Differentially Private Sparse Linear Regression with Heavy-tailed Responses

Xizhi Tian[1,2,3], Meng Ding[4], Touming Tao[5],
Zihang Xiang[1,2], Di Wang[†,1,2]

[1] Provable Responsible AI and Data Analytics (PRADA) Lab
[2] King Abdullah University of Science and Technology
[3] Utrecht University
[4] University at Buffalo
[5] Technical University Berlin

Abstract. As a fundamental problem in machine learning and differential privacy (DP), DP linear regression has been extensively studied. However, most existing methods focus primarily on either regular data distributions or low-dimensional cases with irregular data. To address these limitations, this paper provides a comprehensive study of DP sparse linear regression with heavy-tailed responses in high-dimensional settings. In the first part, we introduce the DP-IHT-H method, which leverages the Huber loss and private iterative hard thresholding to achieve an estimation error bound of $\tilde{O}\Big(s^{*\frac{1}{2}} \cdot \Big(\frac{\log d}{n}\Big)^{\frac{\zeta}{1+\zeta}} + s^{*\frac{1+2\zeta}{2+2\zeta}} \cdot \Big(\frac{\log^2 d}{n\varepsilon}\Big)^{\frac{\zeta}{1+\zeta}}\Big)$ under the (ε, δ)-DP model, where n is the sample size, d is the dimensionality, s^* is the sparsity of the parameter, and $\zeta \in (0, 1]$ characterizes the tail heaviness of the data. In the second part, we propose DP-IHT-L, which further improves the error bound under additional assumptions on the response and achieves $\tilde{O}\Big(\frac{(s^*)^{3/2}\log d}{n\varepsilon}\Big)$. Compared to the first result, this bound is independent of the tail parameter ζ. Finally, through experiments on synthetic and real-world datasets, we demonstrate that our methods outperform standard DP algorithms designed for "regular" data.

Keywords: Differential Privacy · Sparse Linear Regression · Heavy-tailed Data

1 Introduction

Differential Privacy (DP) [18] has received significant attention and is now widely considered the *de facto* standard to protect privacy in data analysis. DP provides a rigorous mathematical framework to ensure that the inclusion or exclusion of any single individual's data in a dataset does not significantly affect the output of an analysis, thereby preserving privacy. A large body of research has explored

[†] Corresponding Author.

various DP guarantees, and these concepts have been successfully adopted in industry [14, 43].

Linear regression in the DP model has been extensively studied for many years, becoming one of the most thoroughly explored topics in machine learning and DP communities. A substantial body of research has addressed the problem from various perspectives. Early investigations of DP linear regression were closely tied to more general frameworks such as DP Stochastic Convex Optimization (DP-SCO) and Empirical Risk Minimization (DP-ERM), as explored in the seminal works of [11, 12, 55, 49, 46, 51, 22, 58, 39, 40, 16, 59, 37]. Subsequently, numerous methods have been proposed to address DP linear regression under different settings. For instance, several studies [4, 19, 25, 53, 15] focus on low-dimensional scenarios in the central DP model, while others [9, 28, 29, 42, 56] extend these ideas to high-dimensional sparse linear regression—an increasingly relevant setting for modern, structured data. The local DP model has also attracted attention [17, 47, 50, 52, 60, 61], imposing stricter privacy requirements but potentially offering stronger protection.

Despite these advances, most prior work assumes *regular* data distributions, typically requiring that features and responses are bounded or sub-Gaussian. Classical approaches such as output perturbation [12] or objective/gradient perturbation [5] rely on these assumptions to guarantee an $O(1)$-Lipschitz loss for all data points. In practice, however, these conditions often fail—particularly in domains like biomedicine and finance, where heavy-tailed distributions commonly arise [6, 24, 57]. In such cases, Lipschitz conditions may be violated; for example, in linear regression with squared loss $\ell(w, (x, y)) = (w^\top x - y)^2$, heavy-tailed features can lead to unbounded gradients, invalidating assumptions used in many DP proofs. Although gradient truncation or trimming [1] has been suggested as a remedy, a thorough analysis of its convergence properties under DP constraints has been lacking. This gap underscores the need for private and robust methods specifically tailored to heavy-tailed data.

Recent works have begun addressing DP linear regression in the presence of heavy-tailed data [3, 27, 31, 48, 54]. However, some of these approaches suffer from a polynomial dependence on the data dimension d, making them unsuitable for high-dimensional settings where $d \gg n$. Thus, a natural question is: What are the theoretical behaviors of DP linear regression in high-dimensional sparse cases with heavy-tailed data?

Our Contributions. In this paper, we study the setting of high-dimensional sparse linear regression with heavy-tailed response under DP constraints. Specifically, we consider the scenario where the feature vector is sub-Gaussian and the response only has a finite $(1 + \zeta)$-th moment with $\zeta \in (0, 1]$. We propose novel DP linear regression algorithms that are robust to heavy-tailed data and capable of handling high-dimensional sparse problems effectively. Specifically:

1. For general heavy-tailed responses, we propose the DP-IHT-H algorithm, addressing the unexplored challenge of adapting Huber loss-based linear regression to differential privacy. Specifically, it achieves an error bound under

(ϵ, δ)-DP given by

$$\tilde{O}\left(s^{*\frac{1}{2}} \cdot \left(\frac{\log d}{n}\right)^{\frac{\zeta}{1+\zeta}} + s^{*\frac{1+2\zeta}{2+2\zeta}} \cdot \left(\frac{\log^2 d}{n\varepsilon}\right)^{\frac{\zeta}{1+\zeta}}\right),$$

where n is the sample size, d is the data dimensionality, and s^* represents the sparsity of the parameter.

2. To further improve the error bound, we propose the DP-IHT-L algorithm, which leverages the ℓ_1 loss function instead. By adding some mild assumptions on the response noise, the DP-IHT-L algorithm achieves a lower gradient bound and reduces the magnitude of the added noise. This enhancement leads to improved error bounds and greater stability, regardless of the value of ζ. Specifically, under (ϵ, δ)-DP, the error is bounded by $\tilde{O}\left(\frac{(s^*)^{3/2}\log d}{n\epsilon}\right)$, which is independent of the moment and matches best-known results of the sub-Gaussian case [21, 10].
3. Through extensive experiments on both synthetic and real-world datasets, we demonstrate that our methods outperform DP algorithms designed for regular data. Moreover, in some cases, DP-IHT-L further improves upon DP-IHT-H.

2 Related Work

As we mentioned, DP linear regression has been extensively studied. Still, most existing methods assume that the underlying data distribution is sub-Gaussian or bounded, rendering them unsuitable for heavy-tailed data. In contrast, in the non-private setting, recent advances have addressed Stochastic Convex Optimization (SCO) and Empirical Risk Minimization (ERM) under heavy-tailed data [7, 20, 30, 32, 41, 38]. However, these non-private methods are not directly adaptable to private settings, particularly in our high dimensional sparse scenarios.

The first study addressing DP-SCO with heavy-tailed data was proposed by [48], which introduced three methods based on distinct assumptions. The first method utilizes the Sample-and-Aggregate framework [34], but its stringent assumptions lead to a relatively large error bound. The second method leverages smooth sensitivity [8] but requires the data distribution to be sub-exponential. Additionally, [48] proposed a private estimator inspired by robust statistics, which shares similarities with our approach. Building on the mean estimator proposed in [27], [26, 44] recently investigated DP-SCO and achieved improved (expected) excess population risks of $\tilde{O}\left(\left(\frac{d}{\epsilon n}\right)^{\frac{1}{2}}\right)$ and $\tilde{O}\left(\frac{d}{\epsilon n}\right)$ for convex and strongly convex loss functions, respectively. These results rely on the assumption that the gradient of the loss function has a bounded second-order moment, aligning with the best-known outcomes for heavy-tailed mean estimation. However, these approaches only consider low dimensional case and cannot address high-dimensional or sparse learning problems. Furthermore, their methods are not directly extendable to our models, where the assumption of bounded

second-order moments of the loss function gradient is overly restrictive than our assumption on the finite $(1+\zeta)$-moment for the response.

Recently, [21] proposed a method that requires only that the distributions of x sub-Gaussian and y at least have bounded second-order moments, achieving an error bound of $\tilde{O}\left(\frac{(s^*)^3 \log^2 d}{n\epsilon}\right)$. In our paper, we consider the weaker assumption that y only has the bounded $(1+\zeta)$-th moment. We show that, under some additional assumptions, it is possible to achieve almost the same error as in [21].

3 Preliminaries

Definition 1 (Differential Privacy [18]). *A randomized mechanism M for the data universe $\mathcal{D}$ satisfies (ε, δ)-differential privacy if, for all measurable subsets $S \subseteq \mathrm{Range}(M)$ and for all pairs of adjacent datasets $D, D' \in \mathcal{D}$ (differing by at most one data point), $\Pr\big[M(D) \in S\big] \leq e^{\varepsilon} \Pr\big[M(D') \in S\big] + \delta$, where the probability space is over the randomness of the mechanism M.*

Definition 2 (Laplacian Mechanism). *Given a function $q : X^n \to \mathbb{R}^d$, the Laplacian Mechanism is defined as: $\mathcal{M}_L(D, q, \epsilon) = q(D) + (Y_1, Y_2, \cdots, Y_d)$, where Y_i is i.i.d. drawn from a Laplacian Distribution $\mathrm{Lap}\left(\frac{\Delta_1(q)}{\epsilon}\right)$, and $\Delta_1(q)$ is the ℓ_1-sensitivity of the function q, i.e., $\Delta_1(q) = \sup_{D \sim D'} \|q(D) - q(D')\|_1$. For a parameter λ, the Laplacian distribution has the density function $\mathrm{Lap}(\lambda)(x) = \frac{1}{2\lambda} \exp\left(-\frac{|x|}{\lambda}\right)$. The Laplacian Mechanism preserves ϵ-DP.*

Definition 3 (Sub-Gaussian Vector). *A random vector $\mathbf{X} \in \mathbb{R}^d$ is sub-Gaussian if every one-dimensional projection $\langle \mathbf{X}, \mathbf{v} \rangle$, for any unit vector $\mathbf{v} \in \mathbb{R}^d$, is a sub-Gaussian random variable. That is, there exists $\sigma^2 > 0$ such that for all $\lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda(\langle \mathbf{X}, \mathbf{v} \rangle - \mathbb{E}[\langle \mathbf{X}, \mathbf{v} \rangle])}\right] \leq e^{\frac{\lambda^2 \sigma^2}{2}}$.*

We consider a sparse linear model $y_i = x_i^\top \boldsymbol{\beta}^* + \varepsilon_i$, where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is the underlying unknown parameter vector, and ε_i represents the noise term. In the high dimensional setting, we assume $\boldsymbol{\beta}^*$ is s^*-sparse, *i.e.,* $s^* = |\operatorname{supp}(\boldsymbol{\beta}^*)|$ with $s^* \ll d$. In DP linear regression, given a dataset where each sample is i.i.d. sampled from the linear model, the goal is to develop some (ϵ, δ)-DP estimator $\boldsymbol{\beta}^{priv}$ to make the estimation error $\|\boldsymbol{\beta}^{priv} - \boldsymbol{\beta}^*\|_2$ as small as possible. We adopt the following general assumptions, requiring bounded eigenvalues and a bounded $\boldsymbol{\beta}^*$, which are common in high dimensional setting [38, 41, 9] :

Assumption 1 *We assume the following:*

1. *The covariates $\{x_i\}$ are zero-mean $O(1)$-sub-Gaussian with covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[xx^\top]$. The eigenvalues of $\boldsymbol{\Sigma}$ are bounded as follows: $c_l \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq c_u$.*
2. *$\|\boldsymbol{\beta}^*\|_2 \leq c_u^{1/2} r = L$, where r is a constant.*

In this paper, we focus on linear models with heavy-tailed responses. Unlike previous work on the DP linear model, here we only assume the response (or the noise) has only bounded $1+\zeta$-th moment:

Assumption 2 *We assume the noise ε_i has zero mean, $\mathbb{E}[\varepsilon_i] = 0$, and a finite $(1+\zeta)$-th moment with some $\zeta \in (0, 1]$:*

$$v_\zeta \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big(|\varepsilon_i|^{1+\zeta}\big) \;<\; \infty.$$

4 DP Iterative Hard Thresholding with Huber Loss

In the non-private case, recently [45, 41] show that an estimator based on the Huber loss [23] can achieve the optional estimation rate:

$$\mathcal{L}_\tau(\boldsymbol{\beta}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\ell_\tau\big(y_i - x_i^\top\boldsymbol{\beta}\big), \mathrm{s.t.}\|\boldsymbol{\beta}\|_0 \le s, \|\boldsymbol{\beta}\|_2 \le L, \tag{1}$$

where s is a parameter and the Huber loss with parameter τ is defined as

$$\ell_\tau(x) \;:=\; \begin{cases} \dfrac{x^2}{2}, & \text{if } |x| < \tau, \\ \tau\,|x| \;-\; \dfrac{\tau^2}{2}, & \text{otherwise.} \end{cases}$$

Compared to the squared loss in the classical linear model, Huber loss is more robust to outliers and heavy-tailed noise, making it highly effective in non-DP scenarios. However, with the DP constraint, it is difficult to privatize such an estimator due to the following key challenges: 1). The optimization problem (non-convex and non-smooth) associated with Huber loss (1) lacks an efficient algorithm for the solution. For instance, [41] proposes an adaptive Huber loss estimator but does not provide an efficient algorithm to solve it. 2). If we directly use the previous DP methods to the Huber loss such as objective perturbation [11] or gradient perturbation methods [5], then we need to introduce a noise with scale $\Omega(d)$, which is extremely large as we consider the high dimensional setting where $d \gg n$.

