4.2), the second term represents the convergence rate of the non-private Huber estimator, while the first term accounts for the privacy cost, influenced by a carefully chosen robustification parameter τ. The robustification parameter τ is specifically chosen based on the effective sample size $n\epsilon$, along with the intrinsic dimensionality and noise level. This selection aims to achieve a balance among robustness, bias, and privacy. For normally distributed errors, Cai et al. (2021) derived a minimax lower bound for (ϵ,δ)-DP estimation of $\boldsymbol{\beta}^{*}$ under ℓ_2-risk in the high dimensions. The lower bound is of the order $\sigma_0\{\sqrt{s^{*}n^{-1}\log p}+s^{*}(n\epsilon)^{-1}\log p\}$, which holds under the conditions $\epsilon\in(0,1)$, $s^{*}\ll p^{1-\omega}$, and $\delta<n^{-1-\omega}$ for some fixed constant $\omega>0$. In contrast, the slower term $\{s(n\epsilon)^{-1}\log p\}^{1-1/(2+t)}$ (ignoring logarithmic factors) in (4.2) highlights the joint effects of heavy-tailedness and privacy considerations more explicitly. Moreover, each iteration in Algorithm 3 consists of an intermediate update with complexity $O(np)$ and a noisy update (NoisyHT) with complexity $O(ps)$. Since $s\ll n$ and $T\asymp\log(n\epsilon)$, the overall complexity for Algorithm 3 is $O\{np\log(n\epsilon)\}$.

Remark 6 .

The result in (4.2) shows that the sparse (ϵ,δ)-DP Huber estimator achieves a near-optimal convergence rate for sub-Gaussian errors in high dimensions. Similar to Remark 5, assume that the regression error ε_i is sub-Gaussian, satisfying $\mathbb{E}(e^{\lambda\varepsilon_i}|\mathbf{x}_i)\le e^{\sigma_0^2\lambda^2/2}$ for all $\lambda\in\mathbb{R}$. In this case, we set the robustification parameter as

$$\tau\asymp\sigma_0\sqrt{\log\{\frac{n\epsilon}{s\log(ep/s)+\log n}\}}.$$

The ℓ_2-error of the resulting sparse DP Huber estimator $\boldsymbol{\beta}^{(T)}$ obtained in Algorithm 3, up to constant and logarithmic factors, is upper bounded by

$$\sigma_0\frac{s\log(ep/s)+\log n}{n\epsilon}+\sigma_0\sqrt{\frac{s\log(ep/s)+\log n}{n}}$$

with probability at least $1-Cn^{-1}$, provided that $n\epsilon\gtrsim sT\{\log(pn)\}^{3/2}\sqrt{\log(T/\delta)}$. Compared to Theorem 4.4 in Cai et al. (2021), our results do not require the following assumptions: (i) $\|\boldsymbol{\beta}^{*}\|_2<c_0$ for some constant $c_0>0$, which is needed by Algorithm 4.2 and Theorem 4.4 of Cai et al. (2021), and (ii) $\sqrt{|I|}\,\|\mathbf{x}_{i,I}\|_\infty<c_{\mathbf{x}}$ with probability one for all subsets $I\subseteq[p]$ with $|I|\ll n$,

where $c_{\mathbf{x}}$ is a dimension-free constant. Additionally, we relax the sample size requirement from $n\epsilon \gtrsim (s^*)^{3/2}$ to $n\epsilon \gtrsim s^*$, up to logarithmic factors.

We have analyzed the trade-offs among bias, privacy, and robustness in $\boldsymbol{\beta}^{(T)}$, and compared Algorithm 3 with the privatized OLS method proposed by Cai et al. (2021) under sub-Gaussian noise conditions. Our analysis shows that when ε_i follows a sub-Gaussian distribution, the estimator obtained from Algorithm 3 achieves a near-optimal convergence rate, while requiring less restrictive sample size conditions and accommodating broader classes of $\boldsymbol{\beta}^*$ and $\mathbf{x}_i$, all while maintaining differential privacy guarantees. We now examine the statistical convergence of Algorithm 3 in a non-private setting. Let $\check{\boldsymbol{\beta}}^{(T)}$ be the non-private sparse Huber estimator obtained from Algorithm 3 by eliminating the noise terms $\mathbf{w}_1^t, \ldots, \mathbf{w}_s^t, \mathbf{w}^t$ for all iterations.

Corollary 1, which is of independent interest, establishes the convergence rate of $\check{\boldsymbol{\beta}}^{(T)}$ under conditions similar to those in Theorem 2. The proof of Corollary 1 follows closely that of Theorem 2, with the noise terms $\{\mathbf{w}_1^t, \ldots, \mathbf{w}_s^t, \mathbf{w}^t\}_{t=0}^{T-1}$ set to zero.

Corollary 1 .

Assume that $\boldsymbol{\beta}^{(0)} \in \mathbb{H}(s) \cap \Theta(r_0)$ for some $r_0 > 0$, and the learning rate satisfies $\eta_0 = \eta / (2\lambda_1)$ for some $\eta \in (0,1)$. Then, with $\tau \asymp \sigma_0 n^{1/(2+t)} \{s \log(ep/s) + \log n\}^{-1/(2+t)}$, $r_0 \asymp \tau$, and $T \asymp \log(n / \log p)$, the non-private sparse Huber estimator $\check{\boldsymbol{\beta}}^{(T)}$ satisfies

$$\|\check{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \sigma_0 \sqrt{\frac{s \log(ep/s) + \log n}{n}}$$

with probability at least $1 - Cn^{-1}$, provided that $n \gtrsim s \log p + \log n$ and $s \gtrsim s^*$.