To address these challenges, we propose the DP-IHT-H algorithm. This algorithm: 1). Efficiently leverages the Huber loss to perform a linear regression. 2). Achieves (ϵ, δ)-DP guarantees with an error bound that only logarithmically depends on the dimension d. In our DP-IHT-H algorithm, we first shrink the original feature vector x to make it have bounded ℓ_∞-norm. Next, we calculate the gradient on the shrunken data using the Huber loss and update our vector via gradient descent. Due to the bounded gradient of Huber loss, the error bound can be controlled with high probability, even for heavy-tailed data. Finally, we perform a "Peeling" step [10] to select the top s indices with the largest magnitudes in the vector while preserving DP.

Algorithm 1 DP IHT with Huber Loss (DP-IHT-H)

Input: n-size dataset $\mathbf{D} = \{(y_i, x_i)\}_{i\in[n]}$, Step size η, Sparsity level s, Privacy parameters ε, δ, Huber loss parameter τ, Truncation parameter K, Total steps T.
Output: $\boldsymbol{\beta}^T$

1: Initialize $\boldsymbol{\beta}^0 = 0$.
2: **Clipping:** For each $i \in [n]$, define the truncated sample $\tilde{x}_i \in \mathbb{R}^d$ by

$$\tilde{x}_{i,j} = \operatorname{sign}(x_{i,j}) \min\Big\{|x_{i,j}|, K\Big\}, \quad \forall j \in [d].$$

3: Denote the truncated dataset as $\tilde{D} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n$.
4: Split the data $\tilde{D}$ into T parts $\{\tilde{D}_t\}_{t=1}^T$, each with $m = \frac{n}{T}$ samples.
5: **for** $t = 0$ to $T-1$ **do**
6: $\quad \boldsymbol{\beta}^{t+0.5} = \boldsymbol{\beta}^t - \dfrac{\eta}{m} \displaystyle\sum_{i\in\tilde{D}_t} \ell'_\tau\left(y_i - \tilde{x}_i^\top \boldsymbol{\beta}^t\right) \tilde{x}_i$
7: $\quad \boldsymbol{\beta}^{t+1} = \Pi_L\Big(\text{Peeling}\big(\boldsymbol{\beta}^{t+0.5}, \tilde{D}_t, s, \varepsilon, \delta, \frac{\eta\tau K}{m}\big)\Big)$, where Π_L is the projection onto the L-radius ball.
8: **end for**
9: **return** $\boldsymbol{\beta}^T$

The Peeling algorithm achieves privacy protection by adding Laplace noise twice. First, noise is added to each entry of the vector to perturb the original data, ensuring that no single component can be directly exposed during the selection process. Based on the noisy values, the algorithm iteratively selects the indices corresponding to the s largest magnitudes. After selecting the s indices, additional Laplace noise is added to further obscure the true values of the selected components. The final output is a sparse vector where only the selected components retain their perturbed values, while all other components are set to zero. Compared to using Gaussian noise, this approach results in a smaller noise scale, effectively protecting privacy while maintaining higher accuracy and output quality. In the following, we will provide the privacy and utility guarantees of DP-IHT-H.

Theorem 1. *For any $0 < \epsilon, 0 < \delta < 1$, DP-IHT-H is (ε, δ)-DP.*

Theorem 2. *If Assumption 1 and 2 hold and assume n is sufficiently large such that $n \geq \tilde{\Omega}\big((s^*)^{\frac{2}{3}} \log d\big)$. Setting $s = O\big(s^*\big)$, stepsize $\eta = O(1)$, $K = O\big(\log d\big)$, $T = O\big(\log n\big)$, and $\tau = O\Big(\Big(\frac{n}{T\,(s^*)^{\frac{3}{2}}\,\log^{\frac{1}{2}}\big(\frac{1}{\sigma}\big)\,\log d}\Big)^{\frac{\zeta}{1+\zeta}}\Big)$ in Algorithm 1, then with probability at least $1 - O\Big(d^{-1} + Te^{-c_1 n}\Big)$ for a constant c_1, we obtain the following bound on the final estimate:*

$$\|\boldsymbol{\beta}^T - \boldsymbol{\beta}^*\|_2 = O\left((s^*)^{\frac{1}{2}} \left(\frac{\log d}{n}\right)^{\frac{\zeta}{1+\zeta}} + (s^*)^{\frac{1}{2}} \left(\frac{(s^*)^{\frac{1}{2}}\,\log^2 d\,\log(1/\delta)}{n\,\varepsilon}\right)^{\frac{\zeta}{1+\zeta}}\right).$$

Algorithm 2 Peeling Procedure

Input: Vector $\boldsymbol{x} \in \mathbb{R}^d$ (based on $\mathbf{D}$), Sparsity s, Privacy parameters ε, δ, Noise scale λ.

Output: $\boldsymbol{x}\big|_S + \tilde{\boldsymbol{w}}_S$

1: Initialize $S = \emptyset$.
2: **for** $i = 1$ to s **do**
3: Generate noise $\boldsymbol{w}_i \in \mathbb{R}^d$ where each component is drawn from: $w_{i,j} \sim \mathrm{Lap}\Big(\frac{2\lambda\sqrt{3s\log(1/\delta)}}{\varepsilon}\Big)$.
4: Append $j^* = \arg\max_{j\in[d]\setminus S}\Big(|\boldsymbol{x}_j| + w_{i,j}\Big)$ to S.
5: **end for**
6: Generate $\tilde{\boldsymbol{w}} \in \mathbb{R}^d$ where each component is drawn from: $\tilde{w}_j \sim \mathrm{Lap}\Big(\frac{2\lambda\sqrt{3s\log(1/\delta)}}{\varepsilon}\Big)$.
7: **return** $\boldsymbol{x}\big|_S + \tilde{\boldsymbol{w}}_S$.

Remark. There are two terms in the above error bound. The first one corresponds to the optimal error bound in the non-private case [41]. The second term corresponds to the error due to the noises added to ensure DP. In the case $\epsilon = O(1)$ the overall error bound is dominated by the second term. Moreover, when $\zeta = 1$, i.e., the noise has bounded variance, the error rate becomes to $\tilde{O}\big(\frac{(s^*)^{3/4}\log d}{\sqrt{n\,\varepsilon}}\big)$.

Under similar conditions, that is, assume that x is sub-Gaussian and y has a finite 2ζ -th moment with $\zeta \geq 1$ -[21] establishes an upper bound for the privacy component, given by $\tilde{O}\left((s^*)^{\frac{1+2\zeta}{1+\zeta}} \cdot \left(\frac{\log^3 n\,\log^2 d}{n^2\epsilon^2}\right)^{\frac{\zeta}{1+\zeta}}\right)$. Thus, our results can be considered as an extension to the $(1+\zeta)$-th moment case. For sub-Gaussian x and y a related bound is also provided in [10]: $\tilde{O}\left(\sqrt{\frac{s^*\log d}{n}} + \frac{s^*\log d\sqrt{\log^3 n}}{n\varepsilon}\right)$. However, this result is not directly comparable to ours due to the strong assumptions imposed on the covariance matrix.

5 DP Iterative Hard Thresholding with ℓ_1 Loss

In DP-IHT-H, the algorithm achieves its lowest error bound when $\zeta = 1$. This naturally raises the question: Can we refine the algorithm so that its performance becomes independent of ζ even when $\zeta < 1$? In this section, we propose a new method, DP-IHT-L, to address the shortcomings of DP-IHT-H. The primary issue with DP-IHT-H lies in the dependence of the bounded gradient of the Huber loss on ζ, which results in larger noise being introduced during the "Peeling" step as the best value depends on n and ζ (see Theorem 2). Thus, it is necessary to use other loss functions that do not have such a parameter τ and are with a constant gradient bound. Motivated by [38], in the following we will show the ℓ_1 loss (absolute loss function) satisfies this requirement. See Algorithm 3 for details.

The basic idea of DP-IHT-L is similar to that of DP-IHT-H. First, we clip x to ensure it has an $\tilde{O}(1)$ bounded ℓ_∞-norm. Then we update $\boldsymbol{\beta}^t$, but instead of using the gradient of the Huber loss, we replace it with the gradient of the ℓ_1 loss, thereby avoiding the dependence on the bound related to ζ. Finally, we apply the "Peeling" algorithm to introduce noise and preserve differential privacy.

Theorem 3. *For $0 < \epsilon$ and $0 < \delta < 1$, the DP-IHT-L algorithm is (ε, δ)-DP.*

To get our bound, we pose additional assumptions on the noise.

Assumption 3 *We assume the following: Let the noise terms $\{\varepsilon_i\}_{i=1}^n$ be i.i.d. with density $h_\varepsilon(\cdot)$ and distribution function $H_\varepsilon(\cdot)$, and define $\gamma = \mathbb{E}\big[|\varepsilon_i|\big]$. There exist constants $b_0, b_1 > 0$ (possibly depending on γ) such that*

$$h_\varepsilon(x) \;\geq\; \tfrac{1}{b_0}, \quad \textit{for all } |x| \leq 8\left(\tfrac{c_u}{c_l}\right)^{\frac{1}{2}}\gamma,$$

$$h_\varepsilon(x) \;\leq\; \tfrac{1}{b_1}, \quad \textit{for all } x \in \mathbb{R}.$$

Remark: The lower bound in the density is essentially a (local) Bernstein condition [2] and is easily satisfied by many heavy-tailed distributions (e.g., any t-distribution with degrees of freedom $v > 2$). The upper bound simply requires that the noise distribution does not have unbounded peaks, which is also satisfied by common distributions such as Gaussian, Laplace, and t-distributions with $v > 2$. As a result, Assumption 3 is quite relaxed for a wide range of heavy-tailed settings.

Under Assumption 3, in t-th iteration, the sub-gradient of the loss function

$$\begin{aligned}\mathbf{G}_t &= \sum_{i\in\tilde{D}_t} \operatorname{sign}\Big(x_i^\top \boldsymbol{\beta}^t - y_i\Big)\tilde{x}_i \\ &= \sum_{i\in\tilde{D}_t} \operatorname{sign}\big(x_i^\top(\boldsymbol{\beta}^t - \boldsymbol{\beta}^*)\big)\tilde{x}_i - \sum_{i\in\tilde{D}_t} \operatorname{sign}\big(x_i^\top \varepsilon_i\big)\tilde{x}_i\end{aligned}$$

exhibits two distinct regimes based on the magnitude of $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2$. These regimes determine the behavior and convergence rate of the algorithm:

- **Large Deviation Regime:** When $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2$ is relatively large, the sub-gradient $\mathbf{G}_t$ is dominated by the term $\sum_{i\in\tilde{D}_t} \operatorname{sign}\big(x_i^\top(\boldsymbol{\beta}^t - \boldsymbol{\beta}^*)\big)\tilde{x}_i$ leading to a larger error bound of $\mathbf{G}_t$. Hence the updates in this phase are large, allowing the algorithm to converge rapidly during the early iterations. This ensures efficient progress toward reducing the parameter error.
- **Small Deviation Regime:** Once $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2$ becomes small, the sub-gradient $\mathbf{G}_t$ is primarily influenced by the noise term $\sum_{i\in\tilde{D}_t} \operatorname{sign}\big(x_i^\top \varepsilon_i\big)\tilde{x}_i$. In this case the error bound of $\mathbf{G}_t$ is small leading to a slower convergence rate that ensures refinement near the underlying parameter $\boldsymbol{\beta}^*$.

Based on these differing bounds, we can establish the following convergence result.

Algorithm 3 Differentially Private Iterative Hard Thresholding with General Loss (DP-IHT-L)

Input: n-size dataset $\mathbf{D} = \{(y_i, x_i)\}_{i=1}^n$, Step size η_t, Sparsity level s, Privacy parameters ε, δ, Truncation parameter K, Total steps T.
Output: $\boldsymbol{\beta}^T$

1: **Initialization:** Set $\boldsymbol{\beta}^0 = 0$.
2: **Clipping:** For each $i \in [n]$, define the truncated sample $\tilde{x}_i \in \mathbb{R}^d$ by

$$\tilde{x}_{i,j} = \operatorname{sign}(x_{i,j}) \min\Big\{|x_{i,j}|, K\Big\}, \quad \forall j \in [d].$$

3: Denote the truncated dataset as $\tilde{D} = \{(\tilde{x}_i, y_i)\}_{i=1}^n$.
4: **Data Splitting:** Split $\tilde{D}$ into T disjoint subsets $\{\tilde{D}_t\}_{t=0}^{T-1}$, each containing $m = \frac{n}{T}$ samples.
5: **for** $t = 0$ to $T-1$ **do**
6: $\quad \boldsymbol{\beta}^{t+0.5} = \boldsymbol{\beta}^t - \dfrac{\eta_t}{m} \displaystyle\sum_{i \in \tilde{D}_t} \operatorname{sign}\Big(x_i^\top \boldsymbol{\beta}^t - y_i\Big)\tilde{x}_i$
7: $\quad \boldsymbol{\beta}^{t+1} = \Pi_L\Big(\operatorname{Peeling}\Big(\boldsymbol{\beta}^{t+0.5}, \tilde{D}_t, s, \varepsilon, \delta, \frac{2\eta K}{m}\Big)\Big)$
8: **end for**
9: **return** $\boldsymbol{\beta}^T$

Theorem 4. *Under Assumption 1, 2 and 3 and assume* $n \geq O\big(c_u\, c_l^{-1} s^* \log d\big)$. *Set* $s = \Omega\big((c_u/c_l)^8 (b_0/b_1)^8 s^*\big)$, $K = O(\log d)$, *and the initial step size* η_0 *satisfying*

$$\left[\frac{c_l^{1/2}\,\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2}{8\,n\,c_u},\ \frac{3c_l^{1/2}\,\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2}{8\,n\,c_u}\right],$$

then for the sequence $\{\boldsymbol{\beta}^t\}$ *generated by Algorithm 3, there are two distinct convergence phases:*

Phase One: *When* $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2 \geq 8\, c_l^{-1/2}\,\gamma$, *using* $\eta_t = (1-c_1)^t \eta_0$ *with* $c_1 = O(c_l\, c_u^{-1})$ *ensures*

$$\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\|_2 \ \leq\ (1-c_1)^{t+1}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \ +\ O\big(\sqrt{\boldsymbol{W}}\big).$$

Phase Two: *Once* $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2 \leq 8\, c_l^{-1/2}\,\gamma$, *switching to a constant step size* $\eta_t = O\big(c_l^{1/2} b_1^2 (n b_0 c_u)^{-1}\big)$ *yields*

$$\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\|_2 \ \leq\ (1 - c_2^*)\,\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2 \ +\ O\big(\sqrt{\boldsymbol{W}}\big).$$

Here,

$$O\big(\sqrt{\boldsymbol{W}}\big) \ =\ O\Big(\frac{T\big(s^*\big)^{3/2}\,\log d\,\big(\log(1/\delta)\big)^{1/2}\,\log(T/n)}{n\,\varepsilon}\Big)$$

represents the cumulative effect of the Laplace noise, and $c_2^* \in (0,1)$ *indicates a strict contraction rate once* $\boldsymbol{\beta}^t$ *is sufficiently close to the true parameter vector* $\boldsymbol{\beta}^*$.

Because the sub-gradient operates under two different regimes, the overall estimation error follows a two-phase pattern. In Phase One, the sub-gradient is governed by the smoothness property, leading to rapid convergence from larger errors. In Phase Two, the strong convexity takes over once the estimator is sufficiently close to $\boldsymbol{\beta}^*$, and the convergence becomes linear in nature. The $O\big(\sqrt{\boldsymbol{W}}\big)$ term arises from comparing against the exact parameter $\boldsymbol{\beta}^*$, which is not affected by the additional noise injected via the peeling procedure. With these two phases established, we can derive a global error bound as follows.

Theorem 5. *Consider the same settings as in Theorem 4, with probability at least* $1 - \exp\Big(-C\, s^* \, \log\Big(\frac{2\,d}{s^*}\Big)\Big)$, *after at most*

$$T = O\Big(\log\Big(\tfrac{\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2}{\gamma}\Big) \; + \; \log\Big(n\,\gamma\, b_0^{-1}\, \log\big(\tfrac{2\,d}{s^*}\big)\Big)\Big)$$

iterations, Algorithm 3 produces an estimator $\boldsymbol{\beta}^T$ *satisfying*

$$\|\boldsymbol{\beta}^T - \boldsymbol{\beta}^*\|_2 \;\le\; O\Big(\frac{(s^*)^{3/2}\,\log d\,\big(\log(\frac{1}{\delta})\big)^{1/2}\,\log n}{n\,\varepsilon}\Big).$$

Overall, our DP-IHT-L algorithm achieves an error bound of $\tilde{O}\Big(\frac{(s^*)^{3/2}\log d}{n\varepsilon}\Big)$ for high-dimensional sparse data with heavy-tailed distributions. In comparison, the DP-IHT-H algorithm attains an error of $\tilde{O}\Big(s^{*\frac{1+2\zeta}{2+2\zeta}}\Big(\frac{\log^2 d}{n\varepsilon}\Big)^{\frac{\zeta}{1+\zeta}}\Big)$, which is larger than our previous bound. Similarly, when $\zeta = 1$, our error bound is almost the same as in [21], which is given by $\tilde{O}\Big(s^{*\frac{1+2\zeta}{1+\zeta}}\Big(\frac{\log^3 n \log^2 d}{n^2\varepsilon^2}\Big)^{\frac{\zeta}{1+\zeta}}\Big)$. Note that in the case of $\zeta = 1$, [10] achieves an error bound of $\tilde{O}\Big(\sqrt{\frac{s^*\log d}{n}} + \frac{s^*\log d}{n\varepsilon}\Big)$. However, their analysis requires the strong condition that $\|x_I\|_\infty \le O\Big(\frac{1}{\sqrt{|I|}}\Big)$ for any index set $I \subseteq [d]$. Thus, their results is imcomparable with ours.

6 Experiments

In this section, we evaluate the practical performance of our proposed algorithms on both synthetic and real-world datasets.

6.1 Experimental Setup

Synthetic Data. We generate synthetic data for linear regression following the model

$$y_i = \langle x_i, \boldsymbol{\beta}^* \rangle + \varepsilon_i, \quad i = 1, \ldots, n,$$

where each feature vector $x_i \in \mathbb{R}^d$ is drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The true coefficient vector $\boldsymbol{\beta}^*$ has s^* nonzero entries, which are sampled from a scaled standard normal distribution, with the remaining entries set to zero. To simulate heavy-tailed noise, the error terms $\{\varepsilon_i\}$ are drawn from a Student-t distribution. Specifically, we set the degrees of freedom ν to 1.75 when $\zeta = 0.5$ and to 3 when $\zeta = 1$.

Real-World Data. For additional validation, we evaluate our algorithms on real-world datasets, including: **NCI-60 cancer cell line dataset** [36], with $n = 59$ samples and $d = 14{,}342$ features. And **METABRIC** (Molecular Taxonomy of Breast Cancer International Consortium) dataset [13, 35], with $n = 1{,}904$ samples and $d = 24{,}368$ features. These datasets, commonly found in publicly available biological databases, are known to exhibit heavy-tailed distributions.

Parameter Choices. Unless otherwise specified, the default parameter settings for the DP-IHT-H algorithm are as follows: $s^* = 5$, $\epsilon = 0.5$, $\zeta = 1$, $\delta = 1/n^{1.1}$, and the Huber loss parameter $\tau = 1$. For DP-IHT-L, the same parameters are applied. The truncation parameter K is set to $\log d$. Across all algorithms, we use a constant step size of $\eta = 0.01$.

Evaluation Metrics. We evaluate algorithm performance using the ℓ_2-estimation error, defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. As there is no underlying parameter $\boldsymbol{\beta}^*$ for real-world data, we will use the adaHuher algorithm in [41], which achieves the optimal rate, to approximate $\boldsymbol{\beta}^*$. Each experiment is repeated 20 times, and we report the average results to ensure statistical reliability.

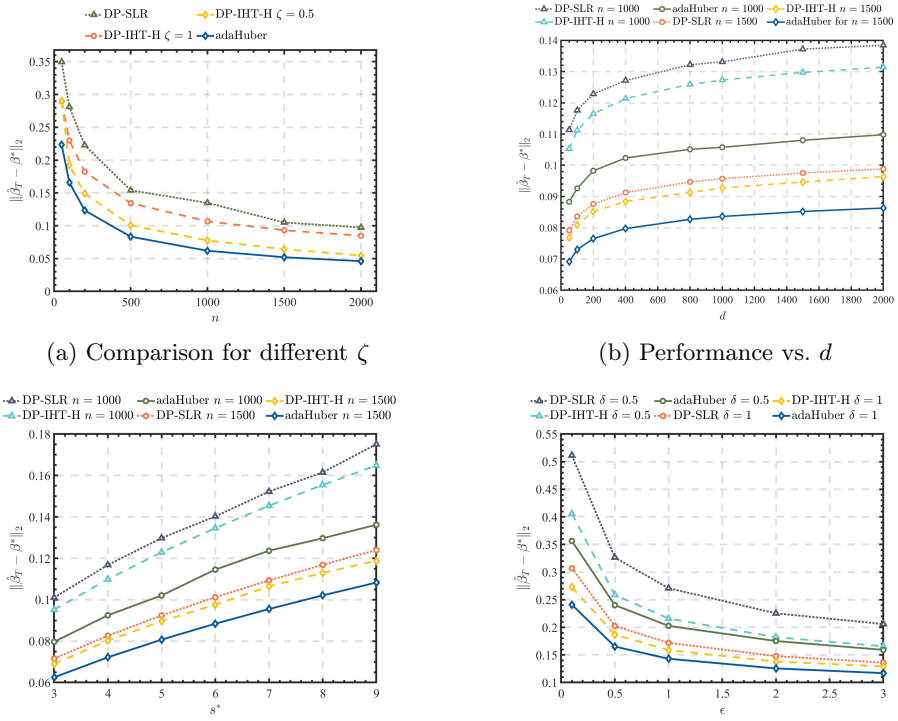
Baselines. We focus on the DP-IHT-H and DP-IHT-L algorithms. As we mentioned above, there is no previous research on the DP sparse model with heavy-tailed response that only has the $1+\zeta$-th moment with $\zeta \in (0, 1)$. For comparison, we include:

- **DP-SLR**: Differentially private sparse linear regression under regular (light-tailed) data [10]. DP-SLR could achieve the almost optimal rate in this setting.
- **adaHuber**: Non-DP linear regression in high-dimensional sparse heavy-tailed settings using the adaptive Huber loss method [41]. adaHuber achieves the optimal rate in the non-private setting.

6.2 Results on Synthetic Data

We first investigate whether DP-IHT-H achieves better performance under heavy tails (i.e., small ζ), whether it outperforms differentially private algorithms designed for regular data distributions, and the performance gap between DP-IHT-H and the non-private optimal algorithm.

Figure 1a shows that when the response variable is heavy-tailed, DP-IHT-H achieves significantly lower errors than DP-SLR, possibly because of the square loss used in DP-SLR is less robust to outliers. Moreover, Figure 1b illustrates that as the dimensionality d increases, the estimation error grows more gradually for DP-IHT-H than for the other two methods. In addition, the figure clearly indicates that DP-IHT-H outperforms the DP-SLR algorithm across various values of d. Figures 1c and 1d further confirm that in heavy-tailed scenarios, DP-IHT-H consistently outperforms the DP-SLR algorithm. Across varying sparsity s^* and privacy budgets ϵ, DP-IHT-H maintains robust performance, demonstrating its advantage in handling heavy-tailed data while preserving differential privacy.



(a) Comparison for different ζ

(b) Performance vs. d

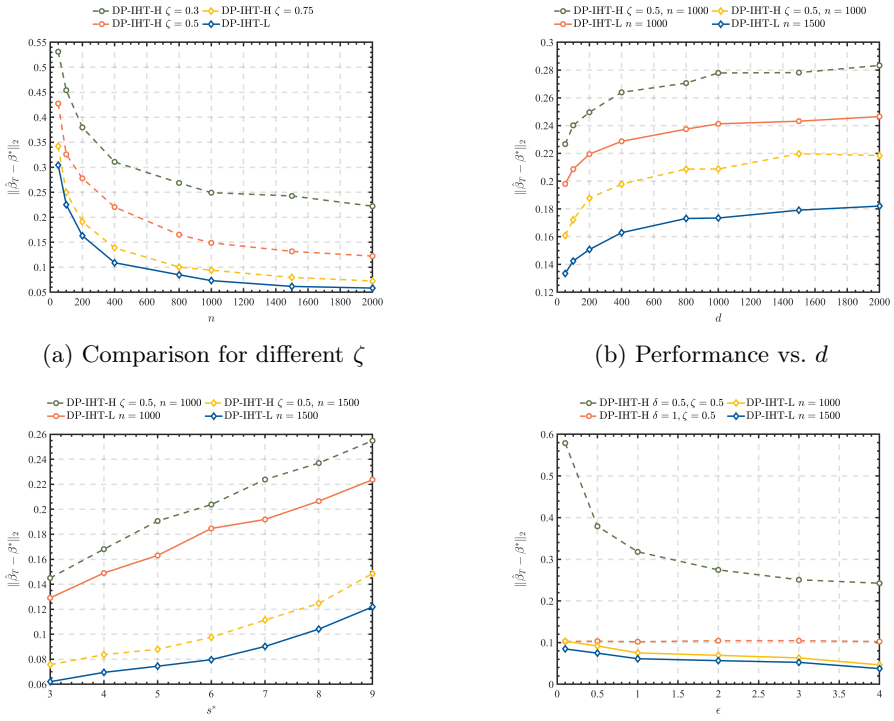
(c) Performance vs. s^*

(d) Performance vs. ϵ

Fig. 1: Comparison of DP-IHT-H, DP-SLR, and adaHuber methods.