5 Numerical Studies

In this section, we conduct simulation studies to evaluate the numerical performance of the DP Huber regression in both low- and high-dimensional settings. The data $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ are generated from the linear model $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \sqrt{b}\varepsilon_i$, where $\mathbf{x}_i = (1, \mathbf{x}_{i,-1}^\top)^\top$ and $\mathbf{x}_{i,-1} \in \mathbb{R}^{p-1}$ follows a distribution that varies across different settings. The noise variables ε_i are i.i.d. and follow either a standard normal distribution $\mathcal{N}(0,1)$ or a heavy-tailed $t_{2.25}$ distribution; the constant b determines the noise scale. The construction of $\boldsymbol{\beta}^*$ differs across scenarios. In the low-dimensional case, each component of $\boldsymbol{\beta}^*$ is independently set to a or $-a$ with equal probability. In the high-dimensional setting, only the first s^* entries are generated in the same manner and the remaining $p - s^*$ components are set to zero. The signal-to-noise ratio (SNR) is defined as

$\mathrm{Var}(\mathbf{x}_i^\top \boldsymbol{\beta}^*)/\{b\mathrm{Var}(\varepsilon_i)\}$. For a fixed design distribution and noise law, larger values of a^2/b correspond to higher SNR.

5.1 Selection of tuning parameters

Recognizing that Algorithms 1 and 3 both rely on a sequence of tuning parameters, we begin by outlining the selection principles adopted in our implementation, as these choices are essential for practical performance and reproducibility. Although our theory permits a broad class of initial values $\boldsymbol{\beta}^{(0)}$ satisfying $\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2 \le r_0$ for some $r_0 > 0$ that may even diverge with *n*, our numerical studies, especially in high-dimensional settings, show that a well-chosen initialization can markedly improve finite-sample behavior. Motivated by Liu et al. (2024), we therefore use, in both low- and high-dimensional scenarios, a private initial estimator obtained by solving a ridge-penalized Huber regression followed by output perturbation. Additional details on this private initialization are provided below. Throughout this section, we fix $\delta = 10n^{-1.1}$ and present tuning choices under the (ϵ, δ)-DP framework. The corresponding rules for ϵ-GDP, along with proofs of the privacy guarantee for the initialization step, are deferred to Section F of the supplementary material.

5.1.1 Selection of tuning parameters in Algorithm 1

We choose the initial value $\boldsymbol{\beta}^{(0)}$ in a data-dependent manner. To ensure that the overall procedure remains (ϵ, δ)-DP, we divide the privacy budget between the initialization step and Algorithm 1, which we denote by $(\epsilon_{\text{init}}, \delta_{\text{init}})$ and $(\epsilon_{\text{main}}, \delta_{\text{main}})$, respectively. Specifically, we set $\epsilon_{\text{init}} = \epsilon_{\text{main}}/5 = \epsilon/6$ and $\delta_{\text{init}} = \delta_{\text{main}}/5 = \delta/6$.

Initialization of $\boldsymbol{\beta}^{(0)}$. We estimate $\boldsymbol{\beta}^{(0)}$ by solving a ridge-penalized Huber regression with row-wise clipping:

$$\boldsymbol{\beta}^{(0)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\frac{1}{n}\sum_{i=1}^{n}\rho_{\tau_0}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}) + \frac{\lambda_0}{2}\|\boldsymbol{\beta}\|_2^2\}, \quad (12)$$

where $\lambda_0 = 0.2$ and $\mathbf{x}_i = (1, \mathbf{x}_{i,-1}^\top)^\top$ with $\mathbf{x}_{i,-1} = \mathbf{x}_{i,-1}\min\{\sqrt{p}/(6\|\mathbf{x}_{i,-1}\|_2), 1\}$. For this initialization step, the privacy parameters are allocated as $(\epsilon_{\text{init}}/4, 0)$ for selecting τ_0 and $(3\epsilon_{\text{init}}/4, \delta_{\text{init}})$ for the output perturbation.

• Selection of τ_0. For each $i \in [n]$, define $\tilde{y}_i = \min\{\log n, \max\{-\log n, y_i\}\}$. Let $m_1 = n^{-1}\sum_{i=1}^{n}\tilde{y}_i + w_1$ and $m_2 = n^{-1}\sum_{i=1}^{n}\tilde{y}_i^2 + w_2$, with $w_1 \sim \text{Laplace}(16(n\epsilon_{\text{init}})^{-1}\log n)$ and $w_2 \sim \text{Laplace}(8(n\epsilon_{\text{init}})^{-1}(\log n)^2)$. Set τ_0 as $\sqrt{m_2 - m_1^2}$ if $m_2 - m_1^2 > 0$ and 2 otherwise.

• Output perturbation. Let $\tilde{\sigma} = 8B\tau_0(3n\epsilon_{\text{init}}\lambda_0)^{-1}\sqrt{2\log(1.25/\delta_{\text{init}})}$ and $B = \sqrt{1+p/36}$. We then set $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}^{(0)} + \tilde{\sigma}\cdot\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Selection of other parameters. We set $\eta_0 = 0.2$, $\gamma = 0.5\sqrt{p+\log n}$, and $T = \lceil 2\log n \rceil$. Recall that the optimal choice of τ satisfies $\tau \asymp \sigma_0\{n\epsilon(p+\log n)^{-1}\}^{1/2}$ (here we fix $\iota = 0$ for simplicity). We then set $\tau = 0.04\tau_0\sqrt{n\epsilon(p+\log n)^{-1}}$, where τ_0 defined in (12) serves as a rough upper bound for σ_0.