We next evaluate the DP-IHT-L algorithm in comparison with DP-IHT-H, focusing on whether DP-IHT-L provides a similar error guarantee for different values of ζ, and whether it outperforms DP-IHT-H for larger ζ. From Figures 2a and 2b, we observe that DP-IHT-L generally outperforms DP-IHT-H, particularly when ζ is smaller. Figure 2b also indicates that when $\zeta = 1$, the difference between the two methods is minimal. As ζ decreases, the performance of DP-IHT-L remains relatively stable. Figure 2c shows that DP-IHT-H can perform better with respect to s^*, consistent with its dependence of $O(s^{*3/4})$ (in contrast to $O(s^{*3/2})$ for DP-IHT-L). Finally, Figure 2d demonstrates that for $\zeta = 1$, both algorithms has less estimation error as ϵ decrease.

Overall, these results indicate that DP-IHT-H and DP-IHT-L are particularly well suited for highly heavy-tailed data, while DP-IHT-L may be preferred in more moderate scenarios due to its stability.



(a) Comparison for different ζ

(b) Performance vs. d

(c) Performance vs. s^*

(d) Performance vs. ϵ

Fig. 2: Comparison of DP-IHT-L and DP-IHT-H across various metrics.

6.3 Real Data Analysis

We evaluate four methods—adaHuber, DP-SLR, DP-IHT-H, and DP-IHT-L—on two genomic datasets to assess their performance in privacy-preserving robust estimation.

NCI-60 Dataset. Following the protocols in [36, 41], we analyze a dataset of protein expression (from 162 antibodies) and RNA transcript levels across 60 cancer cell lines to identify genes affecting *KRT19* expression [33]. After preprocessing, the dataset comprises $n = 59$ samples and $d = 14\,342$ features. For the DP methods, we set $s^* = 5$, $\varepsilon = 0.5$, and $\delta = 1/n^{1.1}$. Table 1 summarizes the results.

The non-private adaHuber achieves the lowest MAE, while the DP-IHT variants perform competitively and surpass DP-SLR, highlighting the benefits of robust estimation in heavy-tailed data.

METABRIC Dataset. We further assess the methods on the METABRIC breast cancer dataset [13, 35], which contains $n = 1\,904$ samples and $d = 24\,368$ features.

Table 1: Results on the NCI-60 dataset.

Method	MAE	Size	Selected Genes
adaHuber	2.07	5	MALL, TM4SF4, ANXA3, ADRB2, NRN1
DP-SLR	2.72	5	MALL, TGFBI, S100A6, LPXN, DSP
DP-IHT-H	2.40	5	MALL, ANXA3, NRN1, CA2, EPS8L2
DP-IHT-L	2.34	5	MALL, NRN1, DSP, AUTS2, EPS8L2

Table 2: Results on the METABRIC dataset.

Method	MAE	Size	Selected Genes
adaHuber	0.92	5	PIK3CA, MUC16, SYNE1, KMT2C, GATA3
DP-SLR	1.22	5	MUC16, CDH1, MAP3K1, NCOR2, CBFB
DP-IHT-H	1.08	5	PIK3CA, MUC16, AHNAK2, MAP3K1, GATA3
DP-IHT-L	1.05	5	PIK3CA, SYNE1, NOTCH1, TG, KMT2C

Here, the parameters are set as $s^* = 5$, $\varepsilon = 1.0$, and $\delta = 1/n^{1.1}$. Table 2 reports the results.

As in the NCI-60 dataset, adaHuber attains the best MAE, and the DP-IHT methods perform comparably while outperforming DP-SLR, even in this lighter-tailed setting.

7 Conclusion

This work presented novel approaches for differentially private linear regression that address the challenges posed by heavy-tailed data distributions. We introduced two algorithms: DP-IHT-H and DP-IHT-L, each designed to handle different tail behaviors. DP-IHT-H demonstrates strong performance for moderately heavy-tailed data, achieving an optimized error bound under (ϵ, δ)-differential privacy. However, its performance degrades for heavier tails, prompting the development of DP-IHT-L, which achieves stable error bounds irrespective of the tail behavior. Extensive experiments on synthetic and real-world datasets verified the effectiveness of our methods, showing their robustness and applicability in diverse scenarios.

Acknowledgements

This work is supported in part by the funding BAS/1/1689-01-01, URF/1/4663-01-01, REI/1/5232-01-01, REI/1/5332-01-01, and URF/1/5508-01-01 from KAUST, and funding from KAUST - Center of Excellence for Generative AI, under award number 5940.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 308–318. ACM (2016)
2. Alquier, P., Cottet, V., Lecué, G.: Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. The Annals of Statistics **47**(4), 2117–2144 (2019)
3. Barber, R.F., Duchi, J.C.: Privacy and statistical risk: Formalisms and minimax bounds (2014)
4. Bassily, R., Feldman, V., Guzmán, C., Talwar, K.: Stability of stochastic gradient descent on nonsmooth convex losses. Advances in Neural Information Processing Systems **33** (2020)
5. Bassily, R., Feldman, V., Talwar, K., Thakurta, A.: Private stochastic convex optimization with optimal rates. In: Advances in Neural Information Processing Systems (2019)
6. Biswas, A., Datta, S., Fine, J.P., Segal, M.R.: Statistical advances in the biomedical science (2007)
7. Brownlees, C., Joly, E., Lugosi, G.: Empirical risk minimization for heavy-tailed losses. The Annals of Statistics **43**(6), 2507–2536 (2015)
8. Bun, M., Steinke, T.: Average-case averages: Private algorithms for smooth sensitivity and mean estimation (2019)
9. Cai, T.T., Wang, Y., Zhang, L.: The cost of privacy in generalized linear models: Algorithms and minimax lower bounds (2020)
10. Cai, T.T., Wang, Y., Zhang, L.: The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. The Annals of Statistics **49**(5), 2825–2850 (2021)
11. Chaudhuri, K., Monteleoni, C.: Privacy-preserving logistic regression. In: Advances in Neural Information Processing Systems. pp. 289–296 (2009)
12. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. Journal of Machine Learning Research **12**, 1069–1109 (2011)
13. Curtis, C., et al.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature **486**(7403), 346–352 (2012)
14. Ding, B., Kulkarni, J., Yekhanin, S.: Collecting telemetry data privately. In: Advances in Neural Information Processing Systems. pp. 3571–3580 (2017)
15. Ding, M., Lei, M., Wang, S., Zheng, T., Wang, D., Xu, J.: Nearly optimal differentially private relu regression. arXiv preprint arXiv:2503.06009 (2025)
16. Ding, M., Lei, M., Zhu, L., Wang, S., Wang, D., Xu, J.: Revisiting differentially private relu regression. Advances in Neural Information Processing Systems **37**, 55470–55506 (2024)
17. Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Minimax optimal procedures for locally private estimation. Journal of the American Statistical Association **113**(521), 182–201 (2018)
18. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis pp. 265–284 (2006)
19. Feldman, V., Koren, T., Talwar, K.: Private stochastic convex optimization: Optimal rates in linear time. In: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing. pp. 439–449 (2020)

20. Holland, M., Ikeda, K.: Better generalization with less data using robust gradient descent. In: International Conference on Machine Learning. pp. 2761–2770 (2019)
21. Hu, L., Ni, S., Xiao, H., Wang, D.: High dimensional differentially private stochastic optimization with heavy-tailed data. In: Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS '22). pp. 227–236. Association for Computing Machinery, New York, NY, USA (2022)
22. Huai, M., Wang, D., Miao, C., Xu, J., Zhang, A.: Pairwise learning with differential privacy guarantees. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 694–701 (2020)
23. Huber, P.J.: Robust estimation of a location parameter. The Annals of Mathematical Statistics **35**(1), 73–101 (1964)
24. Ibragimov, M., Ibragimov, R., Walden, J.: Heavy-Tailed Distributions and Robustness in Economics and Finance, vol. 214. Springer (2015)
25. Iyengar, R., Near, J.P., Song, D., Thakkar, O., Thakurta, A., Wang, L.: Towards practical differentially private convex optimization. In: 2019 IEEE Symposium on Security and Privacy (SP). pp. 299–316. IEEE (2019)
26. Kamath, G., Liu, X., Zhang, H.: Improved rates for differentially private stochastic convex optimization with heavy-tailed data (2021)
27. Kamath, G., Singhal, V., Ullman, J.: Private mean estimation of heavy-tailed distributions. In: Conference on Learning Theory. pp. 2204–2235. PMLR (2020)
28. Kasiviswanathan, S.P., Jin, H.: Efficient private empirical risk minimization for high-dimensional learning. In: International Conference on Machine Learning. pp. 488–497 (2016)
29. Kifer, D., Smith, A., Thakurta, A.: Private convex empirical risk minimization and high-dimensional regression. In: Conference on Learning Theory. pp. 25–1 (2012)
30. Lecue, G., Lerasle, M., Mathieu, T.: Robust classification via mom minimization (2018)
31. Liu, X., Kong, W., Kakade, S., Oh, S.: Robust and differentially private mean estimation (2021)
32. Lugosi, G., Mendelson, S.: Risk minimization by median-of-means tournaments. Journal of the European Mathematical Society (2019)
33. Nakata, B., Takashima, T., Ogawa, Y., Ishikawa, T., Hirakawa, K.: Serum cyfra 21-1 (cytokeratin-19 fragments) is a useful tumour marker for detecting disease relapse and assessing treatment efficacy in breast cancer. British Journal of Cancer **91**(5), 873–878 (2004)
34. Nissim, K., Raskhodnikova, S., Smith, A.: Smooth sensitivity and sampling in private data analysis. In: Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing. pp. 75–84. ACM (2007)
35. Pereira, B., et al.: The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nature Communications **7**, 11479 (2016)
36. Reinhold, W., Varma, S., Sousa, F., et al.: Cellminer: A web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the nci-60 cell line set. Cancer Research **72**(14), 3499–3511 (2012)
37. Shen, H., Wang, C.L., Xiang, Z., Ying, Y., Wang, D.: Differentially private non-convex learning for multi-layer neural networks. arXiv preprint arXiv:2310.08425 (2023)
38. Shen, Y., Li, J., Cai, J.F., Xia, D.: Computationally efficient and statistically optimal robust high-dimensional linear regression (2023)

39. Su, J., Hu, L., Wang, D.: Faster rates of private stochastic convex optimization. In: International Conference on Algorithmic Learning Theory. pp. 995–1002. PMLR (2022)
40. Su, J., Hu, L., Wang, D.: Faster rates of differentially private stochastic convex optimization. Journal of Machine Learning Research **25**(114), 1–41 (2024)
41. Sun, Q., Zhou, W., Fan, J.: Adaptive huber regression. Journal of the American Statistical Association **115**(529), 254–265 (2020)
42. Talwar, K., Thakurta, A., Zhang, L.: Nearly-optimal private lasso. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. pp. 3025–3033 (2015)
43. Tang, J., Korolova, A., Bai, X., Wang, X., Wang, X.: Privacy loss in apple's implementation of differential privacy on macos 10.12. CoRR (2017)
44. Tao, Y., Wu, Y., Cheng, X., Wang, D.: Private stochastic convex optimization and sparse learning with heavy-tailed data revisited. In: 31st International Joint Conference on Artificial Intelligence, IJCAI 2022. pp. 3947–3953. International Joint Conferences on Artificial Intelligence Organization (2022)
45. Tong, H.: Functional linear regression with huber loss. Journal of Complexity **74**, 101696 (2023)
46. Wang, D., Chen, C., Xu, J.: Differentially private empirical risk minimization with non-convex loss functions. In: International Conference on Machine Learning. pp. 6526–6535. PMLR (2019)
47. Wang, D., Gaboardi, M., Smith, A., Xu, J.: Empirical risk minimization in the non-interactive local model of differential privacy. Journal of Machine Learning Research **21**(200), 1–39 (2020)
48. Wang, D., Xiao, H., Devadas, S., Xu, J.: On differentially private stochastic convex optimization with heavy-tailed data. In: International Conference on Machine Learning. pp. 10081–10091. PMLR (2020)
49. Wang, D., Xu, J.: Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 1182–1189 (2019)
50. Wang, D., Xu, J.: On sparse linear regression in the local differential privacy model. In: International Conference on Machine Learning. pp. 6628–6637. PMLR (2019)
51. Wang, D., Xu, J.: Escaping saddle points of empirical risk privately and scalably via dp-trust region method. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 90–106. Springer (2020)
52. Wang, D., Xu, J.: On sparse linear regression in the local differential privacy model. IEEE Transactions on Information Theory **67**(2), 1182–1200 (2020)
53. Wang, D., Xu, J.: Differentially private ℓ_1-norm linear regression with heavy-tailed data. In: 2022 IEEE International Symposium on Information Theory (ISIT). pp. 1856–1861. IEEE (2022)
54. Wang, D., Xu, J.: Private least absolute deviations with heavy-tailed data. Theoretical Computer Science **1030**, 115071 (2025)
55. Wang, D., Ye, M., Xu, J.: Differentially private empirical risk minimization revisited: Faster and more general. Advances in Neural Information Processing Systems **30** (2017)
56. Wang, L., Gu, Q.: A knowledge transfer framework for differentially private sparse learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 6235–6242 (2020)
57. Woolson, R.F., Clarke, W.R.: Statistical Methods for the Analysis of Biomedical Data, vol. 371. John Wiley & Sons (2011)