5.1.2 Selection of tuning parameters in Algorithm 3

Similar to Section 5.1.1, we divide the overall privacy budget between the initialization step and Algorithm 3, using $(\epsilon_{\text{init}}, \delta_{\text{init}})$ for initialization and $(\epsilon_{\text{main}}, \delta_{\text{main}})$ for the main algorithm. We set $\epsilon_{\text{init}} = 2\epsilon_{\text{main}} = 2\epsilon/3$ and $\delta_{\text{init}} = \delta_{\text{main}} = \delta/2$, which ensures that the resulting estimator satisfies the desired (ϵ, δ)-DP guarantee.

Initialization of $\boldsymbol{\beta}^{(0)}$. In high-dimensional sparse models, only the covariates corresponding to the true support of $\boldsymbol{\beta}^*$ carry meaningful signal. Guided by this observation, we construct the initial estimator $\boldsymbol{\beta}^{(0)}$ in two stages. First, we obtain a DP estimate of the support using a clipped gradient procedure. Second, conditional on the selected support, we compute a private initial estimator restricted to that subset of covariates. To ensure privacy, we allocate the budget for these two components as $(\epsilon_{\text{init}}/2, 0)$ for the support recovery step and $(\epsilon_{\text{init}}/2, \delta_{\text{init}})$ for subsequent private estimation step.

• Support recovery. For each $j \in \{2, \ldots, p\}$, define $u_{i,j} = y_i x_{i,j}$. Set $g_j = |n^{-1}\sum_{i=1}^{n}\tilde{u}_{i,j}|$, $\tilde{u}_{i,j} = u_{i,j}\min\{\sqrt{\log(pn)}/|u_{i,j}|, 1\}$, with the convention $\tilde{u}_{i,j} = 0$ when $u_{i,j} = 0$. Let $s_0 = s-1$. We then apply the Report Noisy Max procedure (Dwork and Roth, 2014, Claim 3.9) to $(g_2, \ldots, g_p)$ in a peeling scheme: run it s_0 times, each time adding independent Laplace noise $\text{Laplace}(2s_0\Delta/\epsilon_{\text{init}})$ with $\Delta = 2n^{-1}\sqrt{\log(pn)}$, and select the top-$s_0$ indices without replacement. This yields the support $\mathcal{S}_0 = (j_{(1)}, \ldots, j_{(s_0)})$.

• Initialize $\boldsymbol{\beta}^{(0)}$. We first obtain an $(\epsilon_{\text{init}}/2, \delta_{\text{init}})$-DP estimator $\boldsymbol{\beta}_{\text{init}}$ by applying the initialization procedure described in Section 5.1.1 to $\{(y_i, \mathbf{x}_{i,\mathcal{S}})\}_{i=1}^n$, where $\mathcal{S} = \{1\} \cup \mathcal{S}_0$. We then embed this estimator into $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$ by setting $\boldsymbol{\beta}^{(0)}_{\mathcal{S}} = \boldsymbol{\beta}_{\text{init}}$ and $\boldsymbol{\beta}^{(0)}_{[p]\setminus\mathcal{S}} = \mathbf{0}$.

Selection of other parameters. We set $\gamma = 0.5\sqrt{\log(pn)}$, take $T = \lceil 2\log n \rceil$, and choose the working sparsity level as $s = \lceil 1.2s^* \rceil$. The learning rate is fixed at $\eta_0 = 0.01$. Recall that the optimal order of the robustification parameter satisfies $\tau \asymp \sigma_0\{n\epsilon(s\log p + \log n)^{-1}\}^{1/2}$ (here we fix $\iota = 0$ for simplicity). We then set $\tau = 0.04\tau_0\sqrt{n\epsilon(s\log p + \log n)^{-1}}$, where the choice of τ_0 is described in Section 5.1.1; consistent with the low-dimensional procedure, this same quantity is also used in constructing the initial estimator $\boldsymbol{\beta}^{(0)}$.

5.2 Low-dimensional setting

In the low-dimensional setting, each entry of $\mathbf{x}_{i,-1}$ is independently drawn from either $\mathcal{N}(0,1)$ or $\text{Uniform}(-\sqrt{3},\sqrt{3})$. The signal and noise scale parameters *a* and *b* take values in $\{0.5,1,2\}$. As a benchmark, we consider the non-private Huber estimator obtained by running Algorithm 1 without initialization, clipping, and noise injection. For this estimator, we set $\eta_0 = 0.5$, $T = \lceil 2\log n \rceil$, and update the robustification parameter according to $\tau = 0.2\hat{\sigma}_0\sqrt{n(p+\log n)^{-1}}$, where $\hat{\sigma}_0^2 = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ with $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$.

We set the dimension $p \in \{5,10,20\}$ and let the sample size *n* range from 2500 to 10000 to examine how the relative ℓ_2-error, $\|\overset{\circ}{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 / \|\boldsymbol{\beta}^*\|_2$, varies with both dimension and sample size. All results are computed over 300 repetitions, and $\overset{\circ}{\boldsymbol{\beta}}$ denotes the estimator being evaluated. We first investigate the effect of the initialization scheme on the accuracy of the (ϵ,δ)-DP estimator using the relative ℓ_2-error across different sample sizes; see Figure 1. Figure 2 further compares the (ϵ,δ)-DP Huber estimators at multiple privacy levels with the non-private Huber estimator. As expected, weaker privacy protection (large ϵ) yields improved estimation accuracy. In addition, when the dimension is fixed at $p = 10$, Table 1 shows that the estimation error of the (ϵ,δ)-DP Huber estimator decreases as the SNR increases. Additional results at $p \in \{5,20\}$, along with finite-sample performance of the ϵ-GDP Huber estimators, are given in supplementary material (see Tables S3, S4 and S6–S8).

Next, we compare the coverage performance of the privatized confidence intervals (CIs) introduced in Section 3.2 with both their non-private counterparts and the private bootstrap procedure described in Algorithm 2 of Ferrando et al. (2022) [1], for which the required data ranges are calibrated using the empirical maxima of each simulated dataset. The parameters (τ_1,γ_1) used to construct $\boldsymbol{\Xi}_{\tau_1,\gamma_1,\epsilon}$ in (7) are set as $\gamma_1 = 0.5\sqrt{p+\log n}$ and $\tau_1 = 0.95\tau_0\sqrt{n\epsilon(p+\log n)^{-1}}$, where τ_0 is selected according to Section 5.1.1. The pair

$$(\varsigma_1,\varsigma_2) = (2\gamma_1^2(n\epsilon)^{-1}\sqrt{2\log(1.25/\delta)}, 2\gamma_1^2\tau_1^2(n\epsilon)^{-1}\sqrt{2\log(1.25/\delta)})$$

follows from Lemma C.8 in the supplementary material. We allocate the total privacy budget (ϵ, δ) across the initialization, main estimation, and inference steps as $(\epsilon_{\text{init}}, \delta_{\text{init}})$, $(\epsilon_{\text{main}}, \delta_{\text{main}})$, and $(\epsilon_{\text{infer}}, \delta_{\text{infer}})$, respectively, with $\epsilon_{\text{init}} = \epsilon_{\text{infer}} = \epsilon_{\text{main}} / 4 = \epsilon / 6$, and the same proportions applied to δ. The comparison results for the setting $(n, p, a, b) = (10000, 5, 1, 1)$ are reported in Table 2. All methods achieve nominal coverage across a range of covariate designs and noise distributions. Notably, under heavy-tailed noise, the privatized CIs are substantially shorter than the private bootstrap intervals of Ferrando et al. (2022).

5.3 High-dimensional setting

In the high-dimensional setting, the covariate vectors $\mathbf{x}_{i,-1}$ are independently generated from $\mathcal{N}(\mathbf{0}, \mathbf{\Psi})$, where the covariance matrix $\mathbf{\Psi} = (\Psi_{j,k})_{j,k \in [p-1]}$ with $\Psi_{j,k} = 0.1^{|j-k|}$. We set $a = b = 1$ and $s^* = 10$. As a benchmark, we consider the non-private sparse Huber estimator, Algorithm 3 run without initialization, clipping, and noise injection. For this estimator, we take $\eta_0 = 0.2$, $T = \lceil 2 \log n \rceil$, and $s = \lceil 1.2 s^* \rceil$, and update the robustification parameter according to $\tau = 0.1 \hat{\sigma}_0 \sqrt{n (s \log p + \log n)^{-1}}$, where $\hat{\sigma}_0^2 = n^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$.

Figure 3 displays the logarithmic relative ℓ_2-error as a function of the sample size for both the (ϵ, δ)-DP and non-private sparse Huber estimators, with $p = 10000$ and $\epsilon \in \{0.5, 0.9\}$. An additional plot for $p = 5000$ is provided in the supplementary material (see Figure S3).

To assess robustness under heavy-tailed noise, we also implement the sparse DP least squares estimator (sparse DP LS) following Algorithm 4.2 of Cai et al. (2021). Because their method is designed for models with zero-mean covariates and no intercept, we restrict attention to the DP estimation of the slope coefficients. Accordingly, we center both the response variables and the covariates before applying Algorithm 4.2 of Cai et al. (2021). In addition to $(\eta_0, s, T, \boldsymbol{\beta}^{(0)})$, the algorithm requires three extra tuning parameters: the truncation level *R*, the feasibility parameter c_0, and the noise scale *B*. As specified in Theorem 4.4 of Cai et al. (2021), we set $R = C_R \sigma \sqrt{2 \log n}$ with $\sigma = 2$, $C_R \in \{0.1, 0.5, 1\}$, $c_0 = 1.01 \sqrt{s^*}$, $B = 4(R + c_0 c_{\mathbf{x}}) c_{\mathbf{x}} / \sqrt{s}$, and $c_{\mathbf{x}} = 0.5 \sqrt{\log(pn)}$. The results for $p = 10000$ and $n \in \{5000, 10000, 15000\}$ are summarized in Table 3. The proposed sparse DP Huber estimator achieves substantially higher statistical accuracy than the sparse DP LS method, while also requiring far less tuning. Moreover, under sub-Gaussian errors, it attains the same privacy level with a smaller noise scale than the sparse DP LS estimator. Additional results for $p = 5000$ are provided in the supplementary material (see Table S5).

Supplementary Materials

Appendix A presents the real-data analysis. Appendix B provides a brief review of adaptive Huber regression. Appendices C-E collect the technical proofs of all main theoretical results. Appendix F presents formal privacy guarantees for the initialization strategy, along with additional simulation studies for (ϵ, δ)-DP Huber estimators and ϵ-GDP Huber estimators.

Acknowledgments

We thank the Co-Editor, the Associate Editor, and three anonymous referees for their valuable comments and suggestions, which have greatly improved our paper.

Disclosure Statement

The authors report there are no competing interest to declare.

Funding

J. Chang and L. Yang were supported in part by the National Natural Science Foundation of China (Grant nos. 72495122 and 72125008).

Notes

[1] https://github.com/ceciliaferrando/PB-DP-CIs

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318.

Asi, H., Ullman, J., and Zakynthinou, L. (2023). From robustness to privacy and back. In *International Conference on Machine Learning*, 1121–1146.

Avella-Medina, M. (2021). Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 116, 969–983.

Avella-Medina, M., Bradshaw, C., and Loh, P.-L. (2023). Differentially private inference via noisy optimization. *The Annals of Statistics*, 51, 2067–2092.

Awan, J. and Slavković, A. (2018). Differentially private uniformly most powerful tests for binomial data. *Advances in Neural Information Processing Systems*, 31.