58. Xue, Z., Yang, S., Huai, M., Wang, D.: Differentially private pairwise learning revisited. In: 30th International Joint Conference on Artificial Intelligence, IJCAI 2021. pp. 3242–3248. International Joint Conferences on Artificial Intelligence Organization (2021)
59. Zhang, R., Lei, M., Ding, M., Xiang, Z., Xu, J., Wang, D.: Improved rates of differentially private nonconvex-strongly-concave minimax optimization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 22524–22532 (2025)
60. Zhu, L., Ding, M., Aggarwal, V., Xu, J., Wang, D.: Improved analysis of sparse linear regression in local differential privacy model. arXiv preprint arXiv:2310.07367 (2023)
61. Zhu, L., Manseur, A., Ding, M., Liu, J., Xu, J., Wang, D.: Truthful high dimensional sparse linear regression. arXiv preprint arXiv:2410.13046 (2024)

A Proof of Theorem 1

Proof. Consider two neighboring datasets D and D' that differ by exactly one record. At each iteration of DP-IHT-H, we update

$$\boldsymbol{\beta}^{t+0.5} \;=\; \boldsymbol{\beta}^{t} \;-\; \eta\,\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^{t}).$$

Within both the quadratic and absolute-value regimes of the Huber function, the gradient satisfies $\|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^t)\|_\infty \le \tau K$ because of clipping each feature x_i to $\|\tilde{x}_i\|_\infty \le K$. Hence,

$$\begin{aligned}\left\|\boldsymbol{\beta}^{t+0.5} - \boldsymbol{\beta}'^{t+0.5}\right\|_\infty &= \left\|-\eta\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^t) + \eta\nabla\mathcal{L}_\tau(\boldsymbol{\beta}'^t)\right\|_\infty \\ &\le \eta\left\|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^t) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta}'^t)\right\|_\infty.\end{aligned}$$

Because D and D' differ by only one sample, the change in their gradients is on the order of $\frac{\tau K}{m}$. Therefore,

$$\left\|\boldsymbol{\beta}^{t+0.5} - \boldsymbol{\beta}'^{t+0.5}\right\|_\infty \;\le\; \eta\cdot\frac{\tau K}{m}.$$

By applying the composition property of differential privacy ("Peeling" is (ε,δ)-DP was proved in lemma 3.3 of [10]) and noting that each iteration's update is influenced by only one sample to a bounded extent, we conclude that every step is (ε,δ)-DP. Consequently, the entire procedure maintains (ε,δ)-DP.

B Lemmas for Theorem 2

There are two main preparations for Theorem 2. The first part relates to the properties of Restricted Strong Convexity (RSC) and Restricted Strong Smoothness (RSS). The second part concerns the properties of the "Peeling" algorithm.

We start by discussing key properties related to RSC and RSS.

Lemma 1 (Restricted Strong Convexity). *Under Assumption 1 and 2 for* $\tau \ge 2\max\{(4v_\zeta)^{\frac{1}{1+\zeta}}, 4A_1^2 r\}$ *and* $n \gtrsim (\tau/r)^2(d+t)$*, with probability at least* $1-e^{-t}$*, we have:*

$$\langle\nabla\mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle \ge c_l\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2$$

uniformly over $\boldsymbol{\beta}\in\Theta_0(r) = \{\boldsymbol{\beta}\in\mathbb{R}^d : \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}\le r\}$.

Proof. The proof of this lemma follows directly from Lemma 4 in [41].

Lemma 2 (Restricted Strong Smoothness). *Under Assumption 1 and 2 with* $0 < c_l < c_u$*, we have:*

$$\langle\nabla\mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle \le c_u\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2.$$

Proof. By the mean value property:

$$\|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2 = \|\nabla^2 L_\tau(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2,$$

where $\bar{\boldsymbol{\beta}}$ is a point on the line segment between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$.

From Assumption 2, we know:

$$\begin{aligned}\langle u, \nabla^2 L_\tau(\bar{\boldsymbol{\beta}})u\rangle &= \langle u, \frac{1}{n}\sum_{i=1}^n x_i x_i^\top 1(|y_i - x_i^\top\bar{\boldsymbol{\beta}}| \le \tau)u\rangle \\ &\le u^\top S_n u \le c_u u^\top u,\end{aligned}$$

where S_n is the empirical covariance matrix.

Since $\nabla^2 L_\tau(\bar{\boldsymbol{\beta}})$ is positive definite, we also have:

$$\|\nabla^2 L_\tau(\bar{\boldsymbol{\beta}})^\top u\|_2^2 \le c_u^2\|u\|_2^2 \quad \Rightarrow \quad \|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2^2 \le c_u^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2.$$

Thus:

$$\langle\nabla\mathcal{L}_\tau(\boldsymbol{\beta})-\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle \le \|\nabla\mathcal{L}_\tau(\boldsymbol{\beta})-\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le c_u\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2.$$

Lemma 3. *Under assumption 1 and 2 for $0 < \delta \le 1$, with probability at least $1 - 2e^{-t}$, we have:*

$$\|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty \lesssim 2L\sqrt{\frac{v_\zeta\tau^{1-\delta}t}{n}} + L\frac{\tau t}{2n} + Lv_\zeta\tau^{-\delta}.$$

By tuning $\tau \asymp (n/t)^{\frac{1}{1+\zeta}}$, it also holds that:

$$\|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty \lesssim \tau L\frac{t}{n}.$$

Proof. Under Assumption 2 with $0 < \delta < 1$, define $\xi_i = \psi_\tau(\varepsilon_i)$, where ε_i is the regression error. Then:

$$\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*) = -\frac{1}{n}\sum_{i=1}^n \xi_i\boldsymbol{x}_i.$$

For each $1 \le j \le d$:

$$|\mathbb{E}(\xi_i x_{ij})| = |x_{ij}| \cdot |\mathbb{E}(\xi_i)| \le Lv_\zeta\tau^{-\delta}.$$

Using Bernstein's inequality, with probability at least $1 - 2e^{-t}$:

$$\left|\frac{1}{n}\sum_{i=1}^n(\xi_i x_{ij} - \mathbb{E}\xi_i x_{ij})\right| \lesssim 2L\sqrt{\frac{v_\zeta\tau^{1-\delta}t}{n}} + L\frac{\tau t}{2n}.$$

Combining these results gives:

$$\|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty \lesssim 2L\sqrt{\frac{v_\zeta\tau^{1-\delta}t}{n}} + L\frac{\tau t}{2n} + Lv_\zeta\tau^{-\delta}.$$

By setting $\tau \asymp (n/t)^{\frac{1}{1+\zeta}}$, we further simplify:

$$\|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty \lesssim \tau L\frac{t}{n}.$$

Then we introduce two key properties of the "Peeling" algorithm (Algorithm 2).

Lemma 4. *([10]) Let $\tilde{P}_s$ be defined as in Algorithm 2. For any index set I, any $\boldsymbol{v} \in \mathbb{R}^I$ and $\hat{\boldsymbol{v}}$ such that $\|\hat{\boldsymbol{v}}\|_0 \leq \hat{s} \leq s$, we have that for every $c > 0$,*

$$\left\|\tilde{P}_s(\boldsymbol{v}) - \boldsymbol{v}\right\|_2^2 \leq (1 + 1/c)\frac{|I| - s}{|I| - \hat{s}}\|\hat{\boldsymbol{v}} - \boldsymbol{v}\|_2^2 + 4(1 + c)\sum_{i\in[s]}\|\boldsymbol{w}_i\|_\infty^2$$

Proof. Let $\psi : R_2 \to R_1$ be a bijection. By the selection criterion of Algorithm 2, for each $j \in R_2$ we have $|v_j| + w_{ij} \leq \left|v_{\psi(j)}\right| + w_{i\psi(j)}$, where i is the index of the iteration in which $\psi(j)$ is appended to S. It follows that, for every $c > 0$,

$$\begin{aligned} v_j^2 &\leq \left(\left|v_{\psi(j)}\right| + w_{i\psi(j)} - w_{ij}\right)^2 \\ &\leq (1 + 1/c)v_{\psi(j)}^2 + (1 + c)\left(w_{i\psi(j)} - w_{ij}\right)^2 \leq (1 + 1/c)v_{\psi(j)}^2 + 4(1 + c)\|\boldsymbol{w}_i\|_\infty^2 \end{aligned}$$

Summing over j then leads to

$$\|\boldsymbol{v}_{R_2}\|_2^2 \leq (1 + 1/c)\|\boldsymbol{v}_{R_1}\|_2^2 + 4(1 + c)\sum_{i\in[s]}\|\boldsymbol{w}_i\|_\infty^2$$

Lemma 5. *([10]) Let $\bar{P}_s$ be defined as in Algorithm 2. For any index set I, any $\boldsymbol{v} \in \mathbb{R}^I$ and $\hat{\boldsymbol{v}}$ such that $\|\hat{\boldsymbol{v}}\|_0 \leq \hat{s} \leq s$, we have that for every $c > 0$,*

$$\left\|\tilde{P}_s(\boldsymbol{v}) - \boldsymbol{v}\right\|_2^2 \leq (1 + 1/c)\frac{|I| - s}{|I| - \hat{s}}\|\hat{\boldsymbol{v}} - \boldsymbol{v}\|_2^2 + 4(1 + c)\sum_{i\in[s]}\|\boldsymbol{w}_i\|_\infty^2 .$$

Proof. Let T be the index set of the top s coordinates of v in terms of absolute values. We have

$$\begin{aligned} \left\|\tilde{P}_s(\boldsymbol{v}) - \boldsymbol{v}\right\|_2^2 &= \sum_{j\in S^c} v_j^2 = \sum_{j\in S^c\cap T^c} v_j^2 + \sum_{j\in S^c\cap T} v_j^2 \\ &\leq \sum_{j\in S^c\cap T^c} v_j^2 + (1 + 1/c)\sum_{j\in S\cap T^c} v_j^2 + 4(1 + c)\sum_{i\in[s]}\|\boldsymbol{w}_i\|_\infty^2 . \end{aligned}$$

The last step is true by observing that $|S \cap T^c| = |S^c \cap T|$ and applying lemma 4. Now, for an arbitrary $\hat{\boldsymbol{v}}$ with $\|\hat{\boldsymbol{v}}\|_0 = \hat{s} \leq s$, let $\hat{S} = \mathrm{supp}(\hat{\boldsymbol{v}})$. We have

$$\frac{1}{|I| - s}\sum_{j\in T^c} v_j^2 = \frac{1}{|T^c|}\sum_{j\in T^c} v_j^2 \overset{(*)}{\leq} \frac{1}{\left|(\hat{S})^c\right|}\sum_{j\in(\hat{S})^c} v_j^2 = \frac{1}{|I| - \hat{s}}\sum_{j\in(\hat{S})^c} v_j^2 \leq \frac{1}{|I| - \hat{s}}\sum_{j\in(\hat{S})^c}\|\hat{\boldsymbol{v}} - \boldsymbol{v}\|_2^2$$

The (*) step is true because T^c is the collection of indices with the smallest absolute values, and $|T^c| \leq \left|\hat{S}^c\right|$. We then combine the two displays above to conclude that

$$\begin{aligned}\left\|\tilde{P}_s(\boldsymbol{v})-\boldsymbol{v}\right\|_2^2 &\leq \sum_{j\in S^\bullet\cap T^c} v_j^2+(1+1/c)\sum_{j\in S\cap T^c} v_j^2+4(1+c)\sum_{i\in[s]}\|\boldsymbol{w}_i\|_\infty^2\\ &\leq (1+1/c)\sum_{j\in T^c} v_j^2+4(1+c)\sum_{i\in[s]}\|\boldsymbol{w}_i\|_\infty^2\\ &\leq (1+1/c)\frac{|I|-s}{|I|-\hat{s}}\|\hat{\boldsymbol{v}}-\boldsymbol{v}\|_2^2+4(1+c)\sum_{i\in[s]}\|\boldsymbol{w}_i\|_\infty^2\end{aligned}$$

Lemma 6. *Under assumptions 1 and 2 and event $\mathcal{E}_1=\{\Pi_R(y_i)=y_i,\forall i\in[n]\}$, RSC and RSS properties implies that there exists an absolute constant ρ such that*

$$\mathcal{L}_n\left(\boldsymbol{\beta}^{t+1}\right)-\mathcal{L}_n(\hat{\boldsymbol{\beta}})\leq\left(1-\frac{c_l}{\rho c_u}\right)\left(\mathcal{L}_n\left(\boldsymbol{\beta}^t\right)-\mathcal{L}_n(\hat{\boldsymbol{\beta}})\right)+c_3\left(\sum_{i\in[s]}\left\|\boldsymbol{w}_i^t\right\|_\infty^2+\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}^t\right\|_2^2\right)$$

for every t, where $\boldsymbol{w}_1^t,\boldsymbol{w}_2^t,\cdots,\boldsymbol{w}_s^t$ are the Laplace noise added to $\boldsymbol{\beta}^t-\left(\eta^0/n\right)\sum_{i=1}^n\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}^t-\Pi_R\left(y_i\right)\right)\boldsymbol{x}_i$ when the support of $\boldsymbol{\beta}^{t+1}$ is iteratively selected by "Peeling", S^{t+1} is the support of β^{t+1}, and $\tilde{w}^t$ is the noise vector added to the selected s-sparse vector.