Barber, R. F. and Duchi, J. C. (2014). Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*.

Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. *Symposium on Foundations of Computer Science*, 464–473.

Blumensath, T. and Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27, 265–274.

Brown, G., Hayase, J., Hopkins, S., Kong, W., Liu, X., Oh, S., Perdomo, J. C., and Smith, A. (2024). Insufficient statistics perturbation: Stable estimators for private least squares extended abstract. In *Conference on Learning Theory*, 750–751.

Cai, B., Daskalakis, C., and Kamath, G. (2017). Priv'it: Private and sample efficient identity testing. In *International Conference on Machine Learning*, 635–644.

Cai, T. T., Wang, Y., and Zhang, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49, 2825–2850.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48, 1148–1185.

Chang, J., Hu, Q., Kolaczyk, E. D., Yao, Q., and Yi, F. (2024). Edge differentially private estimation in the β-model via jittering and method of moments. *The Annals of Statistics*, 52, 708–728.

Chaudhuri, K. and Hsu, D. (2012). Convergence rates for differentially private statistical estimation. *International Conference on Machine Learning*, 1327–1334.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12, 1069–1109.

Dong, J., Roth, A., and Su, W. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society Series B*, 84, 3–37.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2013). Local privacy and statistical minimax rates. *Symposium on Foundations of Computer Science*, 429–438.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113, 182–201.

Dwork, C. and Lei, J. (2009). Differential privacy and robust statistics. *Symposium on Theory of Computing*, 371–380.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, 265–284.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9, 211–407.

Dwork, C., Su, W., and Zhang, L. (2021). Differentially private false discovery rate control. *Journal of Privacy and Confidentiality*, 11, 1–44.

Fan, J., Li, Q., and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society Series B*, 79, 247–265.

Feldman, V., Mironov, I., Talwar, K., and Thakurta, A. (2018). Privacy amplification by iteration. *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*.

Ferrando, C., Wang, S., and Sheldon, D. (2022). Parametric bootstrap for differentially private confidence intervals. *International Conference on Artificial Intelligence and Statistics*, 151, 1598–1618.

Georgiev, K. and Hopkins, S. (2022). Privacy induces robustness: Information-computation gaps and sparse mean estimation. *Advances in Neural Information Processing Systems*, 35, 6829–6842.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

Hardt, M. and Talwar, K. (2010). On the geometry of differential privacy. *Symposium on Theory of Computing*, 705–714.

Hopkins, S. B., Kamath, G., Majid, M., and Narayanan, S. (2023). Robustness implies privacy in statistical estimation. *Symposium on Theory of Computing*, 497–506.

Hu, L., Ni, S., Xiao, H., and Wang, D. (2022). High dimensional differentially private stochastic optimization with heavy-tailed data. *Symposium on Principles of Database Systems*, 227–236.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35, 73–101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1, 799–821.

Jain, P., Tewari, A., and Kar, P. (2014). On iterative hard thresholding methods for high-dimensional M-estimation. *Advances in Neural Information Processing Systems*, 27, 685–693.

Kairouz, P., Oh, S., and Viswanath, P. (2017). The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63, 4037–4049.

Kamath, G., Liu, X., and Zhang, H. (2022). Improved rates for differentially private stochastic convex optimization with heavy-tailed data. *International Conference on Machine Learning*, 162, 10633–10660.

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

Karwa, V. and Vadhan, S. (2018). Finite sample differentially private confidence intervals. *Innovations in Theoretical Computer Science Conference*, 94, 44:1–44:9.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. (2008). What can we learn privately? *2008 49th Annual IEEE Symposium on Foundations of Computer Science*.

Kifer, D., Smith, A., and Thakurta, A. (2012). Private convex empirical risk minimization and high-dimensional regression. *Conference on Learning Theory*, 23, 25–40.

Lei, J. (2011). Differentially private M-estimators. *Advances in Neural Information Processing Systems*, 24.

Liu, L., Li, T., and Caramanis, C. (2019). High dimensional robust estimation of sparse models via trimmed hard thresholding. *arXiv preprint arXiv:1901.08237v1*.

Liu, W., Mao, X., Zhang, X., and Xin, Z. (2024). Efficient sparse least absolute deviation regression with differential privacy. *IEEE Transactions on Information Forensics and Security*, 19, 2328–2339.

Liu, X., Jain, P., Kong, W., Oh, S., and Suggala, A. (2023). Label robust and differentially private linear regression: Computational and statistical efficiency. *Advances in Neural Information Processing Systems*, 36, 23019–23033.

Liu, X., Kong, W., Kakade, S., and Oh, S. (2021). Robust and differentially private mean estimation. *Advances in Neural Information Processing Systems*, 34, 3887–3901.

Liu, X., Kong, W., and Oh, S. (2022). Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, 1167–1246.

Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. *Symposium on Theory of Computing*, 75–84.

She, Y., Shen, J., and Barbu, A. (2023). Slow kill for big data learning. *IEEE Transactions on Information Theory*, 69, 5936–5955.

Sheffet, O. (2017). Differentially private ordinary least squares. In *International Conference on Machine Learning*, 3105–3114.

Sheffet, O. (2019). Old techniques in differentially private linear regression. In *Algorithmic Learning Theory*, 789–827.

Smith, A. (2011). Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, 813–822.

Sun, Q., Zhou, W.-X., and Fan, J. (2020). Adaptive Huber regression. *Journal of the American Statistical Association*, 115, 254–265.

Varshney, P., Thakurta, A., and Jain, P. (2022). (Nearly) Optimal private linear regression for sub-Gaussian data via adaptive clipping. In *Conference on Learning Theory*, 1126–1166.

Vershynin, R. (2018). *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Vu, D. and Slavkovic, A. (2009). Differential privacy for clinical trial data: Preliminary evaluations. In *International Conference on Data Mining Workshops*, 138–143.

Wang, D., Xiao, H., Devadas, S., and Xu, J. (2020). On differentially private stochastic convex optimization with heavy-tailed data. *Proceedings of the 37th International Conference on Machine Learning*.

Wang, L. and Gu, Q. (2019). Differentially private iterative gradient hard thresholding for sparse learning. *International Joint Conferences on Artificial Intelligence Organization*, 3740–3747.

Wang, Y.-X. (2018). Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. *Uncertainty in Artificial Intelligence*.

Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P. (2019). Subsampled Renyi differential privacy and analytical moments accountant. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 89, 1226–1235.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105, 375–389.

Yu, M., Ren, Z., and Zhou, W.-X. (2024). Gaussian differentially private robust mean estimation and inference. *Bernoulli*, 30, 3059–3088.

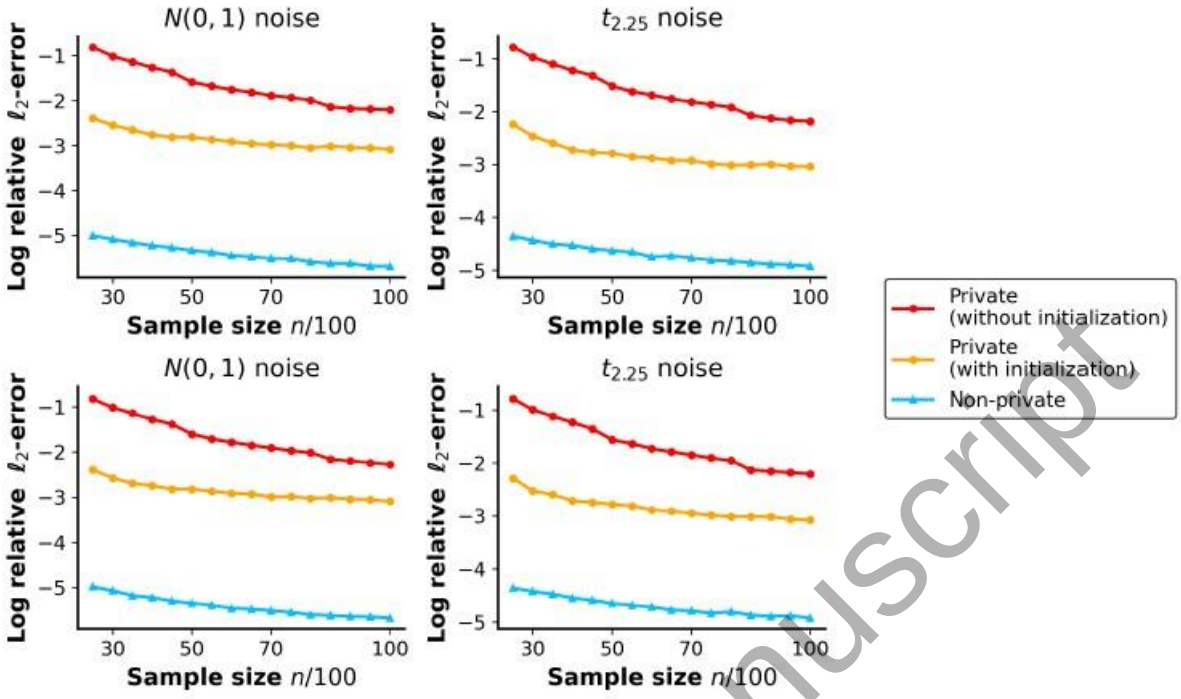

Figure 1: Plots of the average logarithmic relative ℓ_2-error over 300 repetitions as a function of the sample size under different design settings. The top two panels correspond to Gaussian covariates, and the bottom two panels to uniform covariates, with $\epsilon = 0.9$, $p = 10$, $a = 2$, and $b = 0.5$. The noise distribution is either $\mathcal{N}(0,1)$ or $t_{2.25}$.