Proof. For convenience, we notate as below:

▷ Let $S^t=\operatorname{supp}\left(\boldsymbol{\beta}^t\right),S^{t+1}=\operatorname{supp}\left(\boldsymbol{\beta}^{t+1}\right),S^*=\operatorname{supp}(\hat{\boldsymbol{\beta}})$, and define $I^t=S^{t+1}\cup S^t\cup S^*$.

▷ Let $g^t=\nabla\mathcal{L}_n\left(\beta^t\right)$ and $\eta^0=\eta/c_u$, where c_u is the constant in RSC and RSS property.

▷ Let $\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_s$ be the noise vectors added to $\boldsymbol{\beta}^t-\eta^0\nabla\mathcal{L}_n\left(\boldsymbol{\beta}^{\mathrm{t}}\right)$ when the support of $\boldsymbol{\beta}^{t+1}$ is iteratively selected. We define $\boldsymbol{W}=4\sum_{i\in[s]}\|\boldsymbol{w}_i\|_\infty^2$ By RSC and RSS property, we have

$$\mathcal{L}_n\left(\boldsymbol{\beta}^{t+1}\right)-\mathcal{L}_n\left(\boldsymbol{\beta}^t\right)\leq\left\langle\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^t,\boldsymbol{g}^t\right\rangle+\frac{c_u}{2}\left\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^t\right\|_2^2 \tag{2}$$

$$=\frac{c_u}{2}\left\|\boldsymbol{\beta}_{I^t}^{t+1}-\boldsymbol{\beta}_{I^t}^t+\frac{\eta}{c_u}\boldsymbol{g}_{I^t}^t\right\|_2^2-\frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{I^t}^t\right\|_2^2+(1-\eta)\left\langle\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^t,\boldsymbol{g}^t\right\rangle. \tag{3}$$

We first focus on the third term above. In what follows, c denotes an arbitrary constant greater than 1 . We may write $\boldsymbol{\beta}^{t+1}=\tilde{\boldsymbol{\beta}}^{t+1}+\tilde{\boldsymbol{w}}_{S^{t+1}}$, so that $\tilde{\boldsymbol{\beta}}^{t+1}=\tilde{P}_s\left(\boldsymbol{\beta}^t-\eta^0\nabla\mathcal{L}_n\left(\boldsymbol{\beta}^t\right)\right)$ and $\tilde{\boldsymbol{w}}$ is a vector consisting of d i.i.d. Laplace random variables.

$$\begin{aligned}\left\langle\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^t,\boldsymbol{g}^t\right\rangle&=\left\langle\boldsymbol{\beta}_{S^{t+1}}^{t+1}-\boldsymbol{\beta}_{S^{t+1}}^t,\boldsymbol{g}_{S^{t+1}}^t\right\rangle-\left\langle\boldsymbol{\beta}_{S^t\setminus S^{t+1}}^t,\boldsymbol{g}_{S^t\setminus S^{t+1}}^t\right\rangle\\&=\left\langle\overline{\boldsymbol{\beta}}_{S^{t+1}}^{t+1}-\boldsymbol{\beta}_{S^{t+1}}^t,\boldsymbol{g}_{S^{t+1}}^t\right\rangle+\left\langle\tilde{\boldsymbol{w}}_{S^{t+1}},\boldsymbol{g}_{S^{t+1}}^t\right\rangle-\left\langle\boldsymbol{\beta}_{S^t\setminus S^{t+1}}^t,\boldsymbol{g}_{S^t\setminus S^{t+1}}^t\right\rangle\end{aligned}$$

It follows that, for every $c > 1$,

$$\left\langle \boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t, \boldsymbol{g}^t \right\rangle \leq -\frac{\eta}{c_u} \left\| \boldsymbol{g}^t_{S^{t+1}} \right\|_2^2 + c \left\| \tilde{\boldsymbol{w}}_{S^{t+1}} \right\|_2^2 + (1/4c) \left\| \boldsymbol{g}^t_{S^{t+1}} \right\|_2^2 - \left\langle \boldsymbol{\beta}^t_{S^t \setminus S^{t+1}}, \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\rangle \tag{4}$$

Now for the last term in the display above, we have

$$\begin{aligned} -\left\langle \boldsymbol{\beta}^t_{S^t \setminus S^{t+1}}, \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\rangle &\leq \frac{c_u}{2\eta} \left(\left\| \boldsymbol{\beta}^t_{S^t \setminus S^{t+1}} - \frac{\eta}{c_u} \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\|_2^2 - \left(\frac{\eta}{c_u} \right)^2 \left\| \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\|_2^2 \right) \\ &\leq \frac{c_u}{2\eta} \left\| \boldsymbol{\beta}^t_{S^t \setminus S^{t+1}} - \frac{\eta}{c_u} \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\|_2^2 - \frac{\eta}{2c_u} \left\| \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\|_2^2. \end{aligned}$$

We apply lemma 4 to $\left\| \boldsymbol{\beta}_{S^t \setminus S^{t+1}} - \frac{\eta}{c_u} \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\|_2^2$ to obtain that, for every $c > 1$

$$\begin{aligned} &-\left\langle \boldsymbol{\beta}^t_{S^t \setminus S^{t+1}}, \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\rangle \leq \frac{c_u}{2\eta} \left[(1 + 1/c) \left\| \tilde{\boldsymbol{\beta}}^{t+1}_{S^{t+1} \setminus S^t} \right\|_2^2 + (1 + c) \boldsymbol{W} \right] - \frac{\eta}{2c_u} \left\| \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\|_2^2 \\ &= \frac{\eta}{2c_u} \left[(1 + 1/c) \left\| \boldsymbol{g}^t_{S^{t+1} \setminus S^t} \right\|_2^2 + (1 + c) \frac{c_u}{2\eta} \boldsymbol{W} \right] - \frac{\eta}{2c_u} \left\| \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\|_2^2. \end{aligned}$$

Plugging the inequality above back into (Eq.4) yields

$$\begin{aligned} \left\langle \boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t, \boldsymbol{g}^t \right\rangle \leq & -\frac{\eta}{c_u} \left\| \boldsymbol{g}^t_{S^{t+1}} \right\|_2^2 + c \left\| \tilde{\boldsymbol{w}}_{S^{t+1}} \right\|_2^2 + (1/4c) \left\| \boldsymbol{g}^t_{S^{t+1}} \right\|_2^2 \\ & + \frac{\eta}{2c_u} \left[(1 + 1/c) \left\| \boldsymbol{g}^t_{S^{t+1} \setminus S^t} \right\|_2^2 + (1 + c) \frac{c_u}{2\eta} \boldsymbol{W} \right] - \frac{\eta}{2c_u} \left\| \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\|_2^2 \\ \leq & \frac{\eta}{2c_u} \left\| \boldsymbol{g}^t_{S^{t+1} \setminus S^t} \right\|_2^2 - \frac{\eta}{2c_u} \left\| \boldsymbol{g}^t_{S^t \setminus S^{t+1}} \right\|_2^2 - \frac{\eta}{c_u} \left\| \boldsymbol{g}^t_{S^{t+1}} \right\|_2^2 \\ & + (1/c) \left(4 + \frac{\eta}{2c_u} \right) \left\| \boldsymbol{g}^t_{S^{t+1}} \right\|_2^2 + c \left\| \overline{\boldsymbol{w}}_{S^{t+1}} \right\|_2^2 + (1 + c) \frac{c_u}{2\eta} \boldsymbol{W} \end{aligned}$$

Finally, for the third term of (Eq.2) we have

$$\left\langle \beta^{t+1} - \boldsymbol{\beta}^t, \boldsymbol{g}^t \right\rangle \leq -\frac{\eta}{2c_u} \left\| \boldsymbol{g}^t_{S^t \cup S^{t+1}} \right\|_2^2 + (1/c) \left(4 + \frac{\eta}{2c_u} \right) \left\| \boldsymbol{g}^t_{S^{t+1}} \right\|_2^2 + c \left\| \tilde{\boldsymbol{w}}_{S^{t+1}} \right\|_2^2 + (1+c) \frac{c_u}{2\eta} \boldsymbol{W}$$

Now combining this bound with (Eq.2) yields

$$
\begin{aligned}
&\mathcal{L}_n\left(\boldsymbol{\beta}^{t+1}\right)-\mathcal{L}_n\left(\boldsymbol{\beta}^{t}\right)\\
\leq&\frac{c_u}{2}\left\|\boldsymbol{\beta}_{I^t}^{t+1}-\boldsymbol{\beta}_{I^t}^{t}+\frac{\eta}{c_u}\boldsymbol{g}_{I^t}^{t}\right\|_2^2-\frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{I^t}^{t}\right\|_2^2-\frac{\eta(1-\eta)}{2c_u}\left\|\boldsymbol{g}_{S^t\cup S^{t+1}}^{t}\right\|_2^2\\
&+\frac{1-\eta}{c}\left(4+\frac{\eta}{2c_u}\right)\left\|\boldsymbol{g}_{S^{t+1}}^{t}\right\|_2^2+(1-\eta)c\left\|\overline{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2+(1-\eta)(1+c)\frac{c_u}{2\eta}\boldsymbol{W}\\
\leq&\frac{c_u}{2}\left\|\boldsymbol{\beta}_{I^t}^{t+1}-\boldsymbol{\beta}_{I^t}^{t}+\frac{\eta}{c_u}\boldsymbol{g}_{I^t}^{t}\right\|_2^2-\frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{I^t\setminus(S^t\cup S^*)}^{t}\right\|_2^2-\frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{S^t\cup S^*}^{t}\right\|_2^2-\frac{\eta(1-\eta)}{2c_u}\left\|\boldsymbol{g}_{S^t\cup S^{t+1}}^{t}\right\|_2^2\\
&+\frac{1-\eta}{c}\left(4+\frac{\eta}{2c_u}\right)\left\|\boldsymbol{g}_{S^{t+1}}^{t}\right\|_2^2+(1-\eta)c\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2+(1-\eta)(1+c)\frac{c_u}{2\eta}\boldsymbol{W}\\
\leq&\frac{c_u}{2}\left\|\boldsymbol{\beta}_{I^t}^{t+1}-\boldsymbol{\beta}_{I^t}^{t}+\frac{\eta}{c_u}\boldsymbol{g}_{I^t}^{t}\right\|_2^2-\frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{I^t\setminus(S^t\cup S^*)}^{t}\right\|_2^2-\frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{S^t\cup S^*}^{t}\right\|_2^2-\frac{\eta(1-\eta)}{2c_u}\left\|\boldsymbol{g}_{S^{t+1}\setminus(S^t\cup S^*)}^{t}\right\|_2^2\\
&+\frac{1-\eta}{c}\left(4+\frac{\eta}{2c_u}\right)\left\|\boldsymbol{g}_{S^{t+1}}^{t}\right\|_2^2+(1-\eta)c\left\|\overline{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2+(1-\eta)(1+c)\frac{c_u}{2\eta}\boldsymbol{W}.
\end{aligned}
$$

The last step is true because $S^{t+1}\setminus(S^t\cup S^*)$ is a subset of $S^t\cup S^{t+1}$. We next analyze the first two terms, $\frac{c_u}{2}\left\|\boldsymbol{\beta}_{I^t}^{t+1}-\boldsymbol{\beta}_{I^t}^{t}+\frac{\eta}{c_u}\boldsymbol{g}_{I^t}^{t}\right\|_2^2-\frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{I^t\setminus(S^t\cup S^*)}^{t}\right\|_2^2$. Let R be a subset of $S^t\setminus S^{t+1}$ such that $|R|=|I^t\setminus(S^t\cup S^*)|=|\ S^{t+1}\setminus(S^t\cup\ S^*)\ |$.

By the definition of $\tilde{\boldsymbol{\beta}}^{t+1}$ and lemma 4, we have, for every $c>1$,

$$
\frac{\eta^2}{c_u^2}\left\|\boldsymbol{g}_{I^t\setminus(S^t\cup S^*)}^{t}\right\|_2^2=\left\|\tilde{\boldsymbol{\beta}}_{I^t\setminus(S^t\cup S^*)}^{t+1}\right\|_2^2\geq(1-1/c)\left\|\boldsymbol{\beta}_R^t-\frac{\eta}{c_u}\boldsymbol{g}_R^t\right\|_2^2-c\boldsymbol{W}.
$$

It follows that

$$
\begin{aligned}
&\frac{c_u}{2}\left\|\boldsymbol{\beta}_{I^t}^{t+1}-\boldsymbol{\beta}_{I^t}^{t}+\frac{\eta}{c_u}\boldsymbol{g}_{I^t}^{t}\right\|_2^2-\frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{I^t\setminus(S^t\cup S^*)}^{t}\right\|_2^2\\
&\leq\frac{c_u}{2}\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2+\frac{c_u}{2}\left\|\tilde{\boldsymbol{\beta}}_{I^t}^{t+1}-\boldsymbol{\beta}_{I^t}^{t}+\frac{\eta}{c_u}\boldsymbol{g}_{I^t}^{t}\right\|_2^2-\frac{c_u}{2}(1-1/c)\left\|\boldsymbol{\beta}_R^t-\frac{\eta}{c_u}\boldsymbol{g}_R^t\right\|_2^2+\frac{cc_u}{2}\boldsymbol{W}\\
&=\frac{c_u}{2}\left\|\tilde{\boldsymbol{\beta}}_{I^t}^{t+1}-\boldsymbol{\beta}_{I^t}^{t}+\frac{\eta}{c_u}\boldsymbol{g}_{I^t}^{t}\right\|_2^2-\frac{c_u}{2}\left\|\mid\tilde{\boldsymbol{\beta}}_R^{t+1}-\boldsymbol{\beta}_R^t+\frac{\eta}{c_u}\boldsymbol{g}_R^t\right\|_2^2+\frac{c_u}{2}(1/c)\left\|\boldsymbol{\beta}_R^t-\frac{\eta}{c_u}\boldsymbol{g}_R^t\right\|_2^2+\frac{cc_u}{2}\boldsymbol{W}\\
&+\frac{c_u}{2}\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2\\
&\leq\frac{c_u}{2}\left\|\tilde{\boldsymbol{\beta}}_{I^t\setminus R}^{t+1}-\boldsymbol{\beta}_{I^t\setminus R}^{t}+\frac{\eta}{c_u}\boldsymbol{g}_{I^t\setminus R}^{t}\right\|_2^2+\frac{\eta^2}{2cc_u}(1+1/c)\left\|\boldsymbol{g}_{I^t\setminus(S^t\cup S^*)}^{t}\right\|_2^2+\frac{cc_u}{2}\boldsymbol{W}+\frac{c_u}{2}\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2.
\end{aligned}
$$

The last inequality is obtained by applying lemma 4 to $\left\|\boldsymbol{\beta}_R^t - \frac{\eta}{c_u}\boldsymbol{g}_R^t\right\|_2^2$. Now we apply lemma 5 to obtain

$$
\begin{aligned}
&\frac{c_u}{2}\left\|\boldsymbol{\beta}_{I^t}^{t+1} - \boldsymbol{\beta}_{I^t}^t + \frac{\eta}{c_u}\boldsymbol{g}_{I^t}^t\right\|_2^2 - \frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{I^t\setminus(S^t\cup S^*)}^t\right\|_2^2 \\
&\leq \frac{3c_u}{4}\frac{|I^t\setminus R| - s}{|I^t\setminus R| - s^*}\left\|\boldsymbol{\beta}_{I^t\setminus R} - \boldsymbol{\beta}_{I^t\setminus R}^t + \frac{\eta}{c_u}\boldsymbol{g}_{I^t\setminus R}^t\right\|_2^2 + \frac{3c_u}{2}\boldsymbol{W} \\
&+ \frac{\eta^2\left(1+c^{-1}\right)}{2cc_u}\left\|\boldsymbol{g}_{I^\setminus\setminus(S^t\cup S^*)}^t\right\|_2^2 + \frac{cc_u}{2}\boldsymbol{W} + \frac{c_u}{2}\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2 \\
&\leq \frac{3c_u}{4}\frac{2s^*}{s+s^*}\left\|\boldsymbol{\beta}_{I^t\setminus R} - \boldsymbol{\beta}_{I^t\setminus R}^t + \frac{\eta}{c_u}\boldsymbol{g}_{I^t\setminus R}^t\right\|_2^2 + \frac{3c_u}{2}\boldsymbol{W} + \frac{\eta^2}{2cc_u}(1+1/c)\left\|\boldsymbol{g}_{S^{t+1}}^t\right\|_2^2 \\
&+ \frac{cc_u}{2}\boldsymbol{W} + \frac{c_u}{2}\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2
\end{aligned}
$$

The last step is true by observing that $|I^t\setminus R| \leq 2s^* + s$, and the inclusion $I^t\setminus\left(S^t\cup S^*\right) \subseteq S^{t+1}$. We continue to simplify,

$$
\begin{aligned}
&\frac{c_u}{2}\left\|\boldsymbol{\beta}_{I^t}^{t+1} - \boldsymbol{\beta}_{I^t}^t + \frac{\eta}{c_u}\boldsymbol{g}_{I^t}^t\right\|_2^2 - \frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{I^t\setminus(S^t\cup S^*)}^t\right\|_2^2 \\
&\leq \frac{c_u}{2}\frac{3s^*}{s+s^*}\left\|\boldsymbol{\beta}_{I^t} - \boldsymbol{\beta}_{I^t}^t + \frac{\eta}{c_u}\boldsymbol{g}_{I^t}^t\right\|_2^2 + \frac{3c_u}{2}\boldsymbol{W} + \frac{\eta^2}{2cc_u}(1+1/c)\left\|\boldsymbol{g}_{S^{t+1}}^t\right\|_2^2 + \frac{cc_u}{2}\boldsymbol{W} + \frac{c_u}{2}\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2 \\
&\leq \frac{3s^*}{s+s^*}\left(\eta\left\langle\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t, \boldsymbol{g}^t\right\rangle + \frac{c_u}{2}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\right\|_2^2 + \frac{\eta^2}{2cc_u}\left\|\boldsymbol{g}_{I^t}^t\right\|_2^2\right) \\
&+ \frac{\eta^2}{2cc_u}(1+1/c)\left\|\boldsymbol{g}_{S^{t+1}}^t\right\|_2^2 + \frac{(c+3)c_u}{2}\boldsymbol{W} + \frac{c_u}{2}\left\|\overline{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2 \\
&\leq \frac{3s^*}{s+s^*}\left(\eta\mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \eta\mathcal{L}_n\left(\boldsymbol{\beta}^t\right) + \frac{c_u - \eta c_l}{2}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\right\|_2^2 + \frac{\eta^2}{2cc_u}\left\|\boldsymbol{g}_{I^t}^t\right\|_2^2\right) \\
&\quad + \frac{\eta^2}{2cc_u}(1+1/c)\left\|\boldsymbol{g}_{S^{t+1}}^t\right\|_2^2 + \frac{(c+3)c_u}{2}\boldsymbol{W} + \frac{c_u}{2}\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2.
\end{aligned}
$$

Until now, the inequality is true for any $0 < \eta < 1$ and $c > 1$. We now specify the choice of these parameters: let $\eta = 2/3$ and set c large enough so that

$$
\begin{aligned}
\mathcal{L}_n\left(\boldsymbol{\beta}^{t+1}\right) - \mathcal{L}_n\left(\boldsymbol{\beta}^t\right) \leq &\frac{3s^*}{s+s^*}\left(\eta\mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \eta\mathcal{L}_n\left(\boldsymbol{\beta}^t\right) + \frac{c_u - \eta c_l}{2}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\right\|_2^2 + \frac{\eta^2}{2c_u}\left\|\boldsymbol{g}_{I^t}^t\right\|_2^2\right) \\
&- \frac{\eta^2}{4c_u}\left\|\boldsymbol{g}_{S^t\cup S^*}^t\right\|_2^2 - \frac{\eta(1-\eta)}{4c_u}\left\|\boldsymbol{g}_{S^{t+1}\setminus(S^t\cup S^*)}^t\right\|_2^2 \\
&+ \frac{c_u}{2}\left(\frac{3c+7}{2}\right)\boldsymbol{W} + \left(\frac{c}{3} + \frac{c_u}{2}\right)\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2.
\end{aligned}
$$

Such a choice of c is available because c_u is bounded above by an absolute constant thanks to the RSM condition (upper inequality of RSC and RSS property). Now we set $s = 72(c_u/c_l)^2 s^* = \rho L^4 s^*$, where ρ is the absolute constant

referred to in Lemma 8.3 and Theorem 4.4 in [41], so that $\frac{3s^*}{s+s^*} \leq \frac{c_l^2}{24c_u(c_u-\eta c_l)}$, and $\frac{c_l^2}{24c_u(c_u-\eta c_l)} \leq 1/8$ because $c_l < c_u$. It follows that

$$\begin{aligned}\mathcal{L}_n\left(\boldsymbol{\beta}^{t+1}\right) - \mathcal{L}_n\left(\boldsymbol{\beta}^t\right) \leq & \frac{3s^*}{s+s^*}\left(\eta\mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \eta\mathcal{L}_n\left(\boldsymbol{\beta}^t\right)\right) + \frac{c_l^2}{48c_u}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\right\|_2^2 + \frac{1}{36c_u}\left\|\boldsymbol{g}_{I^t}^t\right\|_2^2 \\ & - \frac{1}{9c_u}\left\|\boldsymbol{g}_{S^t\cup S^*}^t\right\|_2^2 - \frac{1}{18c_u}\left\|\boldsymbol{g}_{S^{t+1}\setminus(S^t\cup S^*)}^t\right\|_2^2 \\ & + \frac{c_u}{2}\left(\frac{3c+7}{2}\right)\boldsymbol{W} + \left(\frac{c}{3}+\frac{c_u}{2}\right)\left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2\end{aligned}$$

Because $\left\|\boldsymbol{g}_{I^t}^t\right\|_2^2 = \left\|\boldsymbol{g}_{S^t\cup S^*}^t\right\|_2^2 + \left\|\boldsymbol{g}_{S^{t+1}\setminus(S^t\cup S^*)}^t\right\|_2^2$, we have

$$\begin{aligned}\mathcal{L}_n\left(\boldsymbol{\beta}^{t+1}\right) - \mathcal{L}_n\left(\boldsymbol{\beta}^t\right) \leq & \frac{3s^*}{s+s^*}\left(\eta\mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \eta\mathcal{L}_n\left(\boldsymbol{\beta}^t\right)\right) + \frac{c_l^2}{48c_u}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\right\|_2^2 - \frac{3}{36c_u}\left\|\boldsymbol{g}_{S^t\cup S^*}^t\right\|_2^2 \\ & + \frac{c_u}{2}\left(\frac{3c+7}{2}\right)\boldsymbol{W} + \left(\frac{c}{3}+\frac{c_u}{2}\right)\left\|\overline{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2 \\ \leq & \frac{3s^*}{s+s^*}\left(\eta\mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \eta\mathcal{L}_n\left(\boldsymbol{\beta}^t\right)\right) - \frac{3}{36c_u}\left(\left\|\boldsymbol{g}_{S^t\cup S^*}^t\right\|_2^2 - \frac{c_l^2}{4}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\right\|_2^2\right) \\ & + \frac{c_u}{2}\left(\frac{3c+7}{2}\right)\boldsymbol{W} + \left(\frac{c}{3}+\frac{c_u}{2}\right)\left\|\overline{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2 \text{ (Eq.3).}\end{aligned}$$

To continue the calculations, we consider a lemma from Lemma A.4[**?**]

$$\left\|g_{S^t\cup S^*}^t\right\|_2^2 - \frac{c_l^2}{4}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\right\|_2^2 \geq \frac{c_l}{2}\left(\mathcal{L}_n\left(\boldsymbol{\beta}^t\right) - \mathcal{L}_n(\hat{\boldsymbol{\beta}})\right)$$

It then follows from (Eq.3), the quoted lemma above and the definition of ρ that, for an appropriate constant c_3,

$$\begin{aligned}\mathcal{L}_n\left(\boldsymbol{\beta}^{t+1}\right) - \mathcal{L}_n\left(\boldsymbol{\beta}^t\right) &\leq -\left(\frac{3c_l}{72c_u} + \frac{2s^*}{s+s^*}\right)\left(\mathcal{L}_n\left(\boldsymbol{\beta}^t\right) - \mathcal{L}_n(\hat{\boldsymbol{\beta}})\right) + c_3\left(\boldsymbol{W} + \left\|\overline{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2\right) \\ &\leq -\left(\frac{c_l}{\rho c_u}\right)\left(\mathcal{L}_n\left(\boldsymbol{\beta}^t\right) - \mathcal{L}_n(\hat{\boldsymbol{\beta}})\right) + c_3\left(\boldsymbol{W} + \left\|\tilde{\boldsymbol{w}}_{S^{t+1}}\right\|_2^2\right)\end{aligned}$$

Adding $\mathcal{L}_n\left(\boldsymbol{\beta}^t\right) - \mathcal{L}_n(\hat{\boldsymbol{\beta}})$ to both sides of the inequality concludes the proof.