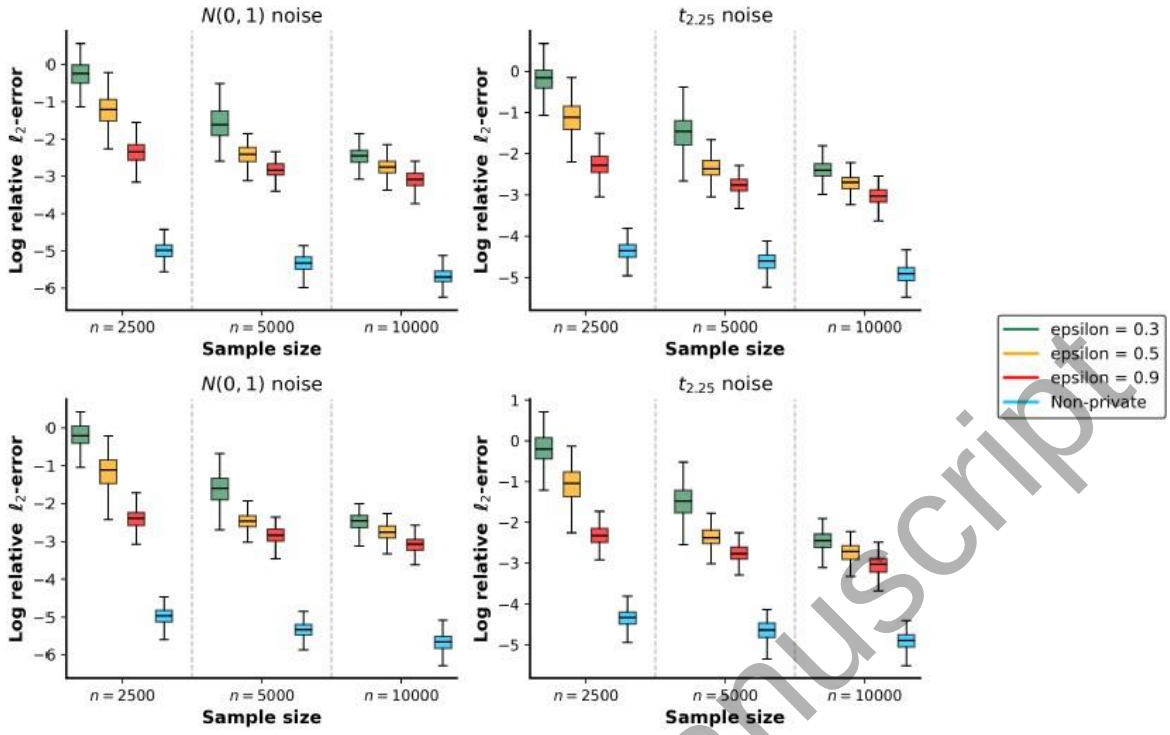

Figure 2: Boxplots of the logarithmic relative ℓ_2-error, based on 300 repetitions, across three sample sizes under four design settings. The top two panels correspond to Gaussian covariates, and the bottom two to uniform covariates, with $p = 10$, $a = 2$, and $b = 0.5$. The noise distribution is either $\mathcal{N}(0,1)$ or $t_{2.25}$.

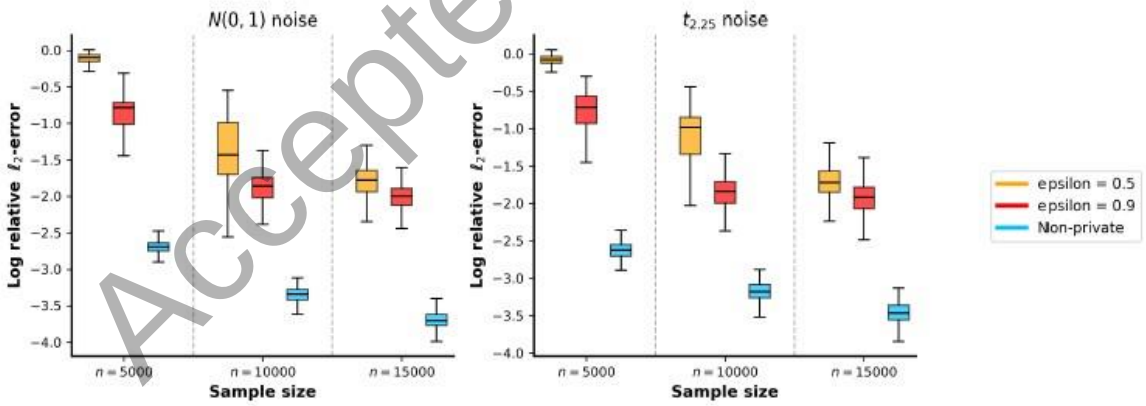

Figure 3: Boxplots of the logarithmic relative ℓ_2-error over 300 repetitions for private and non-private sparse Huber estimators across sample sizes, with $p = 10000$ and $s^* = 10$.

Table 1: Average logarithmic relative ℓ_2-error over 300 repetitions for the (ϵ, δ)-DP Huber estimators under various combinations of (a, b, n, ϵ), with $p = 10$.

			Gaussian design								Uniform design							
			$\mathcal{N}(0,1)$ noise				$t_{2.25}$ noise				$\mathcal{N}(0,1)$ noise				$t_{2.25}$ noise			
a	b	n	non-private	$\epsilon = 0.3$	$\epsilon = 0.5$	$\epsilon = 0.9$	non-private	$\epsilon = 0.3$	$\epsilon = 0.5$	$\epsilon = 0.9$	non-private	$\epsilon = 0.3$	$\epsilon = 0.5$	$\epsilon = 0.9$	non-private	$\epsilon = 0.3$	$\epsilon = 0.5$	$\epsilon = 0.9$
0.5	0.5	2500	−3.624	0.27	−0.650	1.612	3.018	0.35	−0.489	1.451	3.603	0.26	−0.611	1.693	3.026	0.30	−0.476	1.510
		5000	−3.955	0.950	1.745	2.366	3.290	0.780	1.609	2.168	3.966	0.993	1.819	2.393	3.309	0.788	1.618	2.158
		10000	−4.310	1.890	2.392	2.805	3.578	1.754	2.176	2.533	4.292	1.957	2.408	2.811	3.570	1.788	2.202	2.551
	1	2500	−3.277	0.309	−0.540	1.408	2.692	0.437	−0.348	1.234	3.256	0.313	−0.513	1.485	2.703	0.404	−0.304	1.273
		5000	−3.609	0.797	1.540	2.193	2.964	0.598	1.376	1.951	3.619	0.848	1.607	2.212	2.982	0.600	1.385	1.946
		10000	−3.963	1.708	2.225	2.667	3.249	1.518	1.964	2.342	3.945	1.754	2.238	2.676	3.238	1.568	1.987	2.357
	2	2500	−2.931	0.387	−0.386	1.175	2.360	0.572	−0.144	0.974	2.909	0.375	−0.373	1.247	2.372	0.566	−0.101	1.014
		5000	−3.262	0.613	1.314	1.955	2.631	0.368	1.113	1.707	3.273	0.667	1.367	1.975	2.649	0.373	1.116	1.697

		10000	−3.617	1.480	1.998	2.463	2.913	1.262	1.722	2.118	3.598	1.520	2.021	2.473	2.901	1.315	1.741	2.133
1	0.5	2500	−4.317	0.069	0.971	2.041	3.650	0.000	0.836	1.905	4.296	0.051	0.914	2.113	3.649	0.003	−0.803	1.943
		5000	−4.648	1.322	2.170	2.635	3.924	1.163	2.035	2.503	4.659	1.366	2.217	2.645	3.940	1.163	2.036	2.491
		10000	−5.003	2.272	2.611	2.934	4.223	2.130	2.478	2.813	4.985	2.307	2.616	2.934	4.218	2.185	2.504	2.825
	1	2500	−3.970	0.039	0.893	1.889	3.335	0.064	−0.734	1.734	3.949	0.019	0.845	1.967	3.341	0.070	−0.692	1.773
		5000	−4.302	1.212	2.039	2.565	3.610	1.023	1.874	2.390	4.312	1.254	2.086	2.579	3.627	1.029	1.878	2.378
		10000	−4.656	2.162	2.555	2.897	3.902	1.984	2.374	2.726	4.638	2.203	2.563	2.900	3.897	2.046	2.398	2.736
	2	2500	−3.624	0.017	−0.782	1.690	3.018	0.164	−0.596	1.529	3.603	0.036	−0.739	1.761	3.026	0.161	−0.556	1.568
		5000	−3.955	1.063	1.847	2.431	3.290	0.856	1.672	2.232	3.966	1.100	1.890	2.449	3.309	0.869	1.677	2.221
		10000	−4.310	1.997	2.441	2.819	3.578	1.803	2.228	2.601	4.292	2.030	2.453	2.825	3.570	1.863	2.250	2.612
2	0.5	2500	−5.010	0.253	1.233	2.362	4.278	0.204	1.138	2.278	4.989	0.228	1.170	2.418	4.267	0.215	1.102	2.320
		50																

		00	−5.34	11.59	12.43	02.83	14.55	61.48	52.36	82.77	95.35	21.63	52.46	52.83	64.56	11.50	02.37	42.769
		10000	−5.69	62.47	32.77	13.09	84.86	12.40	92.72	33.04	65.67	82.49	42.77	53.09	44.85	52.46	52.74	23.054
	1	2500	−4.66	40.24	51.19	52.29	63.96	40.18	51.07	92.17	24.64	20.22	11.13	62.35	73.95	70.19	41.04	42.217
		5000	−4.99	51.53	62.38	82.82	04.23	91.40	12.28	62.73	65.00	61.57	92.42	52.82	64.24	91.41	92.29	32.727
		10000	−5.34	92.44	92.76	43.09	04.54	32.35	02.68	73.01	95.33	12.47	22.76	93.08	84.53	62.40	92.70	73.027
	2	2500	−4.31	70.23	01.13	42.17	53.65	00.15	30.99	82.02	34.29	60.20	61.08	12.23	93.64	90.16	30.96	62.070
		5000	−4.64	81.44	82.29	52.78	53.92	41.29	02.15	82.66	14.65	91.48	82.33	72.79	43.94	01.31	12.16	82.653
		10000	−5.00	32.38	82.74	23.07	44.22	32.25	32.62	72.97	14.98	52.41	32.74	83.07	54.21	82.31	52.64	52.981

Table 2: Empirical coverage and interval widths (with standard deviations) of the $100(1-\alpha)\%$ CIs for $\boldsymbol{\beta}^*$, constructed using the (ϵ,δ)-DP Huber estimator, its non-private counterpart, and the private bootstrap method. Results are reported for $n=10000$, $p=5$, $\epsilon=0.5$, and $a=b=1$.

		Gaussian design				Uniform design			
		$\mathcal{N}(0,1)$ noise		$t_{2.25}$ noise .		$\mathcal{N}(0,1)$ noise		$t_{2.25}$ noise .	
α		coverage	width (sd)	coverage	width (sd)	coverage	width (sd)	coverage	width (sd)
0.05	private	0.942	0.352 (0.009)	0.943	0.430 (0.017)	0.941	0.349 (0.005)	0.938	0.421 (0.007)
	non-private	0.954	0.039 (< 0.001)	0.953	0.080 (< 0.001)	0.949	0.039 (< 0.001)	0.952	0.080 (< 0.001)
	bootstrap	0.935	0.557 (0.001)	0.949	1.827 (0.011)	0.949	0.101 (< 0.001)	0.941	0.693 (0.001)
0.1	private	0.909	0.296 (0.008)	0.916	0.361 (0.015)	0.905	0.293 (0.004)	0.912	0.354 (0.005)
	non-private	0.903	0.033 (< 0.001)	0.896	0.067 (< 0.001)	0.889	0.033 (< 0.001)	0.897	0.067 (< 0.001)
	bootstrap	0.882	0.453 (0.001)	0.891	1.434 (0.004)	0.899	0.083 (< 0.001)	0.899	0.540 (0.002)

Table 3: Average logarithmic relative ℓ_2-error (for slope coefficients) over 300 repetitions across sample sizes, with privacy levels $(0.5, 10n^{-1.1})$ and $p = 10000$.

Noise	n	non-private sparse Huber	sparse DP Huber	sparse DP LS		
				$C_R = 0.1$	$C_R = 0.5$	$C_R = 1$
$\mathcal{N}(0,1)$	5000	−2.671	−0.063	0.066	0.107	0.164
	10000	−3.327	−1.337	−0.079	0.046	0.084
	15000	−3.665	−1.799	−0.301	−0.043	0.047
$t_{2.25}$	5000	−2.609	−0.039	0.067	0.113	0.166
	10000	−3.158	−1.047	−0.078	0.039	0.091
	15000	−3.437	−1.725	−0.308	−0.038	0.047