C Proof of Theorem 2

Proof. Using the lemmas stated above, we iterate over t and denote

$$\boldsymbol{W}_t = c_3\left(\sum_{i\in[s]}\left\|\boldsymbol{w}_i^t\right\|_\infty^2 + \left\|\tilde{\boldsymbol{w}}_{S^{t+1}}^t\right\|_2^2\right).$$

Then, we have

$$\mathcal{L}_n(\boldsymbol{\beta}^T) - \mathcal{L}_n(\boldsymbol{\beta}^*) \leq \left(1 - \frac{c_l}{\rho c_u}\right)^T \left(\mathcal{L}_n(\boldsymbol{\beta}^0) - \mathcal{L}_n(\boldsymbol{\beta}^*)\right) + \sum_{k=0}^{T-1} \left(1 - \frac{c_l}{\rho c_u}\right)^{T-k-1} \boldsymbol{W}_k$$
$$\leq \left(1 - \frac{c_l}{\rho c_u}\right)^T 8Ac_0^2 + \sum_{k=0}^{T-1} \left(1 - \frac{c_l}{\rho c_u}\right)^{T-k-1} \boldsymbol{W}_k,$$

where the second inequality follows from the RSS and RSC properties and the ℓ_2 bounds on $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}^*$.

On the other hand, by the RSC property we also have the lower bound

$$\mathcal{L}_n(\boldsymbol{\beta}^T) - \mathcal{L}_n(\boldsymbol{\beta}^*) \geq \frac{c_l}{2}\|\boldsymbol{\beta}^T - \boldsymbol{\beta}^*\|_2^2 - \left\langle \nabla \mathcal{L}_n(\boldsymbol{\beta}^*),\, \boldsymbol{\beta}^* - \boldsymbol{\beta}^T \right\rangle.$$

Combining these bounds and noting that for $T \asymp \log n$, we obtain

$$\frac{c_l}{2}\|\boldsymbol{\beta}^T - \boldsymbol{\beta}^*\|_2^2 \leq \|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \sqrt{s+s^*}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^T\|_2 + \frac{1}{n} + \sum_{k=0}^{T-1} \left(1 - \frac{c_l}{\rho c_u}\right)^{T-k-1} \boldsymbol{W}_k. \tag{5}$$

To further bound $\|\boldsymbol{\beta}^T - \boldsymbol{\beta}^*\|_2^2$, we consider the event

$$\mathcal{E}_2 = \left\{ \max_t \boldsymbol{W}_t \leq K \, \frac{L^2\tau^2 (s^*)^2 \log^2 d \, \log\left(\frac{1}{\sigma}\right) t^2}{n^2 \varepsilon^2} \right\},$$

in addition to the event $\mathcal{E}_1$ defined earlier. Then, by applying Eq. (5), Assumptions 1 and 2, and Lemma 6, we deduce that

$$\|\boldsymbol{\beta}^T - \boldsymbol{\beta}^*\|_2 \lesssim (s^*)^{\frac{1}{2}} \left\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\right\|_\infty + \frac{L\tau s^* \, \log d \, \log^{\frac{1}{2}}\left(\frac{1}{\sigma}\right) t}{n\varepsilon}.$$

Regardless of whether the noise is random or fixed, the leading term of $\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)$ is given by

$$C_1 \tau L \frac{t}{n} + C_2 \tau^{-\delta},$$

as established in Lemma 3. Therefore, the optimal tuning for τ is

$$\tau \asymp \left(\frac{t}{n} \left(1 + \frac{s^{*\frac{1}{2}} \log d \, \log^{\frac{1}{2}}\left(\frac{1}{\sigma}\right)}{\varepsilon} \right) \right)^{-\max\left\{\frac{1}{2}, \frac{1}{1+\zeta}\right\}}.$$

It remains to show that the events $\mathcal{E}_1$ and $\mathcal{E}_2$ occur with high probability. As shown in [10], we have

$$\mathbb{P}\left(\mathcal{E}_1^c\right) \leq c_1 \exp\left(-c_2 \, t \log d\right) \quad \text{and} \quad \mathbb{P}\left(\mathcal{E}_3^c\right) \leq 2e^{-2\log d},$$

where $\mathcal{E}_3$ is another event defined in the previous context. This completes the proof.

D Proof of Theorem 4

Proof. Let $\boldsymbol{\beta}^0$ denote the initial estimate and $\boldsymbol{\beta}^*$ the true parameter vector. Define

$$D_t \;=\; (1-c_1)^t \,\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\|_2,$$

where $c_1 \in (0,1)$ is the contraction constant determined by the algorithm's specifications. We prove by induction on t that

$$\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2 \;\le\; D_t,$$

up to an additive error of order $O\big(\sqrt{\boldsymbol{W}}\big)$.

Base Case. For $t = 0$, the claim is immediate:

$$\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\|_2 = D_0.$$

Inductive Step. Suppose that

$$\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2 \le D_t$$

for some $t \ge 0$. We now show that

$$\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\|_2 \;\le\; D_{t+1} + O\big(\sqrt{\boldsymbol{W}}\big).$$

According to the algorithm, the update is performed in two stages. First, a gradient descent step is executed:

$$\boldsymbol{\beta}^{t+0.5} \;=\; \boldsymbol{\beta}^t - \eta_t \,\mathbf{G}_t,$$

where $\mathbf{G}_t$ is a subgradient of $f(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}^t$. By the parallelogram identity, we have

$$\|\boldsymbol{\beta}^{t+0.5} - \boldsymbol{\beta}^*\|_2^2 = \|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2^2 - 2\eta_t \,\langle \boldsymbol{\beta}^t - \boldsymbol{\beta}^*, \mathbf{G}_t\rangle + \eta_t^2 \,\|\mathbf{G}_t\|_2^2.$$

Since f is convex, it follows that

$$\langle \boldsymbol{\beta}^t - \boldsymbol{\beta}^*, \mathbf{G}_t\rangle \ge f(\boldsymbol{\beta}^t) - f(\boldsymbol{\beta}^*).$$

Thus,

$$\|\boldsymbol{\beta}^{t+0.5} - \boldsymbol{\beta}^*\|_2^2 \le \|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2^2 - 2\eta_t \Big[f(\boldsymbol{\beta}^t) - f(\boldsymbol{\beta}^*)\Big] + \eta_t^2 \,\|\mathbf{G}_t\|_2^2.$$

By Lemma 1 in [38], there exists a constant $c_l > 0$ such that

$$f(\boldsymbol{\beta}^t) - f(\boldsymbol{\beta}^*) \ge \frac{n}{4}\, c_l^{1/2} \,\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2,$$

and, moreover, $\|\mathbf{G}_t\|_2 \le n\, c_u^{1/2}$ for some constant $c_u > 0$. Substituting these bounds gives

$$\|\boldsymbol{\beta}^{t+0.5} - \boldsymbol{\beta}^*\|_2^2 \le \|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2^2 - \eta_t \,\frac{n}{2}\, c_l^{1/2} \,\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2 + \eta_t^2 \, n^2 \, c_u.$$

By the induction hypothesis, $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2 \leq D_t$, so that

$$\|\boldsymbol{\beta}^{t+0.5} - \boldsymbol{\beta}^*\|_2^2 \leq D_t^2 - \eta_t \frac{n}{2} c_l^{1/2} D_t + \eta_t^2 n^2 c_u.$$

To obtain $\boldsymbol{\beta}^{t+1}$, the algorithm applies a projection step that incorporates privacy-preserving noise. Let $\boldsymbol{W}_t$ denote the noise introduced in iteration t, so that by Lemma 5 (or a similar projection result),

$$\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}^{t+0.5} - \boldsymbol{\beta}^*\|_2^2 + \boldsymbol{W}_t,$$

where $\boldsymbol{W}_t = 4 \sum_{i=1}^s \|w_i^t\|_\infty^2$. Consequently,

$$\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq D_t^2 - \eta_t \frac{n}{2} c_l^{1/2} D_t + \eta_t^2 n^2 c_u + \boldsymbol{W}_t.$$

By choosing the step size such that

$$\eta_t \in n^{-1} \sqrt{\frac{c_l}{c_u}} \left[\tfrac{1}{8} D_t,\ \tfrac{3}{8} D_t\right],$$

the negative (linear) term dominates the quadratic term, ensuring a geometric decrease of D_t, provided that $\boldsymbol{W}_t$ (and hence the overall noise) remains sufficiently small with high probability. Taking square roots and using standard inequalities, we deduce

$$\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\|_2 \leq (1 - C_1) D_t + \sqrt{\boldsymbol{W}_t} \leq D_{t+1} + \sqrt{\boldsymbol{W}},$$

where

$$O(\boldsymbol{W}) = O\left(\frac{(s^*)^{3/2} \log d \, (\log(1/\delta))^{1/2} \log n}{n \varepsilon} \right)$$

serves as an upper bound on the added noise. This completes the induction for Phase One.

Transition to Phase Two. Once $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2$ becomes sufficiently small, the algorithm enters Phase Two, in which a constant step size is adopted. The analysis in this phase follows analogously to that of Phase One (again leveraging Lemma 1 in [38]), and yields the contraction

$$\|\boldsymbol{\beta}_{l+1} - \boldsymbol{\beta}^*\|_2 \leq \left(1 - \widetilde{c}_2 \frac{c_l^2}{2 c_u^2} \frac{b_1^2}{b_0^2}\right) \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2,$$

where the constants are chosen appropriately.

Finally, by ensuring that additional perturbations (e.g., from noise or minor discrepancies in the index sets) remain controlled (as in Theorem 9 of [38]), we conclude that the overall algorithm converges geometrically. This completes the proof of Theorem 4.

E Proof of Theorem 5

Proof. The proof follows by combining the two-phase convergence analysis from Theorem 4.

Phase One: When

$$\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2 \geq 8\, c_l^{-1/2}\gamma,$$

with the decaying step size $\eta_t = (1-c_1)^t\eta_0$, Theorem 4 implies that

$$\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\|_2 \leq (1-c_1)^{t+1}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 + O\big(\sqrt{\boldsymbol{W}}\big).$$

Thus, after

$$T_1 = O\Big(\log\Big(\frac{\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2}{\gamma}\Big)\Big)$$

iterations, the error is reduced to

$$\|\boldsymbol{\beta}^{T_1} - \boldsymbol{\beta}^*\|_2 \leq 8\, c_l^{-1/2}\gamma.$$

Phase Two: Once the iterate enters the region

$$\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2 \leq 8\, c_l^{-1/2}\gamma,$$

a constant step size is employed so that

$$\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\|_2 \leq (1-c_2^*)\,\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2 + O\big(\sqrt{\boldsymbol{W}}\big)\big).$$

This contraction implies that after an additional

$$T_2 = O\Big(\log\Big(n\,\gamma\, b_0^{-1}\,\log\Big(\frac{2\,d}{s^*}\Big)\Big)\Big)$$

iterations the estimation error satisfies

$$\|\boldsymbol{\beta}^{T_1+T_2} - \boldsymbol{\beta}^*\|_2 \leq O\big(\sqrt{\boldsymbol{W}}\big)\big).$$

Conclusion: Combining the two phases, the total number of iterations required is

$$T = T_1 + T_2 = O\Big(\log\Big(\frac{\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2}{\gamma}\Big) + \log\Big(n\,\gamma\, b_0^{-1}\,\log\Big(\frac{2\,d}{s^*}\Big)\Big)\Big),$$

and the final estimator $\boldsymbol{\beta}^T$ satisfies

$$\|\boldsymbol{\beta}^T - \boldsymbol{\beta}^*\|_2 \leq O\Big(\frac{(s^*)^{3/2}\,\log d\,\big(\log(1/\delta)\big)^{1/2}\,\log(T/n)}{(T/n)\,\varepsilon}\Big).$$

Notice that the term

$$O\Big(\frac{(s^*)^{3/2}\,\log d\,\big(\log(1/\delta)\big)^{1/2}\,\log(T/n)}{(T/n)\,\varepsilon}\Big)$$

originates from the cumulative effect of the Laplace noise injected via the peeling procedure. More precisely, under the event

$$\mathcal{E}_2 = \left\{ \max_t \boldsymbol{W}_t \leq K \, \frac{L^2 \tau^2 \left(s^*\right)^2 \log^2 d \, \log\left(\frac{1}{\sigma}\right) \, t^2}{(T/n)^2 \varepsilon^2} \right\},$$

which holds with high probability, the noise magnitude in each iteration is well controlled. This completes the proof.